eConf: C030908
SLAC-R-703

PROCEEDINGS OF THE CONFERENCE ON STATISTICAL PROBLEMS IN PARTICLE PHYSICS, ASTROPHYSICS AND COSMOLOGY

PHYSTAT2003

Stanford Linear Accelerator Center, Stanford, California

September 8th – 11th, 2003

EDITORS

L. Lyons (Oxford)

R. Mount (SLAC)

R. Reitmeyer (SLAC)

Proceedings supported by the U.S. Department of Energy under

Contract No. DE-AC03-76SF00515

This document, and the material and data contained therein, was developed under sponsorship of the United States Government. Neither the United States nor the Department of Energy, nor the Leland Stanford Junior University, nor their employees, nor their respective contractors, subcontractors, or their employees, makes any warranty, express or implied, or assumes any liability of responsibility for accuracy, completeness, or usefulness of any information, apparatus, product or process disclosed, or represents that its use will not infringe privately owned rights. Mention of any product, its manufacturer, or suppliers shall not, nor is it intended to, imply approval, disapproval, or fitness of any particular use. A royalty-free, nonexclusive right to use and disseminate same for any purpose whatsoever, is expressly reserved to the United States and the University.

Prepared for the Department of Energy under contract number DE-AC03-76S00515 by Stanford Linear Accelerator Center, Stanford University, Stanford, California. Printed in the United States of America. Copies may be obtained by requesting SLAC-R-703 from the following address:

Stanford Linear Accelerator Center
Technical Publications Department
2575 Sand Hill Road, MS-68
Menlo Park, CA 94025
E-mail address: epubs-l@slac.stanford.edu

This document is also available online at `http://www.slac.stanford.edu/econf/C030908/`.

Preface

PHYSTAT2003 was a Conference held at SLAC on September 8th – 11th, 2003. Its theme was "Statistical Problems in Particle Physics, Astrophysics and Cosmology". It followed on from the Confidence Limits Workshops held at CERN and Fermilab in 2000, and a conference in 2002 at the Institute for Particle Physics Phenomenology in Durham, UK. PHYSTAT2003 differed from the earlier meetings in two important respects. While the first three meetings had been attended largely by particle physicists, at SLAC there was also involvement of astrophysicists and cosmologists. Secondly, the SLAC meeting really benefited from a strong presence of statisticians. They enhanced the usefulness of the Conference in many ways: they were involved in the planning stage, gave invited and contributed talks, and were simply "there" to discuss statistical issues with physicists in the breaks between the sessions.

Because of the different disciplines of the participants, the Conference started with talks by Eric Feigelson and Roger Barlow, on astrophysics and particle physics respectively. These were to introduce members of the audience to the range of statistical issues that were actively being used in fields with which they were not directly acquainted. These were followed by the Keynote Address by Bradley Efron, President of the American Statistical Association. Invited talks at the conference were divided almost equally between statisticians and physicists, and there were 2 sets of parallel sessions for contributed talks. The meeting ended with a Panel Discussion dealing with questions submitted by attendees, and with John Rice's Conference Summary.

Also included in these proceedings are articles by Sergei Bityukov, Sergei Redin, Ercan Kuroğlu, and Diego Herranz who had registered for the conference, but were unable to attend because of problems in obtaining visas to visit the USA.

I would like to take this opportunity to thank many people. First and foremost I want to express my deep appreciation to SLAC Director Jonathan Dorfan and to Director of Research Persis Drell. From the very beginning they were extremely positive about having PHYSTAT2003 at SLAC, and their spiritual and financial support were most important.

The Scientific Committee (listed on page 326.) was full of useful guidance in meetings that we had at SLAC, the CERN cafeteria and via e-mail. The smooth running of the conference was largely due to the efficient work of the Local Organizing Committee. Its Chairman Richard Mount did sterling work in smoothing over a whole range of potential difficulties. Among many other activities, Arla LeCount made sure that we all did the jobs we were supposed to, as well as bearing the brunt of the Conference's e-mail enquiries. The very efficient setting up and maintenance of the web site was due to great work by David Lee. I am most grateful to all of them. Also, the friendly and effective help provided by Jane, Jennifer and Bernice at the Conference was very much appreciated.

The value of the Conference, to a large extent, was determined by the excellent quality of the speakers. We were all particularly grateful to the statisticians who devoted time and energy patiently explaining to us that we should be aware of books that have been written since the famous tome by Kendall and Stuart, that chi-squared is not the only way of testing goodness of fit, etc., etc. My thanks extend to all the speakers of invited and contributed talks, especially for their efforts in making their material accessible to all members of the audience.

I hope that these proceedings will provide a useful record of most of the talks at the conference. Beck Reitmeyer of the SLAC Technical Publications Department was responsible for converting a random series of computer files into these Proceedings. She coped remarkably calmly with the elastic sequence of "final" deadlines, which were necessary to coax written text out of some of the speakers. She was ably helped by Yasuko Weyhrauch. I wish to thank members of the Scientific Committee who reviewed the articles appearing here. In some cases this involved active participation in the production of the final text.

Final thanks go to all participants at PHYSTAT2003 for keeping the speakers and Panel members on their toes, and for helping to make it a productive and interactive meeting. I hope that the next PHYSTAT meeting can build on the contacts and advances made at SLAC.

Louis Lyons

Table of Contents

SESSION 1 Session Chair: Bob Cousins

SESSION 2 Session Chair: John Rice

SESSION 3:A (Parallel) Session Chair: Bill Murray

SESSION 3:B (Parallel) Session Chair: Pekka Sinervo

SESSION 4:A (Parallel) Session Chair: Glen Cowan

SESSION 4:B (Parallel) Session Chair: Jim Linneman

SESSION 5 Session Chair: Jogesh Babu

SESSION 6 Session Chair: Eric Feigelson

SESSION 7:A (Parallel) Session Chair: Roger Barlow

SESSION 7:B (Parallel) Session Chair: Volker Blobel

SESSION 11:B (Parallel) Session Chair: Jeffrey Scargle

SESSION 12:A (Parallel) Session Chair: Tom Loredo

SESSION 12:B (Parallel) Session Chair: Nancy Reid

SESSION 13 Session Chair: Jerry Friedman

SESSION 14 Session Chair: Fred James

SESSION 15 Session Chair: Louis Lyons

SESSION 16 Session Chair: Brad Efron

Not Presented at the Conference

* *Manuscript not received.*

Statistical Challenges in Modern Astronomy

E. D. Feigelson
Department of Astronomy & Astrophysics, Penn State University, University Park PA 16802, USA
G. J. Babu
Department of Statistics, Penn State University, University Park PA 16802, USA

Despite centuries of close association, statistics and astronomy are surprisingly distant today. Most observational astronomical research relies on an inadequate toolbox of methodological tools. Yet the needs are substantial: astronomy encounters sophisticated problems involving sampling theory, survival analysis, multivariate classification and analysis, time series analysis, wavelet analysis, spatial point processes, nonlinear regression, bootstrap resampling and model selection. We review the recent resurgence of astrostatistical research, and outline new challenges raised by the emerging Virtual Observatory. Our essay ends with a list of research challenges and infrastructure for astrostatistics in the coming decade.

1. THE GLORIOUS HISTORY OF ASTRONOMY AND STATISTICS

Astronomy is perhaps the oldest observational science[1]. The effort to understand the mysterious luminous objects in the sky has been an important element of human culture for at least 10^4 years. Quantitative measurements of celestial phenomena were carried out by many ancient civilizations. The classical Greeks were not active observers but were unusually creative in the applications of mathematical principles to astronomy. The geometric models of the Platonists with crystalline spheres spinning around the static Earth were elaborated in detail, and this model endured in Europe for 15 centuries. But it was another Greek natural philosopher, Hipparchus, who made one of the first applications of mathematical principles that we now consider to be in the realm of statistics. Finding scatter in Bablylonian measurements of the length of a year, defined as the time between solstices, he took the middle of the range – rather than the mean or median – for the best value.

This is but one of many discussions of statistical issues in the history of astronomy. Ptolemy estimated parameters of a non-linear cosmological model using a minimax goodness-of-fit method. Al-Biruni discussed the dangers of propagating errors from inaccurate instruments and inattentive observers. While some Medieval scholars advised against the acquisition of repeated measurements, fearing that errors would compound rather than compensate for each other, the usefulnes of the mean to increase precision was demonstrated with great success by Tycho Brahe.

During the 19th century, several elements of modern mathematical statistics were developed in the context of celestial mechanics, where the application of Newtonian theory to solar system phenomena gave astonishingly precise and self-consistent quantitative inferences. Legendre developed L_2 least squares parameter estimation to model cometary orbits. The least-squares method became an instant success in European astronomy and geodesy. Other astronomers and physicists contributed to statistics: Huygens wrote a book on probability in games of chance; Newton developed an interpolation procedure; Halley laid foundations of actuarial science; Quetelet worked on statistical approaches to social sciences; Bessel first used the concept of “probable error”; and Airy wrote a volume on the theory of errors.

But the two fields diverged in the late-19th and 20th centuries. Astronomy leaped onto the advances of physics – electromagnetism, thermodynamics, quantum mechanics and general relativity – to understand the physical nature of stars, galaxies and the Universe as a whole. A subfield called “statistical astronomy” was still present but concentrated on rather narrow issues involving star counts and Galactic structure [30]. Statistics concentrated on analytical approaches. It found its principle applications in social sciences, biometrical sciences and in practical industries (*e.g.*, Sir R. A. Fisher’s employment by the British agricultural service).

2. STATISTICAL NEEDS OF ASTRONOMY TODAY

Contemporary astronomy abounds in questions of a statistical nature. In addition to exploratory data analysis and simple heuristic (usually linear) modeling common in other fields, astronomers also often interpret data in terms of complicated non-linear models based on deterministic astrophysical processes. The phenomena studied must obey known behaviors of atomic and nuclear physics, gravitation and mechanics, thermodynamics and radiative processes, and so forth. ‘Modeling’ data may thus involve both the se-

[1] The historical relationship between astronomy and statistics is described in references [15], [38] and elsewhere. Our *Astrostatistics* monograph gives more detail and contemporary examples of astrostatistical problems [3].

lection of a model family based on an astrophysical understanding of the conditions under study, and a statistical effort to find parameters for the specified model. A wide variety of issues thus arise:

- Does an observed group of stars (or galaxies or molecular clouds or γ-ray sources) constitute a typical and unbiased sample of the vast underlying population of similar objects?
- When and how should we divide/classify these objects into 2, 3 or more subclasses?
- What is the intrinsic physical relationship between two or more properties of a class of objects, especially when confounding variables or observational selection effects are present?
- How do we answer such questions in the presence of observations with measurement errors and flux limits?
- When is a blip in a spectrum (or image or time series) a real signal rather than a random event from Gaussian (or often Poissonian) noise or confounding variables?
- How do we interpret the vast range of temporally variable objects: periodic signals from rotating stars or orbiting extrasolar planets, stochastic signals from accreting neutron stars or black holes, explosive signals from magnetic reconnection flares or γ-ray bursts?
- How do we model the points in 2, 3, $\ldots$, 6-dimensional points representing photons in an image, galaxies in the Universe, Galactic stars in phase space?
- How do we quantify continuous structures seen in the sky such as the cosmic microwave background, the interstellar and intergalactic gaseous media?
- How do we fit astronomical spectra to highly non-linear astrophysical models based on atomic physics and radiative processes, including confidence limits on the best-fit parameters?

From a superficial examination of the astronomical literature[2], we can show that such questions are very common today. Of $\simeq 15,000$ refereed papers published annually, 1% have “statistics” or “statistical” in their title, 5% have “statistics’ in their abstract, 10% treat time-variable objects, $5-10$% (est.) present or analyze multivariate datasets, and $5-10$% (est.) fit parametric models. Accounting for overlaps, we roughly estimate that around $\simeq 3,000$ distinct studies each year require non-trivial statistical methodologies. Roughly 10% of these are principally involved with statistical methods; indeed, some of these purport to develop new methods or improve on established ones.

3. ASTROSTATISTICS TODAY

We thus find that astronomy and astrophysics today require a vast range of statistical capabilities. In statistical jargon, it helps for astronomers to know something about: sampling theory, survival analysis with censoring and truncation, measurement error models, multivariate classification and analysis, harmonic and autoregressive time series analysis, wavelet analysis, spatial point processes and continuous surfaces, density estimation, linear and non-linear regression, model selection, and bootstrap resampling. In some cases, astronomers need combinations of methodologies that have not yet been fully developed (§7 below).

Faced with such a complex of challenges, mechanical exposure to a wider variety of techniques is a necessary but not sufficient prerequisite for high-quality statistical analyses. Astronomers also need to be imbued with established principles of statistical inference; *e.g.*, hypothesis testing and parameter estimation, nonparametric and parametric inference, Bayesian and frequentist approaches, and the assumptions underlying and applicability conditions for any given statistical method.

Unfortunately, we find that the majority of the thousands of astronomical studies requiring statistical analyses use a very limited set of classical methods. The most common tools used by astronomers are: Fourier transforms for temporal analysis (developed by Fourier in 1807), least squares regression and χ^2 goodness-of-fit (Legendre in 1805, Pearson in 1900, Fisher in 1924), the nonparametric Kolmogorov-Smirnov 1- and 2-sample nonparametric tests (Kolmogorov in 1933), and principal components analysis for multivariate tables (Hotelling in 1936).

Even traditional methods are often misused. Feigelson & Babu [9] found that astronomers use interchangeably up to 6 different fits for bivariate linear least squares regression: ordinary least squares (OLS), inverse regression, orthogonal regression, major axis regression, the OLS mean, and the OLS bisector. Not only did this lead to confusion in comparing studies (*e.g.*, in measuring the expansion of the Universe via Hubble’s constant, H_o), but astronomers did not realize that the confidence intervals on the fitted parameters can not be correctly estimated with standard analytical formulae. Similarly, Protassov et al. [24] found that the majority of astronomical applications of the

[2] Such bibliometric measures are easily accomplished as the entire astronomical research literature is on-line (in full text at subscribing institutions) through the NASA-supported *Astrophysics Data System*, http://adsabs.harvard.edu/abstract_service.html.

F test, or more generally the likelihood ratio test, are inconsistent with asymptotic statistical theory.

But, while the *average* astronomical study is limited to often-improper usage of a limited repertoire of statistical methods, a significant *tail of outliers* are much more sophisticated. The maximization of likelihoods, often developed specially for the problem at hand, is perhaps the most common of these improvements. Bayesian approaches are also becoming increasingly in vogue.

In a number of cases, sometimes buried in technical appendices of observational papers, astronomers independently develop statistical methods. Some of these are rediscoveries of known procedures; for example, Avni et al. [2] and others recovered elements of survival analysis for treatments of left-censored data arising from nondetections of known objects. Some are quite possibly mathematically incorrect; such as various revisions to χ^2 for Poissonian data that assume the resulting statistic still follows the χ^2 distribution. On rare occasions, truly new and correct methods have emerged; for example, astrophysicist Lynden-Bell [19] discovered the maximum-likelihood estimator for a randomly truncated dataset, for which the theoretical validity was later established by statistician Woodroofe [31].

A growing group of astronomers, recognizing the potential for new liaisons with the accomplishments of modern statistics, have promoted astrostatistical innovation through cross-disciplinary meetings and collaborations. Fionn Murtagh, an applied mathematician at Queen's University (Belfast) with long experience in astronomy, and his colleagues have run conferences and authored many useful monographs (*e.g.*, [16], [17], [22] and [27]). We at Penn State have run a series of *Statistical Challenges in Modern Astronomy* meetings with both communities in attendance (*e.g.*, [3] and [10]). Alanna Connors has organized brief statistics sessions at large astronomy meetings, and we have organized brief astronomy sessions at large *Joint Statistical meetings*. We wrote a short volume called *Astrostatistics* [3] intended to familiarize scholars in one discipline with relevant issues in the other discipline. Other series conferences are devoted to technical issues in astronomical data analysis but typically have limited participation by statisticians. These include the dozen *Astronomical Data Analysis Software and Systems* (*e.g.*, [23]), several Erice workshops on *Data Analysis in Astronomy* (*e.g.*, [8]), and the new SPIE *Astronomical Data Analysis* conferences (*e.g.*, [26]).

Most importantly, several powerful astrostatistical research collaborations have emerged. At Harvard University and the Smithsonian Astrophysical Observatory, David van Dyk worked with scientists at the *Chandra*[3] X-ray Center on several issues, particularly Bayesian approaches to parametric modeling of spectra in light of complicated instrumental effects. At Carnegie Mellon University and the University of Pittsburgh, the Pittsburgh Computational Astrophysics group addressed several issues, such as developing powerful techniques for multivariate classification of extremely large datasets and applying nonparametric regression methods to cosmology. Both of these groups involved academics, researchers and graduate students from both fields working closely for several years to achieve a critical mass of cross-disciplinary capabilities.

Other astrostatistical collaborations must be mentioned. David Donoho (Statistics at Stanford University) works with Jeffrey Scargle (NASA Ames Research Center) and others on applying advanced wavelet methods to astronomical problems. James Berger (Statistics at Duke University) has worked with astronomers William Jefferys (University of Texas), Thomas Loredo (Cornell University), and Alanna Connors (Eureka Inc.) on Bayesian methodologies for astronomy. Bradley Efron (Statistics at Stanford University) has worked with astrophysicist Vehé Petrosian (also at Stanford) on survival methods for interpreting γ-ray bursts. Philip Stark (Statistics at University of California, Berkeley) has collaborated with solar physicists in the *GONG* program to improve analysis of oscillations of the Sun (helioseismology). More such collaborations exist in the U.S., Europe and elsewhere.

4. THE VIRTUAL OBSERVATORY: A NEW IMPERATIVE FOR ASTROSTATISTICS

A major new trend is emerging in observational astronomy with the production of huge, uniform, multivariate databases from specialized survey projects and telescopes[4]. But they are heterogeneous in character, reside at widely dispersed locations, and ac-

[3] The *Chandra* X-ray Observatory is one of NASA's Great Observatories. It was launched in 1999 with a total budget around $2 billion.

[4] An enormous collection of catalogs, and some of the underlying imaging and spectral databases, are already available on-line. Access to many catalogs is provided by Vizier (http://vizier.u-strasbg.fr). The NASA Extragalactic Database (NED, http://ned.ipac.caltech.edu), SIMBAD stellar database (http://simbad.u-strasbg.fr), and ADS (footnote 2) give integrated access to many catalogs and bibliographic information. Raw data are available from all U.S. space-based observatories; see, for example, the Multi-mission Archive at Space Telescope (MAST, http://archive.stsci.edu) and High Energy Astrophysics Science Archive Research Center (HEASARC, http://heasarc.gsfc.nasa.gov).

cessed through different database systems. Examples include:

1. $10^8 - 10^9$-object catalogs of stars and stellar extragalactic objects (*i.e.*, quasars). These include the all-sky photographic optical USNO-B1 catalog, the all-sky near-infrared 2MASS catalog, and the wide-field Sloan Digital Sky Survey (SDSS). Five to ten photometric values, each with measured heteroscedastic measurement errors (*i.e.*, different for each data point), are available for each object.

2. $10^5 - 10^6$-galaxy redshift catalogs from the 2-degree Field (2dF) and SDSS spectroscopic surveys. The main goal is characterization of the hierarchical, nonlinear and anisotropic clustering of galaxies in a 3-dimensional space. But the datasets also include spectra for each galaxy each with 10^3 independent measurements.

3. $10^5 - 10^6$-source catalogs from various multiwavelength wide-field surveys such as the NRAO Very Large Array Sky Survey in one radio band, the InfraRed Astronomical Satellite Faint Source catalog in four infrared bands, the Hipparcos and Tycho catalogs of star distances and motions, and the X-ray Multimirror Mission Serendipitous Source Catalogue in several X-ray bands now in progress. These catalogs are typically accompanied by large image libraries.

4. $10^2 - 10^4$-object samples of well-characterized pre-main sequence stars, binary stars, variable stars, pulsars, interstellar clouds and nebulae, nearby galaxies, active galactic nuclei, gamma-ray bursts and so forth. There are dozens of such samples with typically $10 - 20$ catalogued properties and often with accompanying 1-, 2- or 3-dimensional images or spectra.

5. Perhaps the most ambitious of such surveys is the planned Large-aperture Synoptic Survey Telescope (LSST) which will survey much of the entire optical sky every few nights. It is expected to generate raw databases in excess of 10 PBy (petabyte) and catalogs with 10^{10} entries.

An international effort known as the Virtual Observatory (VO) is now underway to coordinate and federate these diverse databases, making them readily accessible to the scientific user [6, 29]. Considerable progress is being made in the establishment of the necessary data and metadata infrastructure and standards, interoperability issues, data mining, and technology demonstration prototype services[5]. But scientific discovery requires more than effective recovery and distribution of information. After the astronomer obtains the data of interest, tools are needed to explore the datasets. How do we identify correlations and anomalies within the datasets? How do we classify the sources to isolate subpopulations of astrophysical interest? How do we use the data to constrain astrophysical interpretation, which often involve highly non-linear parametric functions derived from fields such as physical cosmology, stellar structure or atomic physics? These questions lie under the aegis of statistics.

A particular problem relevant to statistical computing is that, while the speed of CPUs and the capacity of inexpensive hard disks rise rapidly, computer memory capacities grow at a slower pace. Combining the largest optical/near-infrared object catalogs today produces a table with > 1 billion objects and up to a dozen columns of photometric data. Such large datasets effectively preclude use of all standard multivariate statistical packages and visualization tools (*e.g.*, R and GGobi) which are generally designed to place the entire database into computer memory. Even sorting the data to produce quantiles may not be computationally feasible.

The Virtual Observatory of the 21st century thus presents new challenges to statistical capability in two ways. First, some new methodological developments are needed (§5). Second, efficient access to both new and well-established statistical methods are needed. No single existing software package can provide the vast range of needed methods. We are now involved in developing a prototype system called *VOStat* to provide statistical capabilities to the VO astronomer. It is based on concepts of Web services and distributed Grid computing. Here, the statistical software and computational resources, as well as the underlying empirical databases, may have heterogeneous structures and can reside at distant locations.

5. SOME GRAND METHODOLOGICAL CHALLENGES FOR THE COMING DECADE

While it is risky to prognosticate the directions of future research, and judgments will always differ regarding the relative importance of research goals, we can outline a few "grand challenges" for astrostatistical research for the next decade or two.

5.1. Multivariate analysis with measurement errors and censoring

Traditional multivariate analysis is designed mainly for applications in the social and human sciences where the sources of variance are largely unknowable.

[5] See http://www.ivoa.net and /http://us-vo.org for entry into Virtual Observatory projects.

Measurement errors are usually ignored, or are considered to be exogenous variables in the parametric models [12]. But astrophysicists often devote as much effort to precise determination of their errors as they devote to the measurements of the quantities of interest. The instruments are carefully calibrated to reduce systematic uncertainties, and background levels and random fluctuations are carefully evaluated to determine random errors. Except in the simple case of bivariate regression [1, 5, 9], this information on measurement errors is usually squandered.

While heteroscedastic measurement errors with known variances is common in all physical sciences, only astronomy frequently has nondetections when observations are made at new wavelengths of known objects. These are datapoints where the signal lies below (say) 3 times the noise level. Here again, modern statistics has insufficient tools. Survival analysis for censored data assumes that the value below which the data point must lie is known with infinite precision, rather than being generated from a distribution of noise. Astronomer Herman Marshall [20] makes an interesting attempt to synthesize measurement errors and nondetections, but statistician Leon Gleser [14] argues that he has only recovered Fisher's failed theory of fiducial distributions. Addressing this issue in a self-consistent statistical theory is a profound challenge that lies at the heart of interpreting the data astronomers obtain at the telescope.

5.2. Statistical inference and visualization with very-large-N datasets

The need for computational software for extremely large databases – multi-terabyte image and spectrum libraries and multi-billion object catalogs – is discussed in section 4. A suite of approximate methods based on flowing data streams or adaptive sampling of large datasets resident on hard disks should be sought. Visualization methods involving smoothing, multidimensional shading and variable transparency, should be brought into the astronomer's toolbox. Here, considerable work is being conducted by computer scientists and applied mathematicians in other applied fields so that independent development by astrostatisticians might not be necessary to achieve certain goals.

5.3. A cookbook for construction of likelihoods and Bayesian computation

While the concepts of likelihoods and their applications in maximum likelihood estimation, Bayes Theorem and Bayes factors are becoming increasingly well-known in astronomical research, the applications to real-life problems is still an art for the expert rather than a tool for the masses. Part of the problem is conceptual; astronomers need training in how to construct likelihoods for familiar parametric situations (*e.g.*, power law distributions or a Poisson process). Part of the problem is computational; astronomers need methods and software for the oft-complex computations. Many such methods, such as Markov chain Monte Carlo, are already well-established and can be directly adopted for astronomy [13]. For example, astronomers are often not fully aware of the broad applicability of the EM Algorithm for maximizing likelihoods [21][6].

5.4. Links between astrophysical theory and wavelets

Wavelet analysis has become a powerful and sophisticated tool for the study of features in data. Originally intended mainly for modelling time series, astronomers also use it increasingly for spatial analysis of images [11, 25]. In some ways it can be viewed as a generalization of Fourier analysis in which the basis function need not be sinusoidal in shape and, most importantly, the pattern need not extend over the entire dataset. Wavelets are thus effective in quantitatively describing complicated overlapping structures on many scales, and can also be used for signal denoising and compression. In addition, wavelets have a strong mathematical foundation.

Despite its increasing popularity in astronomical applications, wavelet analysis suffers a profound limitation in comparison with Fourier analysis. A peak in a Fourier spectrum is immediately interpretable as a vibrational, rotational or orbital rotation of solid bodies. A bump or a continuum slope in a wavelet decomposition often has no analogous physically intuitive interpretation. We therefore recommend that astrophysicists seek links between physical theory – often involving continuous media such as turbulent plasmas in the interstellar medium and hierarchical structure formation in the early Universe – and wavelets. One fascinating example is the demonstration that the wavelet spectrum and Lyapunov exponent of the quasi-periodic X-ray emission from Sco X-1, which reflects the processes in an accretion disk around a neutron star, exhibit a transient chaotic behavior similar to that of water condensing and dripping onto an automobile windshield or a dripping handrail [32].

[6] The seminal study of the EM Algorithm is Dempster, Laird & Rubin in 1977 [7], which is one of the most frequently cited papers in statistics. However, the method was independently derived three years earlier by astronomer Leon Lucy [18] as an "iterative technique for the rectification of observed distributions" based on Bayes' Theorem. This study is widely cited in the astronomical literature; its most frequent application is in image deconvolution where it is known as the Lucy-Richardson algorithm.

5.5. Time series models for astrophysical phenomena

The quasi-periodic oscillation of Sco X-1 is only one of many examples of complex accretional behavior onto neutron stars and black holes seen in X-ray and γ-ray astronomy. The accreting Galactic black hole GRS 1915+105 exhibits a bewildering variety of distinct states of stochastic, quasi-periodic and explosive behaviors. The prompt emission from gamma-ray bursts shows a fantastic diversity of temporal behaviors from simple smooth fast-rise-exponential-decays to stochastic spiky profiles. Violent magnetic reconnection flares on the surfaces of the Sun and other magnetically active stars also show complex behaviors. Many of these datasets are multivariate with time series available in several spectral bands often showing lags or hardness ratio variations of astrophysical interest.

There are also important astronomical endeavors which seek astrophysically interesting signals amidst the oft-complex noise characteristics of the detectors. The Arecibo, Parkes and VLA radio telescopes, for example, conduct searches for new radio pulsars or for extraterrestrial intelligences in nearby planetary systems. The Laser Interferometer Gravitational-Wave Observatory (LIGO) and related detectors search for both continuing periodic signals and brief bursts from perturbations in space-time predicted by Einstein's General Relativity. Here the signals sought are orders of magnitude fainter than instrumental variations.

6. INFRASTRUCTURE NEEDED TO ADVANCE ASTROSTATISTICS

The current quality of statistical analyses in astronomical research often begs for improvment. There is both inadequate research on important new challenges (§5) and inadequate application of known advanced methods to astronomical problems (§3). Astronomy clearly needs needs a strong and rapid surge of energy in statistical expertise. Three types of activities should be promoted:

Cross-training In the U.S., the typical curriculum leading to a career in astronomical research requires zero or one course in statistics at the undergraduate level, and zero at the graduate level. Analogously, the curriculum of statisticians includes virtually no coursework in astronomy or other physical science. While statisticians can learn basics from "Astronomy 101" courses given at all universities, the statistical training of astronomers is not as easily accomplished. New curricular products summarizing the applicable statistical subfields, short training workshops for graduate students and young scientists, and effective statistical consulting are all needed.

Increased collaborative research While several astrostatistical research groups are making exciting progress (§3), the total effort is too small to impact the bulk of astronomical research. Very roughly, astrostatistical funding is currently \$1M of the \$1B spent annually on astronomical research. This fraction is far below that spent in biomedical or other non-physical-science fields. Though top academic leaders of statistics have expressed great enthusiasm for astronomy and astrostatistics, we can not pull them away from biostatistics and business applications without a major increase in funding. We might seek, for example, $10 - 20$ cross-disciplinary research groups active at any one time at the end of a decade's growth.

Statistical software For various policy and cultural reasons, astronomers rarely purchase the large commercial statistical software packages, preferring to write their own software as needs arise. This approach has contributed to the narrow methodological scope of astronomical research. Avenues for improving this situation are emerging. *R* is a large statistical software package with the flexible command-line interface preferred by astronomers that has recently emerged (http://www.r-project.org). A wide variety of specialized packages and codes are also available on-line (http://www.astro.psu.edu/statcodes). The new Web services concept being developed within the context of a Virtual Observatory permits coordinated access to heterogeneous software developed specifically for astronomical applications.

At Penn State, we are in the early stages of developing a Center for Astrostatistics to help attain these goals (http://www.astrostatistics.psu.edu). This is an inter-disciplinary Center to serve the astronomy and statistics communities around the nation and worldwide, seeking to bring advances in statistics into the toolbox of astronomy and astrophysics. The Center's Web site will maintain the popular *StatCodes*, build an instructional library of *R* programs, coordinate with the nascent *VOStat* Web service, and develop an archive of annotated links to selected statistical literature applicable to astronomy (and vice versa). The site is also planned to include tutorial handbooks and curricular products developed specifically for astrostatistical needs.

Acknowledgments

We thank the referee, Peter Shawhan (Caltech), for very helpful improvements to the paper. This work was supported by NSF grant DMS-0101360.

References

[1] Akritas, M. G. & Bershady, M. A. 1996, Astrophys. J. 470, 709

[2] Avni, Y., Soltan, A., Tananbaum, H., & Zamorani, G. 1980, Astrophys. J. 238, 800

[3] Babu, G. J. & Feigelson, E. D. 1996, Astrostatistics, London, Chapman & Hall

[4] G. J. Babu & Feigelson, E. D. 1997, Statistical Challenges in Modern Astronomy II, New York, Springer

[5] Boggs, P. T., Byrd, R. H. & Schnabel, R. B. 1987, SIAM J. Sci. Statist. Comput. 8, 1052

[6] Brunner, R. J., Djorgovski, S. G. & Szalay, A. S. (eds.) 2001, Virtual Observatories of the Future, San Francisco, Astron. Soc. Pacific

[7] Dempster, A. P., Laird, N. M. & Rubin, D. B. 1977, J. Roy. Statist. Soc. Ser. B, 39, 1

[8] Di Gesu, V., Duff, M. J. B., Heck, A., Maccarone, M. C., Scarsi, L., & Zimmerman, H. U. 1997, Data Analysis in Astronomy IV,

[9] Feigelson, E. D. & Babu, G. J. 1992, Astrophys. J. 397, 55

[10] Feigelson, E. D. & Babu, G. J. 2003, Statistical Challenges in Modern Astronomy, New York, Springer

[11] Freeman, P. E., Kashyap, V., Rosner, R., & Lamb, D. Q. 2002, Astrophys. J. Suppl., 138, 185

[12] Fuller, W. A. 1987, Measurement Error Models, New York, Wiley

[13] Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. 1995, Bayesian Data Analysis, London, Chapman & Hall

[14] Gleser, L. J. 1992, in Statistical Challenges in Modern Astronomy (E. D. Feigelson & G. J. Babu, eds.), New York, Springer, 263

[15] Hald, A. 1990, A History of Probability and Statistics and Their Applications before 1750, New York, Wiley

[16] Heck, A. & Murtagh, F. 1993, Intelligent Information Retrieval: The Case of Astronomy and Related Space Sciences, Dordrecht, Kluwer

[17] Jaschek, C. & Murtagh, F. 1990, Errors, Bias and Uncertainties in Astronomy, Cambridge, Cambridge Univ. Press

[18] Lucy, L. B. 1974, Astron. J. 79, 745

[19] Lynden-Bell, D. 1971, Mon. Not. Roy. Astro. Soc. 155, 95

[20] Marshall, H. L 1992, in Statistical Challenges in Modern Astronomy (E. D. Feigelson & G. J. Babu, eds.), New York, Springer, 247

[21] McLachlan, G. J. & Krishnan, T. 1997, The EM Algorithm and Extensions, New York, Wiley

[22] Murtagh, F. & Heck, A. 1987, Multivariate Data Analysis, Dordrecht, Reidel

[23] Payne, H. E., Jedrzejewski, R. I., & Hook, R. N. 2003, Astronomical Data Analysis Software and Systems, San Francisco, Astro. Soc. Pacific

[24] Protassov, R., van Dyk, D. A., Connors, A., Kashyap, V. L., & Siemiginowska, A. 2002, Astrophys. J., 571, 545

[25] Scargle, J. D. 1998, Astrophys. J., 504, 405

[26] Starck, J.L. & Murtagh, F. (eds.) 2002, Astronomical Data Analysis II, Proc. SPIE, vol 4847

[27] Starck, J.L. & Murtagh, F. 2003, Astronomical Image and Data Analysis, New York, Springer

[28] Stigler, S. M. 1986, *The History of Statistics: The Measurement of Uncertainty before 1900*, Cambridge, Harvard Univ. Press

[29] Szalay, A. & Gray, J. 2001, Science, 293, 2037

[30] Trumpler, R. J. & Weaver, H. A., 1953, *Statistical Astronomy*, Berkeley, Univ. of California Press

[31] Woodroofe, M. 1985, Annals Statist. 13, 163

[32] Young, K. & Scargle, J. D. 1996, Astrophys. J., 468, 617

Introduction to Statistical Issues in Particle Physics

Roger Barlow
Manchester University, UK and Stanford University, USA

An account is given of the methods of working of Experimental High Energy Particle Physics, from the viewpoint of statisticians and others unfamiliar with the field. Current statistical problems, techniques, and hot topics are introduced and discussed.

1. PARTICLE PHYSICS

1.1. The Subject

Particle Physics emerged as a discipline in its own right half a century ago. It pioneered 'big science'; experiments are performed at accelerators of increasing energy and complexity by collaborations of many physicists from many institutes. It has evolved a research methodology within which statistics is of great importance, although it has done so without strong links to the statistics community – a fault that this conference exists to remedy. Thus although a statistician will be familiar with the research methods and statistical issues arising in, say, agricultural field trials or clinical testing, they may be interested in a brief description of how particle physicists do research, and the statistical issues that arise.

Particle physics is also known as High Energy Physics [1] and the names are sometimes merged to give High Energy Particle Physics. Whatever it is called, its field of study is all the 'Elementary' Particles that have been discovered:

- The 6 quarks (u, d, s, c, b, t)
- The 6 leptons (e, μ, τ, ν_e, ν_μ, ν_τ)
- The intermediate bosons: W, Z, γ, g
- The 100+ hadrons made from two quarks(π, K, $D_s(2317)$...) or three quarks (p,n, Λ...) or five quarks (Θ^+...)

To this long list must be added the corresponding list of antiparticles. However this is not all: the domain of particle physics also includes all the particles that have not yet been discovered – some of which never will be discovered:

- Higgs boson(s)
- squarks and sleptons
- Winos and Zinos/charginos and neutralinos
- further hadrons
- etcetera etcetera...

This list of proposed particles is limited only by the imagination of the theorists who propose them – which is no limitation at all.

For each species of particle we want to establish:

- Does it exist?
- If it does exist, what are its properties: its mass, its lifetime, its charge, magnetic moment and so on?
- If its lifetime is not infinite, what particles does it decay into? What are the branching fractions to different decay modes? What are the distributions in the parameters (energies and directions) of the particles in the final state? Do they agree with our theoretical models?
- What happens when it collides with another particle? What processes can occur and with what probabilities (expressed as cross sections)? What are the distributions of the parameters of the particles produced? Answers will depend on the target particle and the collision energy.

1.2. Template for an Experiment

To study some phenomenon X, which could be any of the above, a particle physics experiment goes through the following stages:

- Arrange for instances of X

  This may involve a beam of particles, directly from an accelerator or through some secondary system, striking a target; the beam and target particles and the energy are chosen as being favorable for X. It may entail a colliding beam machine like LEP for the Z or *BABAR* for CP violation in the B system or the LHC for the Higgs. It may be done by producing particles and then letting them decay, as in the studies of CP violation in the K^0 system. An extreme example is proton lifetime studies, where one just assembles

[1] The terms are almost equivalent; strictly the phrase 'High Energy' means 'above the threshold for pion production', i.e., the energy at which a collision between two protons can produce three outgoing particles.

a large number of ordinary protons (perhaps as hydrogen in water) in suitably low-background conditions deep underground and waits to observe any decays.

For important studies dedicated experiments (even accelerators) are built. For many more, the experimenter utilizes data taken with an experiment designed primarily for another purpose but also favorable for X. An example is the study of charm mesons at *BABAR*, Belle and CLEO, for which the primary purpose is B physics.

- Record events that might be X

  A detector is built (or an existing detector is utilized). 'Events' – interactions or decays – are observed by a whole range of detectors (tracking detectors like drift chambers and silicon detectors, calorimeters that measure deposited energy). Fast logic and/or online computers select the events that look promising, and these are recorded: the phrase 'written to tape' is used even though today the recording medium is generally disk storage.

- Reconstruct the measurable quantities of the visible particles.

  The electronic signals are combined and interpreted: points are joined to form tracks, and measurement of their curvature in a magnetic field gives the particle momentum. A calorimeter may give the energy, a Cherenkov counter the velocity. From this emerges a reconstructed 'event' as a list of the particles produced, their kinematic quantities (energies and directions) and possibly their identity (as pions or kaons or electrons, etc.)

- Select events that could be X by applying *cuts*

  Knowing the pattern one is looking for, one can then select the events that contain the phenomenon being studied.

  A key point is that this selection (and also the electronic selection described above) is not going to be perfect. There will always be a selection *efficiency* which is less than 100%.

  There is also a chance that some of the events that look like X and survive the selection and the cuts are actually from some other process. There will be a *background* which is greater than zero. Statistical techniques are obviously important for the treatment and understanding of efficiency and background.

- Histogram distributions of interesting variables

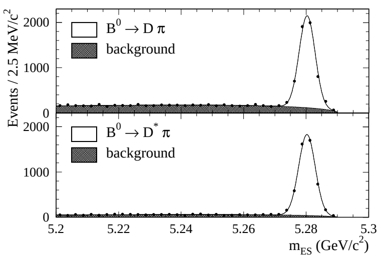


Figure 1: Examples of analysis: B^0 decay to $D\pi$ and $D^*\pi$

  Relevant quantities, sensitive to X, are formed from the kinematic variables of the particles detected and measured. These are typically displayed in a histogram, or histograms. (Joint two-dimensional plots are also common. Sometimes, but rarely, the data at this stage is a single number.)

  These distributions are then compared with the theoretical predictions, of which there may be several. One will be the predicted distribution if X is not present. Another may be the prediction if X is present in the amount, and with the properties, predicted by an expected theory such as the 'Standard Model' [1] of Particle Physics. There may also be predictions obtained within the framework of a particular model, but with one or more parameters adjusted to fit the data.

An example of such a result is shown in Figure 1 (taken from [2]). In the top plot the phenomenon X is the decay of the B^0 meson to $D\pi$, in the lower plot the decay to $D^*\pi$. The distributions show the invariant mass, which is the quantity given by

$$M^2c^4 = \left(\sum_i E_i\right)^2 - \left(\sum_i \vec{p}_i c\right)^2 \qquad (1)$$

where the sums run over the two final-state particles. If the two observed particles do indeed come from the decay of a B^0 particle then this quantity should be 5.28 GeV/c^2, though this is smeared out by experimental resolution. The plots show the predictions of a theory in which this decay does not occur (and all events are background) and also a prediction in which the decay is produced, with a normalization adjusted to give the best fit to the data. The result of this fit gives the number of signal events, from which the branching ratio can be obtained (though in fact that was not done in this example).

If that looks trivial, a harder example is the decay $B \to \pi^0\pi^0$, taken from [3] and shown in Figure 2. (To be fair, things are not quite as bad as this 1 dimensional plot implies.)

In this confrontation of theory with experiment, one can then ask: is there any evidence for X or is the null

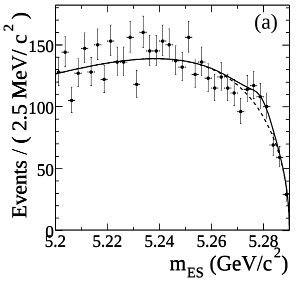


Figure 2: Another example of an analysis: $B^0 \to \pi^0\pi^0$

hypothesis unrefuted? Given that there is X, what is the best estimate for the normalization (and perhaps other) parameter(s) involved in our model for X? Are these results compatible with the standard prediction for X? These are familiar statistical questions.

1.3. Statistics in HEP

From the above description we can bring out some features of the way statistics is used in HEP.

Firstly, everything is a counting experiment. To measure a branching ratio or a cross section, one counts the number of events produced and observed. To measure the mass of a particle one uses a histogram where the number of entries in each bin is a random Poisson process. (The data of Figure 1 could be used to fit the mass of the B^0 meson, were it not already well known.) Poisson statistics is of paramount importance. Even the Gaussian (Normal) distribution plays its main rôle as the large N limit of the Poisson. (There are exceptions to this generalization, but they occur in the details of the reconstruction of particle quantities.)

This unpredictability is not due to any lack of knowledge on our part: not sampling error, or measurement error, or due to unconsidered effects. It is true and absolute randomness, driven by the fundamental nature of quantum mechanics. We know that, for instance, a K_s^0 particle may decay into two charged pions or two neutral pions, with probabilities of 69% and 31% respectively. That is all we can ever know. A sample of K^0 particles will decay to $\pi^+\pi^-$ and $\pi^0\pi^0$ in a ratio of roughly 2:1 even if they are prepared absolutely identically – we have no hope of ever being able to say which ones are more likely to ‘choose’ one path rather than another. Likewise the timing of a decay is absolutely random in that the probability that a particle existing at time t will decay before time $t + \delta t$ is a constant, independent of the value of t; there is no ‘ageing’ process.

But the Poisson distributions that result are just like any conventional Poisson process. These and other uncertainties, are (almost always) controlled and understood. These distributions have standard deviations known to be $\sqrt{N}$. The Gaussian used for the signal distributions in Figure 1 is well established (it has a mean of 5.28 and a standard deviation of 0.0025 GeV/c^2).

So, in common with the other physical sciences, the distributions involved (signal, backgrounds) are given by functions known up to a few parameters – which can be fitted for. The approach to the data is not descriptive (identifying features, looking for trends) but prescriptive: the distribution is taken as having some functional form, and one has a pretty good idea as to what that functional form is, apart (possibly) from a few adjustable parameters.

1.4. Unused Statistical Methods

A consequence of this knowledge of uncertainty – the fact that we know what it is that we don’t know – is that many techniques commonly used in the broad field of statistics are little used (or not used at all) in particle physics.

Student’s t is unknown. This is a technique used to handle small numbers of values from a distribution of unknown mean and unknown standard deviation, but our uncertainties come from known measurement errors. (If a measurement error is not known, a separate large-number determination is made.) The F test and ANOVA, tools for studying problems with unknown variances, are similarly of little use. The whole experimental design field – Latin squares and similar techniques used to minimize uncontrollable effects – is not needed as such effects are not a problem.

Another set of neglected techniques are those handling Time Series and Markov chains. Changes with time can be relevant in some studies, but it appears in them as another quantity to be measured and histogrammed. The development with time of a particle is basically smooth, punctuated by radical transformations (such as the decay of a particle to two or more lighter ones) which occur at random times.

Non-parametric Statistics are also barely featured, as all these distributions, which are believed to be true idealizations of what ‘really’ happens, or at least good approximations to them, are parametrised.

The notion of a Parent population is not helpful: a sample of particles is taken, but the randomness is (as stated earlier) inherent in the nature of particle behavior and not produced by the sampling. If there is a parent distribution, it is an infinite set of particles produced under these conditions – all the events we might have seen.

The point here is not that particle physics has nothing to learn from standard statistical techniques. The Statistician has many implements in their toolbox. Different fields of application will call for different tools; some of those heavily used in other fields are of less relevance in this one.

2. TOOLS

Having seen that particle physics makes little use of some statistical tools, we take a more detailed look at the ones it does utilize.

2.1. Monte Carlo Simulation

Theoretical distributions for the quantities being studied are predicted by quantum mechanics – perhaps with a few unknown parameters – and are often beautiful and simple. Angular distributions may be flat, or described by a few trigonometric terms; masses often follow a Cauchy function (which the particle physicists call the Breit-Wigner), time distributions may be exponential, or exponential with a sinusoidal oscillation.

These beautiful and simple forms are generally modified by unbeautiful and complicated effects (higher-order calculations in perturbation theory, or the fragmentation of quarks into other particles). Furthermore the measurement and reconstruction process that the detector does for the particles is not completely accurate or completely efficient.

The translation from knowing the distributions in principle to knowing them in practice is done by Monte Carlo simulation. Particles are generated according to the original simple distributions, and then put through repeated random processes to describe the theoretical complications and then the passage of particles through the detector, including probabilities for colliding with nuclei in the beam pipe, slipping through cracks in the acceptance, or other eventualities. A complete software representation of all the experimental hardware has to be coded. The effects of the particles on the detector elements is simulated and the information used to reconstruct the kinematic quantities using the same programs that are run on the real data. This provides the full theoretical distribution function that the data is predicted to follow, albeit as a histogram rather than a smooth curve.

These programs are large and slow to run. Significant resources (both people and machines) are put into them. The generation of 'Monte Carlo data' is a significant issue for all experiments. Cases are known where data has been taken and analyzed but results delayed because of lack of the correct Monte Carlo data [4].

2.2. The Likelihood

Having the parametrized theoretical description of the distribution means the likelihood function is always known, and it assumes an overwhelmingly important position. Writing this function – where the x_i are the data and θ the unknown parameter(s)

$$L(x_1, x_2 \dots x_N|\theta) = \prod P(x_i|\theta) \qquad \ln L = \sum_i \ln P(x_i|\theta)$$

the form $p(x|\theta)$ is totally known, and L (or $\ln L$) follows.

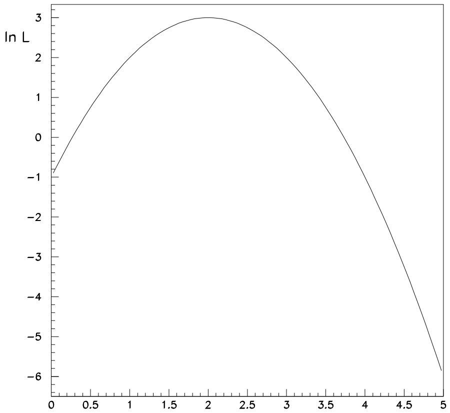


Figure 3: The log likelihood as a function of a parameter

Having the likelihood function, the Maximum Likelihood estimator is then easy to implement, and is very widely used. Even estimators like least-squares are, at least by some, 'justified' as being derivable from Maximum Likelihood. Its (asymptotic) efficiency, and its invariance properties are desirable and useful.

In some cases the ML estimate leads to an algebraic solution but in general, and in complex analysis, the physicist just maps out $\ln L$ for their data set as a function of θ and reads off the ML estimator from the peak, as can be done in Figure 3. This also produces an interval estimate as part of the minimization process. Following the value of $\ln L$ until it falls off by $\frac{1}{2}$ from its maximum gives the 68% central confidence interval. Strictly speaking this is valid only for large N, but this restriction is generally disregarded. Perhaps we should not be so cavalier about doing so.

Maximum Likelihood methods can also be used for functions with several parameters, as illustrated in Figure 4. Confidence regions are mapped out by reading off the likelihood contours. This is done in many analysis and the MINUIT program [5] is widely used in exploring the likelihood and parameter space.

2.3. Fitting Data

Fitting the parametrized curve to the experimental data is done by several techniques.

1) χ^2 using $\sigma^2 = n$ i.e., minimizing $\chi^2 = \sum_i \frac{(y_i - f(x_i|\theta))^2}{n}$ has the advantage that the minimization can be done by differentiating and solving the normal equation, which is especially simple if f is linear in θ. However the use of the observed number rather than the predicted number in the denominator

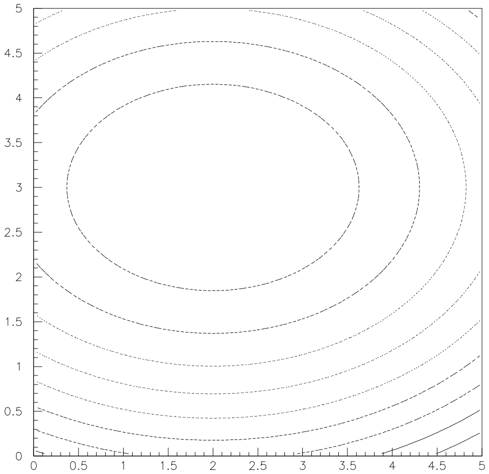


Figure 4: Contours of $\ln L$ in two dimensions

is recognized to lead to bias (downward fluctuations get an undue weight) and this cannot safely be used if n is small. (Actually in many cases what happens is that one of the bins has $n = 0$, and the physicist gets divide-by-zero messages and then starts to worry.)

2) χ^2 using $\sigma^2 = f$ i.e., the predicted value rather than the actual number, avoids the bias (and the divide-by-zero problem) but gives nonlinear equations. It still suffers from using a Gaussian probability as an approximation to a Poisson distribution and is thus not the 'real' maximum likelihood estimator.

3) 'Binned Maximum Likelihood' uses the Poisson likelihood in each bin rather than the χ^2. It is therefore a proper Maximum Likelihood estimator. Efficiency is lost (only) if the bins are wider than the structure of the data.

4) Full maximum likelihood does not use binning at all. It can be useful for very small event samples. For large samples it becomes computationally intensive (as there is a sum over events rather than a sum over bins) though with today's computers this is hardly important. Perhaps a more significant factor for physicists is that it does not have the readily interpretable graphic image given by a histogram and fitted curve.

2.4. Goodness of Fit

Having found a fit, one has to judge whether to believe it. Whether the question is 'Does the curve really describe the data?' or 'Do the data really fit the curve' depends on one's point of view.

The likelihood value does not contain the answer to this question. This appears counter-intuitive and many people have wrestled (unsuccessfully) to produce ways that the likelihood can be used to say something about the quality of the fit.

The $\chi^2 = \sum_i \left(\frac{y_i - f(x_i|\theta)}{\sigma_i}\right)^2$ certainly does give a goodness of fit number. It is heavily used for GoF and 2-sample tests: researchers may quote χ^2 or χ^2/N_D or the probability of exceeding this χ^2.

Alternative measures of goodness of fit have never really caught on. The Kolmogorov-Smirnov test is occasionally used – generally misleadingly, in my opinion. This is a totally robust test but pays the price for that by being weak. If you know anything about the data, e.g., that the numerical value of the parameter means something, then a more powerful test should be available. The KS test is being used to certify that distributions are in agreement when a more powerful approach would show up a difference.

2.5. Toy Monte Carlo

The 'Toy Monte Carlo' has emerged as a technique made possible by modern computing resources. Having obtained a result, it may be hard or impossible to obtain significance levels or confidence regions in the traditional analytic way, for instance if the likelihood function one is studying does not even plausibly resemble a distorted parabola, but instead some shape with multiple maxima.

As an alternative approach, starting with an estimate $\hat{\theta}_{exp}$ from the data, say $\{x_1 \ldots x_N\}$, how can one establish a confidence region? Consider any particular θ. Use the known $L(x|\theta)$ to generate a set of N values of x – an "experiment". Use this in your estimator (whatever that is) to find a corresponding $\hat{\theta}$. Repeating many times gives the probability that this θ will give an estimate below (or above) the experimental one. This is just what the Neyman construction uses. To find a particular confidence region one has to explore the parameter space until one finds the limits one wants.

3. TOPICS

Having explained the basic and generally agreed techniques used, there are a number of topics where advances are being made, or which are the subject of heated discussion and argument, or both.

3.1. Bayesian Probability

The religious war which has been waged over the past few years has now cooled – although some isolated zealots remain on both sides. The 'frequentists' have come to accept that the use of Bayesian techniques can be illuminating and helpful, and sometimes provide more useful information than a frequentist confidence level, especially for measurements of bounded parameters (e.g., masses). The 'Bayesians' are recognizing that Bayesian confidence levels will not *totally* replace the use of frequentist levels, and that they do have to take on board the issue of robustness (or otherwise) under changes of prior.

A real benefit of this debate has been to bring the subject out into the open. The classic statistics texts [6, 7], from which many particle physicists first learned the subject, slide swiftly between the frequentist and Bayesian concepts of probability, never really acknowledging that they are using two very different quantitites.

3.2. Small Signals and Confidence Regions

The 'Energy Frontier' is a cutting edge of particle physics: new, more powerful, accelerators open up new areas for investigation and new particles are discovered. Another cutting edge is the 'Luminosity Frontier': the discovery that processes hitherto thought to be impossible do actually occur, albeit very rarely. The discovery of CP violation [8]: that the probability of the decay $K_L^0 \to \pi^+\pi^-$ was not zero but 0.2%, was enormously important despite the smallness of the figure. Many of today's experiments are looking for phenomena which are known to be exceedingly rare, at the parts-per-million level at best.

Although the implications can be spelt out quite simply and dramatically – 'If the AMS experiment sees even one $\overline{^{12}C}$ nucleus, our entire view of the universe will change.'– in practice things are not so clear-cut because of the presence of background. Also one has to be able to handle not just the dramatic discoveries, but the much more frequent useful analysis that make no discovery but push back the limits and the region in which any discovery may be made.

An experiment that sees no events will note the standard result from Poisson statistics that an observed number of zero translates to a limit on the true value of less than 3 events, with 95% confidence. This can then be converted (using the figures for this particular experiment) into a limit on the branching ratio or cross section for the process concerned, and then possibly into a limit on a mass or coupling constant. If there is an expected background for this process equivalent to, say, 0.2 events, then the amount for the branching ratio limit is reduced to 2.8. But this clearly has problems: suppose the predicted background were 3.1 and no events were observed (unlikely but not impossible), what can one then say about the limit?

There has been a lot of activity and discussion recently in this area. Indeed it sparked off the workshop [9] of which this conference is the successor. The standard frequentist (Neyman) construction may result in statements about results in the non-physical region (here, a negative number of signal events) which, though statistically correct, appear nonsensical. Bayesian methods avoid this problem, as does the frequentist technique proposed by Feldman Cousins [10] which switches smoothly and automatically between quoting central and one-sided confidence regions.

3.3. When to Claim a Discovery?

Another area of discussion is over the form of reporting non-zero signals. When the number of signal events is much larger than the expected background, or a fitted parameter is significantly different from the theoretical prediction, then clearly the experiment can claim a discovery. If the numbers or parameter values are compatible, the experiment quotes an upper limit. But there is an area in between where the probability of the null hypothesis giving the result is small enough to be interesting, but not so small as to be completely negligible. The experiment must not be rash, phoning the New York Times with a discovery which turns out to be a statistical fluctuation, nor must it be too cautious or the subject can never progress. Such results are bound to occur – the probability that an experiment will produce a value in this region is by definition small-but-not-negligible, or better. Given the large number of busy experiments reporting results, this is a real problem.

Some experiments have policies such as 4σ for 'evidence for', 5σ for 'discovery of' – significance levels are often presented in terms of the equivalent discrepancy in standard deviations. Is it possible to report a two-sided result (as the Feldman Cousins technique will sometimes produce) and yet not claim a discovery? 'We report with 95% confidence that the branching ratio lies in the range (2.3 to 3.4) 10^{-6} but we're not actually claiming to have seen it.' Such 'discoveries' are reported in a way which must be affected by the prior (subjective) probability, in exactly the way the Bayesians describe. Statistically identical data on the decays $B^+ \to \pi^+\pi^0$ and $B^+ \to \pi^-\pi^0$ would be reported completely differently.

3.4. Blind Analysis

In recent years particle physicists have become aware of practitioner bias. This has been fuelled particularly by reports from the Particle Data Group, which has the job of reporting and combining the measurements of particle properties [11], who show how some values change significantly over time, but never by more than one standard deviation. Another source of disquiet was the Electroweak measurements from LEP and the SLC which agree with each other and with the Standard Model far *too* well with a χ^2 per degree of freedom well below 1 [12].

This practitioner bias is *against* claiming differences from the null hypohesis. The experiment template presented in section 1.2 often continues

- Extract result, usually by fitting parametrized distribution(s) to data.
- Compare your result with that of accepted theory and/or other experiments.
- If it disagrees, look for a bug in your analysis. You will probably find one. Keep searching and fixing until the agreement is acceptable.

The mistake in method is that the experimenter stops looking for bugs when they have agreement, not when they honestly believe that all (substantial) biases are accounted for. To guard against this the data can be 'blinded'. There are two techniques used, covering two types of situation

- In the extraction of a result, this can be encoded by some unknown offset.
- Choosing the cuts which select the data is done on Monte Carlo data, or on real data in sidebands – regions close to but not actually including the region where the signal is expected. Otherwise the temptation to nudge a cut slightly to include a few more events is too great.

3.5. Systematic Errors

In the early days of particle physics, the 50s and 60s, a typical experiment would get handfuls of events – a few hundred if lucky – from painstaking analysis of bubble chamber pictures. Statistical errors were thus $\sim 10\%$ and were so large that the effect of systematic uncertainties was generally small.

In the 70s and 80s, the development of counter experiments led to event samples in the tens of thousands. Statistical errors were now at the per cent level, and systematic errors began to be more important.

The current generation of experiments – the Z factory at LEP, the B factories, Deep Inelastic Scattering at HERA – deal with millions of events. Statistical errors are at the level of $\sim 0.1\%$ and we have learned how to talk about 'parts per mille'.

Systematic errors (uncertainties in factors systematically applied in the analysis) can no longer be fudged. The word 'conservative' has been grossly overused in this context. It sounds safe and reassuring; in practice it is usually a sign of laziness or cowardice. The experiment perhaps cannot be bothered to evaluate an uncertainty and makes a guess, and then it inflates that guess to cover the possibility that they'll be caught out, and calls it a 'conservative' estimate of the systematic error.

Particle physicists also confuse the evaluation of systematic errors with overall consistency checks. There is bad practice being spread to and between graduate students. They will identify all the calibration constants and parameters that contribute to the final result and vary those by their appropriate error, and fold the resultant variation into the systematic error. This is correct procedure. But they will also vary quantities like cut values, which should not in principle affect the result, by some arbitrary amount and then solemnly fold those resulting variations into the systematic error. This is nonsense. Looking at what happens when you change a cut value is a good and sensible thing: a (say) looser cut will give a higher efficiency and a higher background and thus more observed events, but after correcting for the new efficiency and background the result should be compatible with the original. This is a useful check that one understands what's going on and that the analysis is consistent. But it does not feed into a numerical uncertainty.

3.6. Unfolding

Measurements of the properties of particles in events are made with finite resolution, so the plots of these quantities, and functions of these quantities, are 'smeared out'. Events move between histogram bins. Sharp peaks become broad, edges become slopes.

The recovery of the original sharp distribution from the observed one is known as 'unfolding'. This is an alternative use of the Monte Carlo simulation process: rather than compare the data with a theoretical prediction smeared by Monte Carlo simulation, one compares the original theory with the de-smeared data. Clearly this is preferable, if it can be done, as the unfolding process depends only on the experiment and not on the original theory, and so once unfolded the data can be compared with any prediction.

It looks to be a simple problem: given an original distribution as a histogram, the probability of migration from any bin i to any bin j, P_{ji}, can be estimated from a Monte Carlo sample (this includes the probability that it may not be accepted: $\sum_j P_{ji} \leq 1$). The matrix is inverted, and then applied to the data histogram to give the reconstructed original.

Unfortunately it is not at all simple [13]. In the matrix inversion the errors on the P_{ji} from finite statistics have devastating consequences and produce unrealistic results. There is a lot of activity in handling this in a sensible way, and in investigating other approaches, such as Maximum Entropy techniques.

3.7. Combining Results

The combination of compatible measurements with different errors is straightforward. However results are sometimes incompatible, or marginally compatible. But something must be done with the results, as the community needs a way of using the combined number. Indeed it is the responsibility of the Particle Data Group [11] to combine measurements and form 'world average' results in a meaningful way.

There is also a problem in combining limits. If two experiments report 95% confidence level upper limits of, say, 0.012 and 0.013, how can one combine these two measurements? This question was put forcefully by the Higgs searches at the end of the LEP run. The four experiments reported results separately compatible and possibly marginally suggestive of a signal from a Higgs boson of mass around 114 GeV/c^2. Did four possibles make a probable? The answer to that statistics question determined whether or not LEP would run another year, at a cost of millions not only in power bills but in its impact on the construction schedule for the LHC. The CERN management decided that the answer in this case was 'no'. History will be their judge.

In combining experiments the likelihood function contains much more information than a simple limit, or value and error. There is a suggestion that these should be routinely published, and we are probably going to see that happening a lot in the future.

3.8. Multivariate Classification

The classification of events (usually 'signal' and 'background') and particles (pion, kaon ...) by means of a cut on a discriminator variable is a basic hypothesis testing problem. However there may be several variables, each containing useful information, and the best choice will be made by combining these in some way.

The Fisher Discriminant has been re-discovered as a technique which is good if the means of distributions differ between the two samples. The Neural Network (feed-forward 'perceptron' configuration) has become a standard item in the toolbox which can handle more general differences, and there are many developments going on in this area.

The use of cuts is deeply engrained. In many cases it is simple and appropriate. However in cases where there are no clean boundaries it may be better to consider all events, weighting them according to their signal-like or background-like nature.

4. CONCLUSIONS

I have given several talks on 'Statistics for Particle Physicists' but 'Particle Physics for Statisticians' has been a new and interesting experience. This has been a very broad view. Particular topics will be considered in detail in the subsequent talks in this conference, in both plenary and parallel sessions. Hopefully the account here will provide you with a map which will help you place them in context.

Acknowledgments

The author gratefully acknowledges the support of the Fulbright Foundation.

References

[1] S. L. Glashow 'Partial Symmetries of Weak Interactions', Nucl. Phys **B22** 579 (1961)
S. Weinberg 'A Model of Leptons', Phys. Rev. Lett. **19** 1264 (1967)
A. Salam 'Weak and Electromagnetic Interactions', Proc. 8^{th} Nobel Symposium, Svartholm 307 (1968).

[2] *BABAR* Collaboration, 'Measurement of time-dependent CP asymmetries in $B^0 \to D^{(*)\pm}\pi^{\mp}$ decays and constraints on $sin(2\beta + \gamma)$' SLAC-PUB-100155, 2003. To be published in Phys. Rev. Lett.

[3] *BABAR* Collaboration, 'Observation of the decay $B^0 \to \pi^0\pi^0$', SLAC-PUB-100092, 2003. To be published in Phys. Rev. Lett.

[4] Details witheld to prevent embarrassment of those involved.

[5] F. James 'MINUIT: Function Minimization and Error Analysis Reference Manual' `http://wwinfo.cern.ch/asdoc/minuit/minmain.html`

[6] J. Orear 'Notes on Statistics for Physicists', University of California report UCRL-8417 (1958) and Cornell report CLNS 82/511 (1982).

[7] A.G. Frodesen et al. 'Probablity and Statistics in Particle Physics', Universitetsforlaget Bergen-Oslo-Tromso (1979).

[8] V.L. Fitch et al., Phys. Rev. Lett. **13** (1964) 138.

[9] Proc. Workshop on Confidence Limits 17-18 January 2000, Ed. F. James, L. Lyons and Y. Perrin. CERN yellow report 2000-005 (2000) `http://user.web.cern.ch/user/Index/library.html`
Fermilab Workshop in Confidence Limits 27-28 March 2000. `http://conferences/fnal.gov/cl2k/`

[10] G.J. Feldman and R.D. Cousins Phys. Rev. **D57** (1998) 37731111.

[11] K. Hagiwara et al, 'The Review of Particle Properties' Phys. Rev. **D66** (2002) 010001.

[12] See e.g., P. Harrison 'Blind Analysis' p 278, Proc. Conf. on Advanced Statistical Techniques in Par-

ticle Physics', Ed. M.R. Whalley and L. Lyons, IPPP/02/39, Durham 2002.

[13] G. Cowan 'A Survey of Unfolding Methods for Particle Physics', p248, Proc. Conf. on Advanced Statistical Techniques in Particle Physics', Ed. M.R. Whalley and L. Lyons, IPPP/02/39, Durham 2002.
V. Blobel, 'An Unfolding Method for High Energy Physics Experiments', p258, *ibid.*

Bayesians, Frequentists, and Physicists

Bradley Efron
Department of Statistics and Department of Health Research and Policy,
Stanford University, Stanford, CA 94305, USA

PHYSTAT2003 brought statisticians together with particle physicists, astrophysicists, and cosmologists. This paper, which is taken from the text of the keynote address, concerns the uneasy relationship between Bayesian and frequentist statistics, with particular attention to the "neutrino problem": how to set confidence limits for a parameter known to be non-negative. Model selection, an objective Bayes technique, gives a different answer than the classic Neyman confidence construction.

1. INTRODUCTION

Ten years ago I gave a talk entitled "Astronomy and biostatistics, the odd couple". It emphasized the mostly unconscious convergence of methods in the two fields, arising from a shared need to account for biased sampling – astronomers because they are stuck on earth and statisticians because humans are such lousy experimental animals, tending to go missing just when you most need their data.

Of course astronomy historically is the most statistical branch of the physics tree. I never expected to attend a conference bringing particle physicists and statisticians under the same roof. The happy existence of PHYSTAT2003 reflects the determination of modern physicists to pursue nature in its most subtle manifestations, where the certainties of mass experimentation give way to small numbers of events observed with great difficulty; in short where statistical inference becomes important.

It is hard to imagine "PHYSTAT1903". Maybe PHYSTAT1803 is slightly more conceivable with Laplace and Gauss arguing the virtues of Bayesian versus frequentist inference. The Bayesian-frequentist argument is certainly a long-lived one, even by the standards of philosophy. It reflects, I believe, two quite different attitudes toward the scientific process: the cautious frequentist desire for objectivity and consensus, versus the individual scientist trying aggressively to make the best sense of past data and the best choice for future direction.

Statistics concerns the efficient accumulation of knowledge – how a scientist learns from experience – so it is not surprising that there is more than one philosophical approach to such a broad problem. I will tread lightly on philosophical matters here. Mainly I want to show how Bayesian and frequentist ideas interact, in particular concerning confidence intervals and the "neutrino problem", where both methodologies show virtues and defects.

2. WHY NOT BAYES?

Bayesianism was the first statistical philosophy, and remains the simplest and most attractive from the point of view of intellectual neatness. Here is a true-life story illustrating Bayesian virtues. A physicist friend of mine and her husband found out, thanks to a sonogram, that they were going to be the parents of twin boys. The doctor told them that one-third of twin pairs are identical, the other two-thirds fraternal, but she wondered if there was a more exact estimate for *her* twins being identical, given that both were to be boys.

This is exactly the problem Bayes solved in 1763. To use standard language, the prior odds of Identical are $(1/3)/(2/3) = 1/2$, while the likelihood ratio (the ratio of probabilities of observing "Both Boys") equals about

$$\frac{P\{\text{Both Boys}|\text{Identical}\}}{P\{\text{Both Boys}|\text{Fraternal}\}} = \frac{1/2}{1/4} = 2.$$

Then Bayes' rule says that the *a posteriori* odds of Identical to Fraternal equal the prior odds times the likelihood ratio,

$$\frac{P\{\text{Identical}|\text{Both Boys}\}}{P\{\text{Fraternal}|\text{Both Boys}\}} = \frac{1}{2} \times 2 = 1.$$

In other words my physicist friend had a 50-50 chance for Identical. (Later she told me that the boys turned out to be "very non-identical".)

Bayes' rule is satisfying, convincing, and fun to use. But using Baye's rule does not make one a Bayesian; *always* using it does, and that's where difficulties begin. "Expert Opinion", which (correctly) gave us the prior odds of one-third to two-thirds for the Twins question, doesn't exist for most genuine scientific problems, or if it does exist may be contentious or just plain wrong. Even the likelihood ratio, which doesn't depend on prior opinions, may be impossible to compute.

Scientists are drawn to Bayesian thinking and language even when they can't use Bayes' theorem. In 2002 a heroic bout of radio telescopy detected subtle polarization in the cosmic microwave background, as predicted by the Big Bang theory. Michael Turner, a leading cosmologist, commented "If you had any doubts that this radiation is from the Big Bang, this should quash them." Even leaving aside the almost theological question of prior odds on the Big Bang theory, this is a case where the denominator of the likelihood ratio,

$$\frac{P\{\text{Polarized}|\text{Big Bang}\}}{P\{\text{Polarized}|\text{Any Other Theory}\}}$$

seems especially obscure. Dr. Turner spoke as a good scientist but a questionable Bayesian.

The 20th century saw the development of a persuasive frequentist theory of statistical inference that continues to dominate scientific practice. (An "objective" Bayesian counter-reformation is stirring, discussed later.) Frequentist statistics does away with the need for prior opinions. This gives frequentism a legitimate claim to objectivity, a considerable virtue in a world with competing scientific teams working at great distances from each other.

Here is a classic frequentist result: data points $x_1, x_2, \ldots, x_n$ are independently observed from a normal distribution with mean μ and variance 1,

$$x_1, x_2, \ldots, x_n \overset{\text{ind}}{\sim} N(\mu, 1), \qquad (1)$$

and we want to say something about the unknown value of μ. The *standard 90% confidence interval* for μ is

$$\mu \in [\bar{x} - 1.645/\sqrt{n},\ \bar{x} + 1.645/\sqrt{n}\ ], \qquad (2)$$

with $\bar{x}$ the mean $\Sigma x_i/n$. Interval (2) contains the unknown μ with 90% probability no matter what μ might be, which is its crucial frequentist property.

Formula (2) and its generalizations are familiar friends, appearing literally millions of times per year in the scientific literature. Nevertheless, there are limitations to the frequentist viewpoint that become more problematical as one moves away from simple situations like (1). One doesn't have to move very far away for difficulties to emerge, as the neutrino problem will show.

3. ACCURACY AND CORRECTNESS

One criticism of frequentist statistics is that it is incomplete: by itself it does not necessarily produce a unique solution to a given inferential problem. In situation (1) for instance the interval

$$\mu \in [\bar{x} - 1.96/\sqrt{n},\ \bar{x} + 1.44/\sqrt{n}\ ]$$

also contains μ with 90% probability. The standard interval (2) is preferred because of its symmetry, dividing its 10% noncoverage probability into 5% errors on either side, but symmetry is not part of the frequentist philosophy.

R.A. Fisher, the founder of modern statistical theory, and the person most responsible for demoting Bayesianism to the back burner, distrusted Jerzy Neyman's confidence interval methodology. He felt that Neyman intervals could be *accurate* (that is, have the claimed coverage probability) without being *correct*.

Here is a simple example of Fisherian "incorrectness". Suppose that in situation (1) the sample size n is either 25 or 100, the choice being determined by the independent flip of a fair coin. It is easy to compute that in this case

$$\bar{x} \pm .262 \qquad (3)$$

is a 90% confidence interval for μ. However it is always incorrect: if $n = 25$ then interval (2) is $\bar{x} \pm .329$, wider than (3), while $n = 100$ gives the narrower interval $\bar{x} \pm .164$.

Fisher called the random variable n an *ancillary statistic*, a quantity conveying no direct information for μ, but whose value determines the accuracy with which μ can be estimated. With typical ingenuity Fisher showed that ancillaries can pop up quite unexpectedly, even in innocuous-looking situations.

Suppose we independently observe $x_1, x_2, \ldots x_{10}$ from a Cauchy distribution with unknown center μ, so each x_i has density

$$f_\mu(x) = \pi^{-1}/[1 + (x - \mu)^2].$$

The maximum likelihood estimate $\widehat{\mu}$ is, to a quite good approximation, normally distributed about μ,

$$\widehat{\mu} \dot{\sim} N(\mu, \sigma^2) \quad \text{with} \quad \sigma = 0.447, \qquad (4)$$

but the obvious 90% confidence interval $\mu \in \widehat{\mu} \pm 1.645 \cdot 0.447$, which is almost perfectly accurate, runs into the same correctness problems as (3).

Here the ancillary is the second derivative of the log likelihood function $\ell(\mu) = \log \prod_{i=1}^{10} f_\mu(x_i)$ evaluated at $\widehat{\mu}$,

$$\ddot{\ell} = \frac{\partial^2}{\partial \mu^2} \ell(\mu)|_{\mu = \widehat{\mu}}.$$

The conditional distribution of $\widehat{\mu}$ given $\ddot{\ell}$ is approximately

$$\widehat{\mu}|\ddot{\ell} \sim N(\mu, (-\ddot{\ell})^{-1}),$$

so the magnitude of $-\ddot{\ell}$ determines the accuracy of $\widehat{\mu}$ for estimating μ. The correct interval is

$$\widehat{\mu} \pm 1.645 \cdot (-\ddot{\ell})^{-\frac{1}{2}}. \qquad (5)$$

Figure 1 shows log likelihood functions for two Cauchy samples each of size 10, both of which have

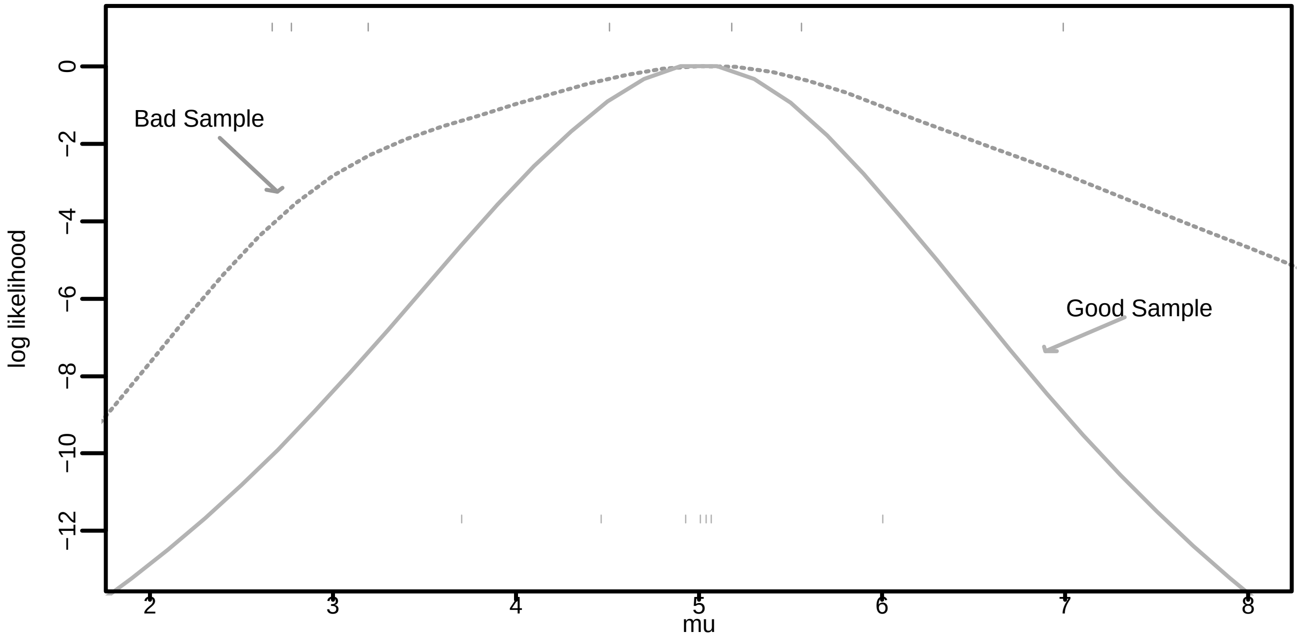


Figure 1: Log likelihood functions for two Cauchy samples of size $n = 10$, each with $\widehat{\mu} = 5.00$; μ can be determined much more accurately in the "good sample", which has its likelihood decreasing more rapidly on either side of $\widehat{\mu}$.

$\widehat{\mu} = 5.00$. However the "good sample" has its likelihood declining much more rapidly on either side of $\widehat{\mu}$, resulting in a shorter interval (5),

$$\text{Good Sample}: \ 5.00 \pm .35; \quad \text{Bad Sample}: \ 5.00 \pm .65. \tag{6}$$

(The difference is the closer spacing of the good sample's observations around $\widehat{\mu} = 5.00$.)

Both (5) and the unconditional interval $\widehat{\mu} \pm 1.645 \cdot 0.447$ are accurate, i.e. have 90% frequentist coverage, but (5) more correctly reflects the information available in the sample at hand. This problem doesn't affect interval (2) since $\bar{x}$ is a sufficient statistic in the normal case (1), but an argument can be made that Cauchy-type situations are actually more common.

It's worth noting that the simplest Bayesian analysis, beginning with a flat prior for μ (an improper uniform distribution over the entire real line) automatically gives results like (6). This is a reminder that Bayesian properties are important to consider even if one eventually intends a frequentist analysis.

4. THE NEUTRINO PROBLEM

Adding a non-negativity constraint,

$$\mu \geq 0 \tag{7}$$

to the normal sampling assumptions (1) brings us to what might be called "the neutrino problem" since it arises pertinaciously in experiments assessing the mass of the neutrino. Here we will state the problem as setting a 95% upper limit for μ under model (1) and (7). There is no loss of generality taking $n = 1$ in (1), so we can express the data simply as a single observation of

$$x \sim N(\mu, 1). \tag{8}$$

The standard one-sided interval for μ in this case is

$$\mu \leq x + 1.645. \tag{9}$$

The standard interval in perfectly accurate, covering μ exactly 95% of the time no matter what its value might be. However it seems incorrect, especially to physicists, since if x is less than -1.645 interval (9) is empty of values satisfying (7). Exactly this case arose in actual neutrino experimentation, see for example Mandelkern (2002).

Various alternative bounds have been suggested, the best-known being the Feldman-Cousins bounds (1998), classic Neyman confidence limits using a likelihood ratio ordering rule. Figure 2 displays four possibilities: the standard bound $x + 1.645$; Feldman-Cousins; the Bayes upper 95% *a posteriori* value beginning with a uniform prior density for μ on $[0, \infty)$; and a bound based on *model selection*, as discussed below, an "objective Bayesian" construction that takes seriously the possibility of μ equaling zero. If a good bound is a small one then Model Selection is the clear winner, the Uniform$[0, \infty)$ Bayesian bound is worst, while Feldman-Cousins and the standard bound are intermediate, differing only for $x < 0$. But of course there is more to the story.

Which upper bound is right? The Bayesian answer would be "Surely that depends upon our prior knowledge concerning μ. If there is nothing scientifically special about $\mu = 0$ except as the lower endpoint of the allowable μ domain, then a case can be made for the Uniform$[0, \infty)$ prior approach. In some situations though there may be strong scientific grounds for suspecting that μ equals zero, or at least is very close to zero compared to the standard deviation 1 of the observation $x \sim N(\mu, 1)$. Then an analysis should reflect those beliefs."

"Model Selection" refers to the last situation. We assume that there are two possible models for the parameter μ in (8),

$$\begin{array}{ll} \textit{Small Model} & \mathcal{M}_0 : \mu = 0 \\ \textit{Big Model} & \mathcal{M}_1 : \mu > 0. \end{array} \tag{10}$$

Having observed $x \sim N(\mu, 1)$ we wish to set a 95% upper bound for μ, including in our calculations the possibility that $\mu = 0$. Because the two models are of different dimensions, a hallmark of the general model selection problem, this is a more intricate task than setting confidence bounds within a single model.

A full Bayesian analysis requires prior probabilities on the two models in (10), and also a prior density on μ when $\mathcal{M}_1$ applies,

$$\textit{A priori} \begin{cases} \mathcal{M}_0 \text{ true with probability } \pi_0 \\ \mathcal{M}_1 \text{ true with probability } \pi_1 = 1 - \pi_0 \end{cases} \tag{11}$$

and

$$\mu \sim g(\cdot) \quad \text{if} \quad \mathcal{M}_1 \quad \text{true,} \tag{12}$$

where $g(\mu)$ is some prior density on $(0, \infty)$.

The Bayes 95% upper limit for μ having observed $x \sim N(\mu, 1)$ from prior situation (11, 12) depends on an integral of the sampling density $\varphi_\mu(x) = (2\pi)^{-\frac{1}{2}} \exp\{-\frac{1}{2}(x - \mu)^2\}$,

$$\varphi_{(1)}(x) \equiv \int_0^\infty \varphi_\mu(x) g(\mu) d\mu,$$

as well as $\varphi_0(x) = (2\pi)^{-\frac{1}{2}} \exp\{-x^2/2\}$. Bayes' rule then provides the *a posteriori* probability of $\mathcal{M}_1$ given x,

$$\text{Prob}\{\mathcal{M}_1 | x\} = \pi_1 \varphi_{(1)}(x) / [\pi_0 \varphi_0(x) + \pi_1 \varphi_{(1)}(x)], \tag{13}$$

and the probability that μ exceeds some positive value "c" given $\mathcal{M}_1$ and x,

$$\text{Prob}\{\mu > c | \mathcal{M}_1, x\} = \int_c^\infty \varphi_\mu(x) g(\mu) d\mu / \varphi_{(1)}(x). \tag{14}$$

Since μ can exceed c only if $\mathcal{M}_1$ is true, the *a posteriori* probability of $\mu > c$ is

$$\text{Prob}\{\mu > c | x\} = (13) \cdot (14). \tag{15}$$

The value of c that makes (15) equal .05 is the Model Selection upper 95% bound.

At this point the non-Bayesian may quail at making prior selections of π_0 and $g(\mu)$. "Objective Bayes", currently the most popular form of Bayesian statistics, attempts to alleviate such fears by restricting attention to priors that enjoy reasonable frequentist properties.

First consider choosing $g(\mu)$ in (12). Without the negativity constraint (1) we might very well take $g(\mu) \equiv 1$, which yields the standard interval (9) as the one-sided Bayes 95% interval. This suggests using

$$g(\mu) = 1 \quad \text{on} \quad (0, \infty) \tag{16}$$

in (12), the prior that gave the dotted curve in Figure 2. (Another choice is mentioned below.) Now $\varphi_{(1)}(x)$ equals the standard normal cumulative distribution function

$$\Phi(x) = \int_{-\infty}^x \varphi_0(x) dx,$$

so (13) becomes

$$\text{Prob}\{\mathcal{M}_1 | x\} = \pi_1 \Phi(x) / [\pi_0 \varphi_0(x) + \pi_1 \Phi(x)]. \tag{17}$$

We can complete the choice of objective prior by selecting π_0, and $\pi_1 = 1 - \pi_0$, to make

$$\text{Prob}\{\mathcal{M}_1 | x_0\} = \frac{1}{2} \tag{18}$$

at some "break-even point" x_0 that is comfortable for frequentists. Efron and Gous (2001) argue that

$$x_0 = 1.28 \tag{19}$$

the 90th percentile of x under $\mathcal{M}_0$, is a choice commensurate with standard Fisherian hypothesis testing. The "Model Select" curve in Figure 2 is formula (15) evaluated for the prior choices (16-19). [Note: solving $P\{\mathcal{M}_1 | x_0\} = \frac{1}{2}$ in (17) determines π_0 and $\pi_1 = 1 - \pi_0$, so that (15) can be evaluated. However π_0 cannot be interpreted as the actual prior probability of $\mathcal{M}_0$; we would get a different π_0, but the same values of (15), if (16) were changed to say $g(\mu) = 2$ on $(0, \infty)$.]

Figure 3 graphs Prob$\{\mathcal{M}_0 | x\}$, the *a posteriori* probability that $\mu = 0$. The values roughly agree with standard frequentist hypothesis testing: observing x equal the 95th $\mathcal{M}_0$ percentile 1.645 gives mild evidence against $\mathcal{M}_0$, Prob$\{\mathcal{M}_0 | x\} = 0.36$; x equal the 99th percentile gives strong evidence, etc. Observing $x = 0$ gives 80% probability that $\mu = 0$. This leaves only 20% for $\mu > 0$, and leads to the notably small 95% upper bound.

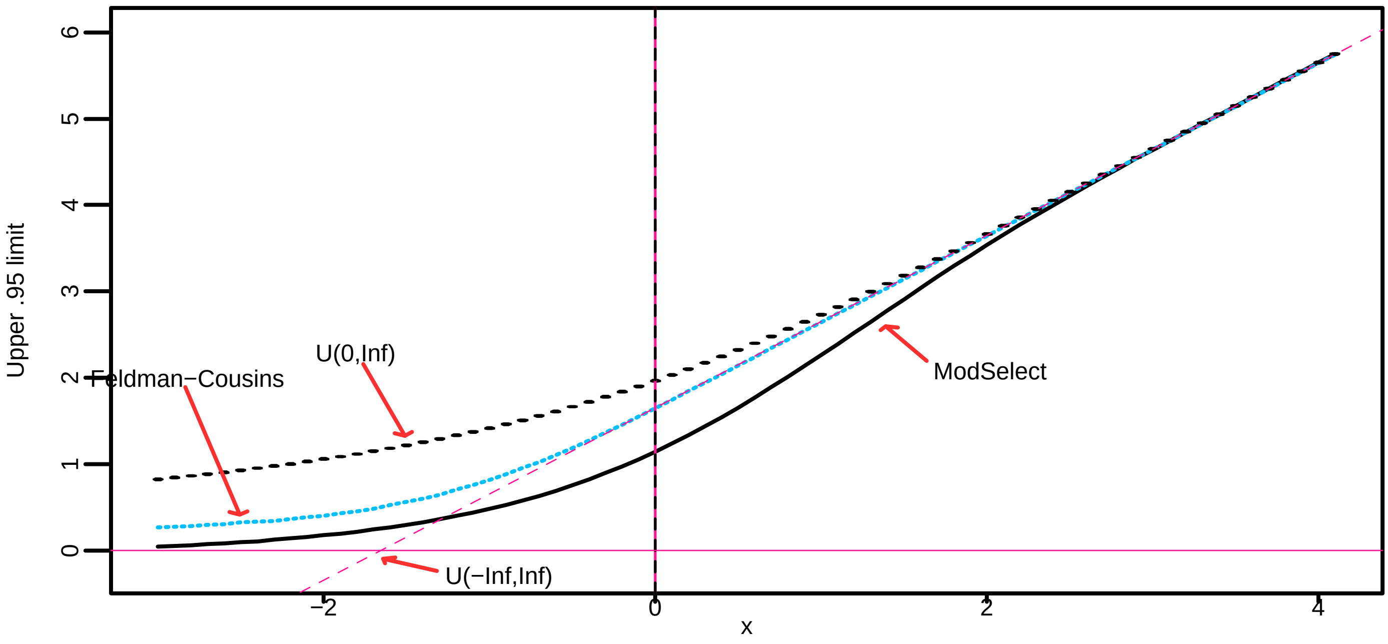


Figure 2: Four possible 95% upper bounds for μ having observed x in model (7,8): standard frequentist and Bayesian bounds $x + 1.645$ (labelled U(-Inf, Inf) in figure); Bayesian bound given a uniform prior on $[0, \infty]$ (labelled U(0, Inf)); Model Selection bound.

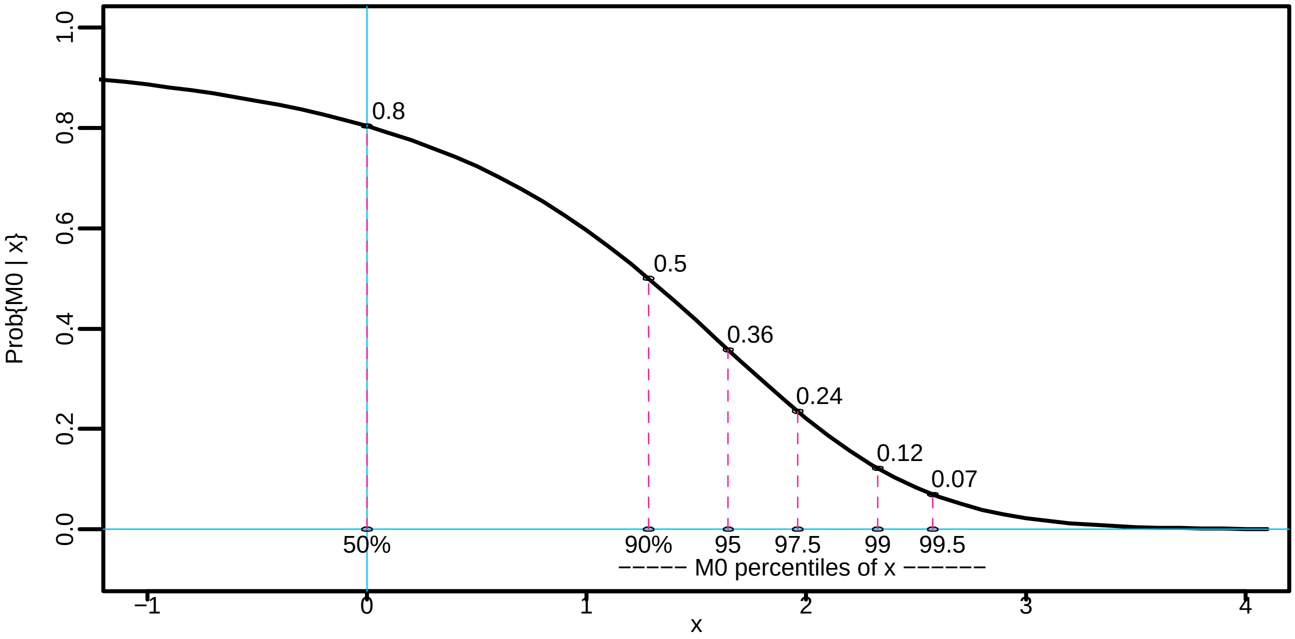


Figure 3: *A posteriori* probability that $\mu = 0$ given x for Model Selection specifications (16)-(19); observing $x = 1.645$ gives 36% probability that $\mu = 0$, etc.

From a classical point of view Model Selection looks like an ungainly combination of hypothesis testing and estimation. We used Bayes' theorem to handle the problem, but the resulting method is not subjective in the sense of employing specific prior knowledge (or guesses) concerning neutrinos. What it does employ is a qualitative belief that values of μ at or near 0 deserve increased weight. "Near" means near compared to the variance 1 of the observation $x \sim N(\mu, 1)$. Mandelkern's discussion suggests that this was and is the case for the actual electron neutrino problem.

It is easy to criticize Bayesian Model Selection methods. Even trying to be objective, we still have had to make an uncomfortable number of prior specifications. Alternative specifications are certainly possible. Choosing

$$g(\mu) = 1 + 3e^{-3\mu} \quad (\text{for } \mu > 0),$$

instead of $g(\mu) = 1$, makes the dotted curve in Figure 2 look much more like the standard bound $x + 1.645$ when $x \geq 0$. However this change doesn't have much effect on the Model Select curve.

Moving the break-even point x_0 closer to zero, which effectively puts less prior probability on $\mu = 0$, changes the Model Selection upper 95% value more dramatically. Taking $x_0 = 0.50$ instead of 1.28 makes the Model Selection upper 95% bounds agree rather nicely with the Feldman-Cousins curve, as seen in Figure 4.

At this point an optimist could say that we have the best of all possible statistical worlds, with close agreement between the Bayesian Model Selection and Neyman frequentist approaches. However the two methods deliver their 95% upper bounds quite differently. At $x = -0.50$, roughly the neutrino result obtained by the "Troisk" group in (1999) according to Mandelkern, the Feldman-Cousins and the Model Selection bounds both equal about 1.15; however Model Selection (with $x_0 = 0.50$) implies a 69% probability that μ equals or is quite near 0. If believed, this might influence subsequent detection strategies.

5. MULTIDIMENSIONAL PROBLEMS AND STEIN'S PARADOX

The neutrino problem is one-dimensional in the sense of involving only a single real-valued parameter. The relationship between Bayesian and frequentist methods becomes more difficult, and more intriguing, when we venture into multi-dimensional situations.

Suppose then that instead of (8) the observed data $\mathbf{x}$ is a p-dimensional normal vector having unknown mean vector $\boldsymbol{\mu}$ and identity covariance matrix I,

$$\mathbf{x} \sim N_p(\boldsymbol{\mu}, I). \tag{20}$$

We wish to make inferences about $\boldsymbol{\mu}$ having observed $\mathbf{x}$. Another way to express (20) is to say that we have p independent versions of (8), each with its own unknown parameter μ_i,

$$x_i \overset{\text{ind}}{\sim} N(\mu_i, 1) \quad i = 1, 2, \ldots, p \tag{21}$$

now removing the requirement that μ_i is not negative.

At first glance this looks easy enough. The maximum likelihood estimate of $\boldsymbol{\mu}$ is $\widehat{\boldsymbol{\mu}} = \mathbf{x}$, i.e. $\widehat{\mu}_i = x_i$ for $i = 1, 2, \ldots, p$. The obvious 90% confidence region for $\boldsymbol{\mu}$ is a sphere centered at $\mathbf{x}$, say

$$C = \{\boldsymbol{\mu} : \|\boldsymbol{\mu} - \mathbf{x}\|^2 < c\},$$

where c is the upper 90% point of a chi-square distribution with p degrees of freedom. The obvious "objective prior" $g(\boldsymbol{\mu}) = 1$ for $\boldsymbol{\mu}$ in $\mathcal{R}^p$ also leads to point estimate $\widehat{\boldsymbol{\mu}} = \mathbf{x}$ as *a posteriori* mean and to C as the *a posteriori* Bayes 90% region, so Bayesian and frequentist methods agree with each other.

It came as a great surprise to statisticians that there is something wrong with these "obvious" estimates and confidence regions. Charles Stein, working with his graduate student Willard James in the early 1960's, produced an estimator $\widetilde{\boldsymbol{\mu}}$ uniformly superior to the MLE $\widehat{\boldsymbol{\mu}} = \mathbf{x}$ in terms of expected squared error,

$$\widetilde{\boldsymbol{\mu}} = \left[1 - \frac{p-2}{\|\mathbf{x}\|^2}\right] \cdot \mathbf{x}. \tag{22}$$

The James-Stein theorem states simply that in dimensions $p \geq 3$,

$$E\{\|\widetilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2\} < E\{\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2\} \tag{23}$$

for every possible choice of the parameter vector $\boldsymbol{\mu}$. Similarly, one can find 90% confidence regions uniformly smaller than C.

Statisticians have a lot invested in the estimator $\widehat{\boldsymbol{\mu}} = \mathbf{x}$, (it underlies analysis of variance and regression theory among other things) so Stein's result was initially viewed as paradoxical. A great deal of subsequent effort has gone into understanding Stein's paradox and tapping its potential for improved estimation in high dimensions.

The difference in squared error seen in (23) can be enormous in realistic applied problems. An example of Stein estimation (applied in a binomial setting rather than (14), but basically the same procedure) appears in Table 1. The batting averages of 18 major league baseball players in their first 45 at bats of the 1970 season are given in column one. They play the role of MLE estimates $\widehat{\mu}_i$ for the underlying true averages μ_i. The second column shows each player's average for the remainder of the season. These typically involved several hundred more at bats so we can use them as surrogates for the true μ_i. Column three gives a version of the James-Stein estimate $\widetilde{\boldsymbol{\mu}}$ (based only on the first 45 at bats) applicable to binomial data. The ratio of squared errors is found to be overwhelming in this case, $\sum_{i=1}^{n}(\widehat{\mu}_i - \mu_i)^2 \Big/ \sum_{i=1}^{n}(\widetilde{\mu}_i - \mu_i)^2 = 3.50$.

Stein's work was carried out in a frequentist framework but results like (23) also disturbed Bayesians.

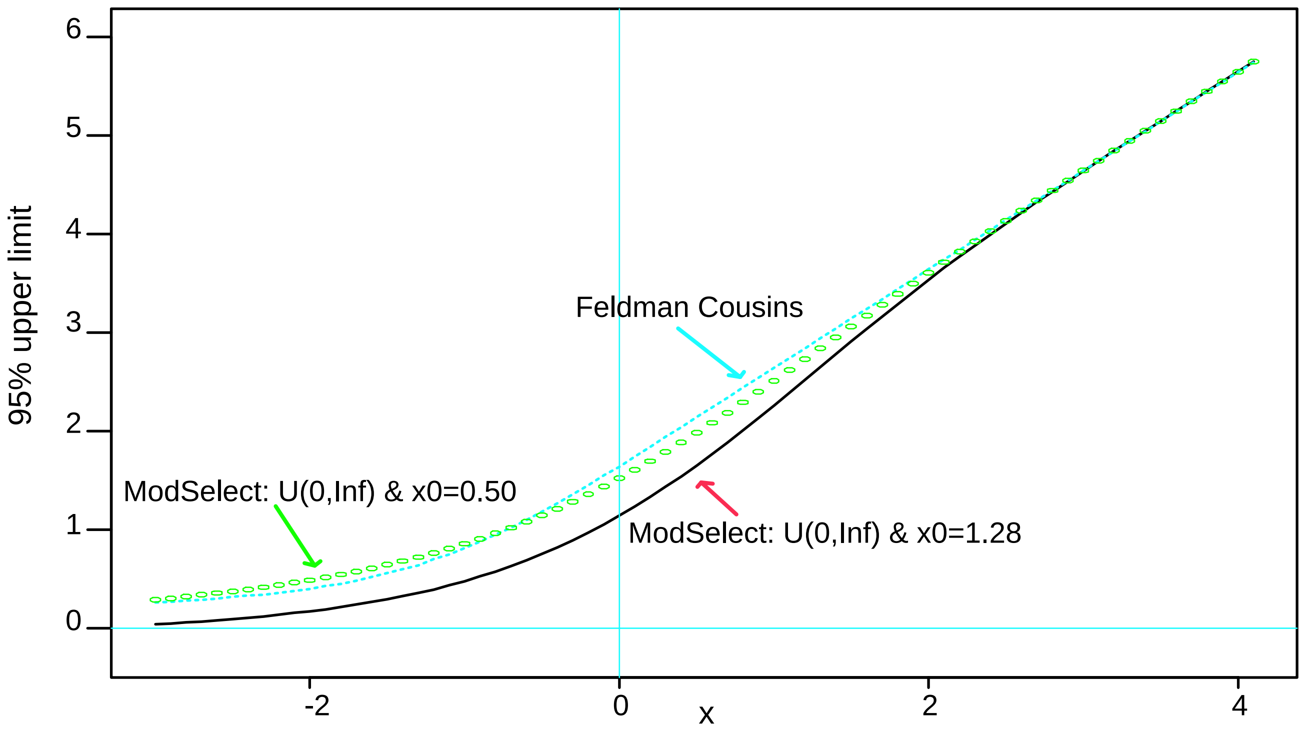


Figure 4: Moving the break-even point x_0, (17), closer to zero makes the Model Selection upper 95% point nearly match the Feldman-Cousins bound.

Why can the estimates and confidence regions derived from the "uformative" prior $g(\mu) \equiv 1$ be uniformly dominated? It turns out that there are better uninformative priors for high-dimensional problems.

Suppose we assume a scaled multivariate normal prior for $\boldsymbol{\mu}$ in (20), all the variances equalling a single unknown value A,

$$\boldsymbol{\mu} \sim N_p(\mathbf{0}, A \cdot I), \tag{24}$$

but then take the scaler A in (24) to itself have a flat "prior prior", say

$$h(A) \equiv 1 \quad \text{for} \quad A > 0. \tag{25}$$

The Bayes estimator $\widehat{\mu}_{\text{Bayes}} = E\{\boldsymbol{\mu}|\mathbf{x}\}$ computed from the *hierarchical prior* (24)-(25) turns out to closely resemble the James-Stein estimator, and likewise dominates the MLE $\widehat{\boldsymbol{\mu}} = \mathbf{x}$ as in (23). Another way to say this is that $\widehat{\boldsymbol{\mu}}_{\text{Bayes}}$ from (24)-(25) has smaller Bayes risk than $\widehat{\boldsymbol{\mu}}_{\text{Bayes}}$ from $g(\boldsymbol{\mu}) \equiv 1$ *no matter what the true prior is assumed to be.* In other words (24)-(25) is more uninformative than a flat prior, at least in dimensions $p \geq 3$. Perhaps the most important general lesson is that the facile use of what appear to be uninformative priors is a dangerous practice in high dimensions.

Hierarchical priors dominate the recent Bayesian literature. They lead to an interesting amalgam of frequentist and Bayesian thinking called *empirical Bayes.* Empirical Bayes ideas neatly motivate the James-Stein estimator (22). For a given value of A in the normal prior (24), the *a posteriori* expectation of $\boldsymbol{\mu}$ given $\mathbf{x} \sim N_p(\boldsymbol{\mu}, I)$ is

$$\widehat{\boldsymbol{\mu}}_A = B\mathbf{x} \quad \text{where} \quad B = A/(A+1).$$

If we don't know A, we can replace B with the estimate

$$\widehat{B} = 1 - (p-2)/\|\mathbf{x}\|^2,$$

which is unbiased for B, in the usual frequentist sense, under assumptions (24) and (20). In this way the James-Stein estimator $\widetilde{\boldsymbol{\mu}} = \widehat{B}\mathbf{x}$ is an obvious frequentist estimate for the unavailable Bayes rule $\widehat{\boldsymbol{\mu}}_A$ we would like to use.

Notice that the James-Stein estimator $\widetilde{\mu}_i$ for μ_i depends on the other observations x_j, $j \neq i$. Even though the component problems are assumed independent in (21), putting them together as in (22) is always better, in an expected total squared-error sense, than using the separate estimators $\widehat{\mu}_i$. Originally this seemed the most paradoxical element of James-Stein estimation. The empirical Bayes argument helps reduce the sense of paradox. Sets of problems that are analyzed together, like the baseball averages, may relate to each other at the parameter level even if their individual statistical errors are independent.

The 21st Century has brought statisticians bigger problems to deal with, involving inferences for hundreds and even thousands of parameters at once.

Table I A baseball example of Stein estimation, 1970 season. The James-Stein estimator based on each player's first 45 at bats does much better than their observed averages at predicting subsequent performance.

	$\widehat{\mu}_i$	μ_i	$\widetilde{\mu}_i$
	First 45	Remainder Season	James-Stein
Clemente	0.400	0.346	0.290
F. Robinson	0.378	0.298	0.286
F. Howard	0.356	0.276	0.282
Johnstone	0.333	0.222	0.277
Berry	0.311	0.273	0.273
Spencer	0.311	0.270	0.273
Kessinger	0.289	0.263	0.268
Alvarado	0.267	0.210	0.264
Santo	0.244	0.269	0.259
Swoboda	0.244	0.230	0.259
Unser	0.222	0.264	0.254
Williams	0.222	0.256	0.254
Scott	0.222	0.303	0.254
Petrocelli	0.222	0.264	0.254
Rodriguez	0.222	0.226	0.254
Campaneris	0.200	0.285	0.249
Munson	0.178	0.316	0.244
Alvis	0.156	0.200	0.239

An uneasy alliance between Bayesian and frequentist methods, epitomized by empirical Bayes, seems to be replacing the time-worn adversarial relationship. We can hope all of this will be settled by the time of PHYSTAT2103.

References

Mandelkern, M. (2002). "Setting Confidence Intervals for Bounded Parameters" *Statistical Science* **17** 149-192. [A statisticians' guide to the "neutrino problem" as it actually occurs.]

Feldman, G. and Cousins, R. (1998). "Unified approach to the classical statistical analysis of small signals" *Physical Review D* **57** 3873-89. [The most popular approach to bounded parameter confidence limits; considers lower as well as upper bounds.]

Efron, B. (1998). "R.A. Fisher in the 21st Century" *Statistical Science* **13** 95-122. [Discusses correctness and the frequentist-Bayesian relationship.]

Efron, B. and Hinkley, D. (1978). "Assessing the accuracy of the MLE: observed versus expected Fisher information" *Biometrika* **65** 457-87. [The Cauchy ancillarity example.]

Efron, B. and Gous, A. (2001). "Scales of evidence for model selection: Fisher versus Jeffreys" *IMS Lecture Series* **38** 208-256. [Jeffreys, a geophysicist and objective Bayesian, suggested a different scale of evidence than the familiar Fisherian .05 etc.]

Efron, B. and Morris, C. (1975). "Data analysis using Stein's estimator and its generalizations" *Jour. Amer. Stat. Assoc.* **70** 311-319. [Includes the baseball example. A popular version appeared in 1977 *Scientific American* **236** 119-127.]

Efron, B. "Robbins, Empirical Bayes, and Microarrays" *Annal. Stat.* **31** 366-378. [Empirical Bayes and high-dimensional problems in biogenetics.]

Using What We Know: Inference with Physical Constraints

Chad M. Schafer and Philip B. Stark
Department of Statistics, University of California, Berkeley, CA 94720, USA

Frequently physical scientists seek a confidence set for a parameter whose precise value is unknown, but constrained by theory or previous experiments. The confidence set should exclude parameter values that violate those constraints, but further improvements are possible: We construct *minimax expected size* and *minimax regret* confidence procedures. The resulting confidence sets include only values that satisfy the constraints; they have the correct coverage probability; and they minimize a measure of average size. We illustrate these approaches with three examples: estimating the mean of a normal distribution when this mean is known to be bounded, estimating a parameter of a bivariate normal distribution arising in a signal detection problem, and estimating cosmological parameters from MAXIMA-1 observations of the cosmic microwave background radiation. In the first two examples, the new methods are compared with two others: a standard approach adapted to force the estimate to conform to the bounds, and the likelihood-ratio testing approach proposed by Feldman and Cousins [1998]. Software that implements the new method efficiently is available online.

1. INTRODUCTION

In many statistical estimation problems parameters are just indices of stochastic models, but in the physical sciences parameters are often physical constants whose values have scientific interest. Previous experiments, theory and physical constraints often limit the possible or plausible values of unknown constants. In cosmology, for example, decades of observation and theoretical research have led to wide agreement on the range of possible values for key cosmological parameters, such as the Hubble constant and the age of the Universe. A good statistical method should use everything we know—data and physical constraints—to make inferences as sharp as possible. This paper looks at the problem of incorporating prior constraints into confidence sets from a frequentist perspective.

There is a duality between hypothesis tests and confidence sets. Suppose that Θ is the set of possible values of the parameter θ (either a scalar or a vector), and let η denote a generic element of Θ. Let $A(\eta)$ be an *acceptance region* for testing the hypothesis that $\theta = \eta$. If the data, a realization of the random variable X, fall within $A(\eta)$, we consider $\theta = \eta$ an adequate explanation of the data, while if the data fall outside $A(\eta)$, we reject the hypothesis $\theta = \eta$. The chance when $\theta = \eta$ that the data fall outside $A(\eta)$ is the probability of *type I error*—the significance level—of the test.

Suppose we have a family of acceptance regions $\{A(\eta) : \eta \in \Theta\}$, each with significance level at most α; that is,

$$P_\eta\{X \notin A(\eta)\} \le \alpha, \ \ \forall \eta \in \Theta. \tag{1}$$

Then the set

$$C_A(x) \equiv \{\eta \in \Theta : x \in A(\eta)\} \tag{2}$$

is a confidence procedure for θ with *confidence level* at least $1 - \alpha$. That is,

$$P_\theta\{C_A(X) \ni \theta\} \ge 1 - \alpha, \ \ \forall \theta \in \Theta. \tag{3}$$

Tailoring the acceptance regions $\{A(\eta)\}$ lets us control properties of the resulting confidence set.

For example, we might want the confidence set to include the smallest possible range of parameter values. That would lead us to pick $A(\eta)$ to minimize the probability when $\theta \ne \eta$ that $X \in A(\eta)$, (the probability of *type II error*). It is generally not possible to minimize these false coverage probabilities simultaneously over all $\theta \in \Theta$. The constraint $\theta \in \Theta$ avoids tradeoffs in favor of impossible models.

Incorporating bounds is simple with Bayesian methods: Use a prior that assigns probability one to the set Θ. However, any prior does more than impose the constraint $\theta \in \Theta$: It also assigns probabilities to all measurable subsets of Θ. In problems with infinite-dimensional parameters, it can be impossible to find a prior that honors the physical constraints [Backus 1987, 1988].

1.1. Expected Size of Confidence Regions as Risk

We want a confidence procedure to produce sets that are as small (accurate) as possible, but still to have coverage probability $1-\alpha$, no matter what value θ has, provided it is in Θ. To quantify size, we use an arbitrary measure ν on Θ (typically ν is ordinary volume—Lebesgue measure). We study how the expected size of the region depends on the true value of the parameter θ. This embeds our problem in statistical decision theory: We compare estimators based on their *risk functions* over $\theta \in \Theta$, where risk is the expected measure of the confidence region.

It is rare that one procedure minimizes the expected size for every $\theta \in \Theta$. (Such procedures are *uniformly most accurate* (UMA) confidence procedures. See Schervish [1995], for example.) Making the expected size small for one value of θ tends to make it larger for other values, so minimizing the expected size for $\theta \notin \Theta$ tends to make the expected size unnecessarily large for some values of $\theta \in \Theta$. We seek the *minimax expected size* (MES) confidence procedure: the procedure that minimizes the maximum expected size for parameter values θ that are members of Θ, the set of possible theories. Thus, the parameter constraint $\theta \in \Theta$ enters in two ways: The confidence region includes only values in Θ, and the expected size is considered only for $\theta \in \Theta$.

MES is the inversion of a family of hypothesis tests that are most powerful against a *least favorable alternative* (LFA), a mixture of theories $\{P_\eta : \eta \in \Theta\}$; those tests are based on likelihood ratios. Evans et al. [2003] establish in some generality that MES is of this form. (Often in decision theory the minimax procedure is the Bayes procedure for the prior that yields the largest Bayes risk.) Typically, we can only approximate MES numerically.

Forming confidence regions by inverting hypothesis tests based on likelihood ratios is not new in the physical sciences. For example, Feldman and Cousins [1998] construct confidence intervals by inverting the likelihood ratio test (LRT). (See Bickel and Doksum [1977], Lehmann [1986], and Schervish [1995] for discussions of LRT.) MES has two advantages over LRT: First, it is optimal in a sense that clearly measures accuracy. Second, although approximating the LFA can be challenging, performing the likelihood ratio test in complex situations can be even more difficult because one must calculate the restricted MLE for all possible data.

On the other hand, LRT has an appealing invariance under reparametrization: The LRT confidence set for a transformation of a parameter is just the same transformation applied to the LRT confidence set for the original parameter. In contrast, a transformation of the MES confidence set is a confidence set for the transformation of the parameter, but typically it is not the same set as the MES confidence set designed for the transformed parameter—it has larger (maximum) expected measure. Bayesian credible regions based on "uninformative" priors also lack this kind of invariance, because a prior that is flat in one parametrization is not flat after a non-affine reparametrization. (None of these methods necessarily produces a confidence *interval* under reparametrizations. For example, a confidence interval for θ^2 that does not include zero would transform to a confidence set for θ that is the union of two disjoint intervals.)

Any procedure that has $1-\alpha$ coverage probability for all $\eta \in \Theta$ has strictly positive expected measure for all $\theta \in \Theta$. Let $r(\theta)$ be the infimum of the risks at θ of all $1-\alpha$ confidence procedures. The *regret* of a confidence procedure at the point θ is the difference between $r(\theta)$ and the risk at θ [DeGroot 1988]. MES is the $1-\alpha$ procedure whose supremal risk over $\eta \in \Theta$ is as small as possible. In contrast, the *minimax regret* procedure (MR) is the $1-\alpha$ confidence procedure for which the supremum of the regret is smallest. MR can be constructed in much the same was as MES, by finding a *least regrettable alternative* (LRA). MES and MR can be quite different, as illustrated in section 2.

The next section gives two simple examples demonstrating MES and MR, and contrasting them with a classical approach and LRT. Section 3 sketches the theory behind MES and MR in more detail. Section 4 applies the approaches to a more complicated problem: estimating cosmological parameters from observations of the cosmic microwave background radiation (CMB). Section 5 describes a computer algorithm for approximating MES and MR in complex problems such as the CMB problem.

2. SIMPLE EXAMPLES

2.1. The Bounded Normal Mean Problem

We observe a random variable X that is normally distributed with mean θ and variance one. We know *a priori* that $\theta \in [-\tau, \tau] = \Theta$. We seek a confidence interval for θ. Evans et al. [2003] discuss this problem in detail, and characterize the MES procedure. Compare the following three approaches:

1. **Truncating the standard confidence interval.** Let z_p be the p^{th} percentile of the standard normal distribution. A simple approach that honors the restriction $\theta \in [-\tau, \tau]$ is to intersect the usual confidence interval $[X - z_{1-\alpha/2}, X + z_{1-\alpha/2}]$ with $[-\tau, \tau]$. The resulting confidence interval corresponds to inverting hypothesis tests whose acceptance regions are

$$A_{\text{TS}}(\eta) = \left[\eta - z_{1-\alpha/2},\ \eta + z_{1-\alpha/2}\right] \tag{4}$$

   for $\eta \in [-\tau, \tau]$. This is an intuitively attractive solution, and it is the only *unbiased* procedure: The parameter value that the interval is most likely to cover is the true value θ. However, some biased procedures have smaller maximum expected length.

2. **Inverting the likelihood ratio test.** Let $\hat{\theta}$ denote the restricted maximum likelihood estimate of θ: the parameter value in Θ for which the likelihood is greatest, given data $X = x$. Acceptance regions for the likelihood ratio test are formed by setting a threshold k_η for the ratio of the likelihood of the parameter η and the likelihood of $\hat{\theta}$; the hypothesis $\theta = \eta$ is rejected if the ratio is too small. The threshold is chosen so that when $\theta = \eta$, the probability that $X \in A(\eta)$ is at least $1-\alpha$. Thus,

$$A_{\text{LRT}}(\eta) = \left\{ x \colon \frac{\phi(x-\eta)}{\phi(x-\hat{\theta})} \geq k_\eta \right\}, \tag{5}$$

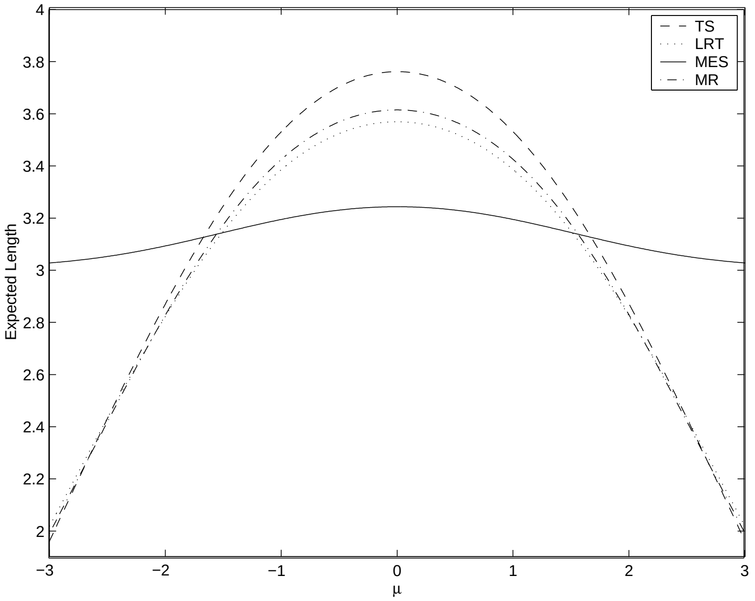


Figure 1: Expected lengths of the 95% confidence intervals for a bounded normal mean as a function of the true value θ for $\tau = 3$.

where $\phi(\cdot)$ is the standard normal density function:

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-(z)^2/2\right) \tag{6}$$

and

$$\hat{\theta} = \begin{cases} -\tau, & x \le -\tau \\ x, & -\tau < x < \tau \\ \tau, & x \ge \tau \end{cases} . \tag{7}$$

3. **Minimax expected size procedure.** Both MES and LRT are based on inverting tests involving likelihood ratios, but the likelihoods in the denominator (the alternative hypotheses) are different. The MES acceptance regions are

$$A_{\mathrm{MES}}(\eta) = \left\{ x \colon \frac{\phi(x-\eta)}{\int_{-\tau}^{\tau} \phi(x-u)\lambda(du)} \ge c_\eta \right\}, \tag{8}$$

where λ is the least favorable alternative and c_η is chosen so that the coverage probability is $1-\alpha$.

Table I Maximum expected lengths of three 95% confidence procedures for estimating the mean of a normal distribution when the mean is known to be in the interval $[-\tau, \tau]$. TS is the truncated standard procedure, LRT is the inversion of the likelihood ratio test (the Feldman-Cousins approach), and MES is the minimax expected size procedure.

τ	**TS**	**LRT**	**MES**
1.75	2.9	2.7	2.6
2.00	3.2	2.9	2.8
2.25	3.4	3.1	3.0
2.50	3.6	3.3	3.1
2.75	3.7	3.5	3.2
3.00	3.8	3.6	3.2
3.25	3.8	3.7	3.3
3.50	3.9	3.7	3.3
3.75	3.9	3.8	3.4
4.00	3.9	3.8	3.4

Table I lists the maximum expected sizes for each of these three procedures for several values of the bound τ. The advantage of MES is larger when τ is larger. Figure 1 compares the expected lengths of the intervals as a function of the true value θ for $\tau = 3$. The MES procedure attains smaller expected length at $\theta = 0$ at the cost of larger expected length when $|\theta|$ is large. When $\tau \le 2z_{1-\alpha}$, as it is here, the MES procedure minimizes the expected size at $\theta = 0$ (equivalently, the regret of MES at $\theta = 0$ is zero: λ assigns probability one to $\theta = 0$). The MES interval in the bounded normal mean problem is a truncated version of the confidence interval proposed by Pratt [1961] for estimating an unrestricted normal mean. Figure 1 also shows the expected size of the MR interval. The expected size at zero is larger for MR than for MES, but that increase is offset by large decreases in expected size for large $|\theta|$. None of the methods dominates the rest for all $\theta \in \Theta$; other considerations are needed to choose among them.

Table II 95% confidence intervals for $\sin 2\beta$ using each of the four methods.

Method	Lower	Upper
TS	-0.07	1.00
LRT	-0.07	1.00
MES	0.00	1.00
MR	-0.08	1.00

The bounded normal mean problem arises in particle physics: Affolder et al. [2000] estimate the violation of charge-conjugation parity (CP) using observations of proton-antiproton collisions in the CDF detector at Fermilab. The parameter that measures CP violation is called $\sin 2\beta$, which must be in the interval $[-1, 1]$. In the model Affolder et al. [2000] use, the MLE of $\sin 2\beta$ has a Gaussian distribution with mean $\sin 2\beta$ and standard deviation 0.44. This standard deviation captures both systematic and random error in the estimate. This is equivalent to the situation described above, with $\tau = 1.0/0.44 \approx 2.27$. The observed measurement was 0.79, and the 95% confidence intervals for $\sin 2\beta$ are shown in table II. These results illustrate the strange behavior of MES in some cases: Since the LFA concentrates its mass on zero, the interval will always include a parameter value arbitrarily close to zero. Figure 2 compares acceptance regions and intervals for the four methods in this case. Note that for MES, the acceptance regions always extend to either $-\infty$ or $+\infty$.

2.2. Estimating a Function of Two Normal Means: An Example in Psychophysics

Suppose $\{X_{ij}\}_{i=1}^{2}{}_{j=1}^{n_j}$ are independent, normally distributed with variance one, that the expected values of $\{X_{1j}\}$ are all μ_1 and that the expected values of $\{X_{2j}\}$ are all μ_2. We know *a priori* that $-b \le \mu_2 \le \mu_1 \le b$. The goal is to estimate the two parameters $\theta_1 = 0.5(\mu_1 + \mu_2)$ and $\theta_2 = \mu_1 - \mu_2$ from observing the signs of $\{X_{ij}\}$. Thus,

$$\Theta \equiv \{(\eta_1, \eta_2) \in \Re^2 \colon -b \le \eta_1 - \eta_2/2 \le \eta_1 + \eta_2/2 \le b\}. \tag{9}$$

Let

$$Y_i \equiv \sum_j 1_{\{X_{ij} \ge 0\}} \quad i = 1, 2. \tag{10}$$

These observable variables are sufficient statistics for the signs of $\{X_{ij}\}$; they are independent; and Y_i has the binomial$(n_i, p_i \equiv \Phi(\mu_i))$ distribution, where Φ is the standard normal cumulative distribution function. We call (p_1, p_2) the *canonical parameters* because of their simple relationship to the distribution of the observations.

This is a stylized version of an estimation problem in signal detection theory [Miller 1996, Kadlec 1999]. A subject is presented with a randomized sequence of noisy auditory stimuli, and is asked to discern which stimuli contain "signal," and which are only "noise." In the standard model, the subject is assumed to have an internal scoring mechanism that assigns a number to each stimulus. If the number is positive, the subject reports that the stimulus contains signal; otherwise, the subject reports that the stimulus is just noise. Moreover, according to the model, scores for different stimuli are independent normal random variables with variance one.

For stimuli that consist of signal and noise, the expected scores are all equal to μ_1, while for stimuli that contain just noise, the expected scores are all equal to μ_2. The quantity of greatest interest is θ_2, the difference between these means, denoted d' in the psychophysics literature. It is a measure of the distance between the distributions of scores with and without noise, and (indirectly) provides an upper bound on the accuracy of signal detection. Of secondary interest is θ_1, the midpoint of the two means, which measures the "bias" in the decision rule: When $\theta_1 = 0$, so that $\mu_1 = -\mu_2$, the chance of claiming that the stimulus contains signal when it does not is equal to the chance of claiming that the signal is just noise when it contains signal. When $\theta_1 > 0$, the subject is biased in favor of claiming that the stimulus contains signal; when $\theta_1 < 0$, the subject is biased in favor of claiming that signal is not present. The restriction $-b \le \mu_2 \le \mu_1 \le b$ derives from the assumption that $\epsilon \le p_2 \le p_1 \le 1 - \epsilon$: The subject is more likely to report that signal is present when it is in fact present, and the subject has a strictly positive chance of misclassifying both types of stimuli. The constraints are related through $b = \Phi^{-1}(1 - \epsilon)$.

2.3. Confidence Regions for (θ_1, θ_2)

We compare methods for obtaining a $1 - \alpha$ confidence region for the parameter vector (θ_1, θ_2). Starting with a "good" confidence region for (p_1, p_2) and then finding its preimage in (θ_1, θ_2) space tends to produce unnecessarily large confidence regions for (θ_1, θ_2) because of the nonlinear relationship between these parametrizations. This distinction between the canonical parameters and the parameters of interest is crucial: We want the confidence region for models to constrain the values of the parameters of interest as well as possible. Whether that region corresponds to a small set of canonical parameters is unimportant.

The first approach we consider is based on the normal approximation to the distribution of the maximum likelihood estimator (MLE). For large enough samples, the MLE is approximately normally distributed with mean $\theta = (\theta_1, \theta_2)$ and covariance matrix

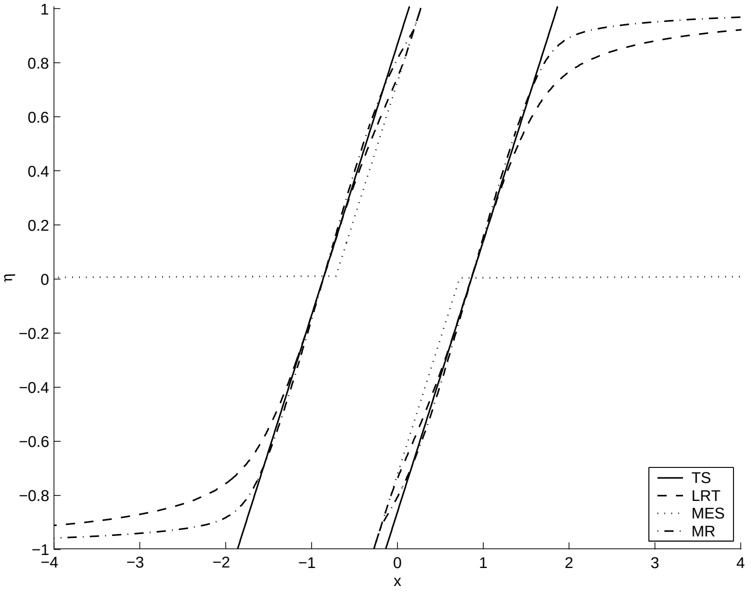


Figure 2: A depiction of 95% confidence intervals for an application of the bounded normal mean problem, the estimation CP violation parameter $\sin 2\beta$. Read across to see the acceptance region $A(\eta)$ for each of the four confidence procedures. Vertical sections are confidence intervals for different data values.

$\mathbf{I}^{-1}(\theta)$, where $\mathbf{I}(\theta)$ is the *Fisher information matrix* [Bickel and Doksum 1977]. In this case,

$$\mathbf{I}(\theta_1, \theta_2) = \begin{bmatrix} w_1 + w_2 & 0.5(w_1 - w_2) \\ 0.5(w_1 - w_2) & 0.25(w_1 + w_2) \end{bmatrix}, \quad (11)$$

where

$$w_i \equiv \frac{n_i \phi^2(\mu_i)}{p_i(1-p_i)}, \quad i = 1, 2, \quad (12)$$

and ϕ is the standard normal density. We can use this asymptotic distribution and the constraint to construct an approximate confidence region for (θ_1, θ_2) by intersecting Θ with an ellipse centered at the MLE. The light gray truncated ellipse in Figure 3 is an approximate 95% confidence region formed using this method. In this case, $n_1 = n_2 = 10$, the observed data are $y_1 = 8$ and $y_2 = 4$, and the bound b is $\Phi^{-1}(.99)$.

Figure 3 also illustrates the MES confidence region. The regular grid of points is the set of parameter values tested; those accepted are plotted as larger dots than those rejected. The MES region is the convex hull of the accepted parameter values. Table III compares the expected size of the confidence regions for these two procedures, along with LRT and MR, for various values of (θ_1, θ_2). MES has the smallest maximum expected size over this sample of parameter values, but small expected size for large θ_2 comes at the cost of increased expected size when θ_2 is small. TS is dominated by the others; there is no clear choice among the other three procedures.

Table III Expected sizes of four approximate 95% confidence regions for the parameter θ_1 in the psychophysics example in section 2.2: truncating the confidence ellipse based on the asymptotic distribution of the MLE (TS), inverting the likelihood ratio test (LRT), minimax expected size (MES), and minimax regret (MR).

$\lvert\boldsymbol{\theta_1}\rvert$	$\boldsymbol{\theta_2}$	**TS**	**LRT**	**MES**	**MR**
0.00	0.00	1.87	1.55	2.42	1.57
0.00	1.50	3.78	3.20	2.68	2.97
0.00	3.00	5.49	3.13	2.84	3.08
0.00	4.50	6.32	2.61	2.73	2.63
1.00	0.00	2.40	1.69	2.58	1.94
1.00	1.00	3.05	2.45	2.68	2.52
1.00	2.00	3.77	2.68	2.68	2.55
2.00	0.00	2.96	1.37	2.48	1.72
2.00	0.50	2.96	1.49	2.50	1.80

3. SOME THEORY

This section presents some of the theory behind MES and MR informally; see also Evans et al. [2003] for a more rigorous and general treatment of MES.

Consider the following estimation problem. The compact set Θ, a subset of $\Re^p$, is the set of possible states of nature—the possible values of an unknown parameter θ. For each $\theta \in \Theta$, there is a distribution P_θ on the space of possible observations $\mathcal{X} = \Re^m$; X is a random variable with distribution P_θ; and x is a generic observed value of X. Each distribution P_θ has

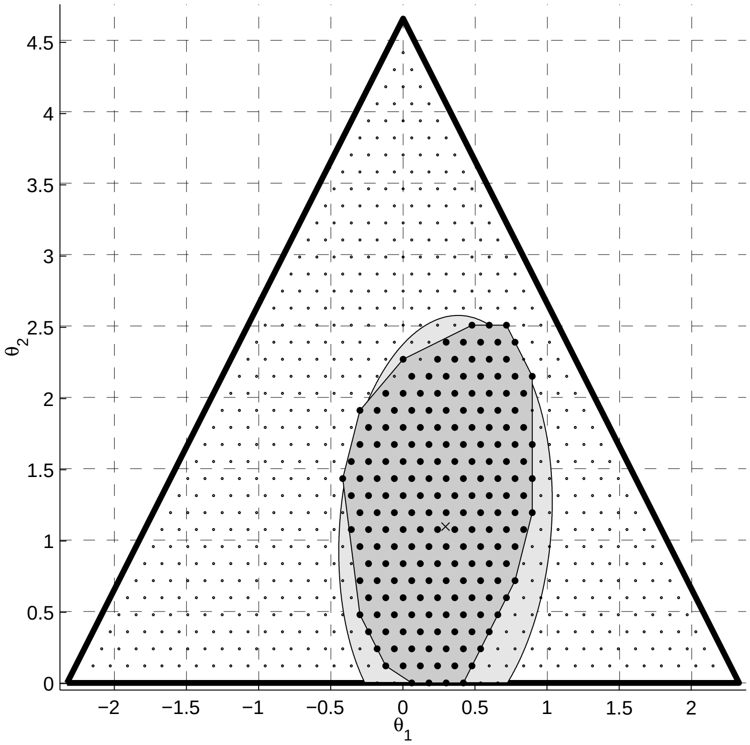


Figure 3: Approximate 95% confidence sets for an estimation problem in psychophysics. In this case $n_1 = n_2 = 10$, $y_1 = 8$, and $y_2 = 4$. The light gray truncated ellipse is a confidence region found using the asymptotic approximation to the distribution of the MLE. The "x" in the center of the ellipse is the MLE. The dots in the grid are the parameter values considered by MES. The larger dots are accepted values; the smaller are rejected. The darker, irregular region is the convex hull of these accepted parameter values, the MES confidence set.

a density $f(x|\theta)$ relative to Lebesgue measure; $f(x|\theta)$ is jointly continuous in x and θ.[1] We seek a confidence set for θ based on the observation $X = x$ and the *a priori* constraint $\theta \in \Theta$.

First consider testing the hypothesis $\theta = \eta$ at level α for an arbitrary fixed value $\eta \in \Theta$. Let $A(\eta)$ be the acceptance region of the test—the set of values $x \in \mathcal{X}$ for which we would not reject the hypothesis. Because the significance level of the test is α,

$$P_\eta\{X \in A(\eta)\} \geq 1 - \alpha. \tag{13}$$

The *power function* β of the test is the chance that the test rejects the hypothesis $\theta = \eta$ when in fact $\theta = \zeta$:

$$\beta(\zeta, \eta) \equiv 1 - P_\zeta\{X \in A(\eta)\}. \tag{14}$$

Because A_η has significance level α, $\beta(\eta, \eta) \leq \alpha$. Subject to that restriction, when testing a particular alternative hypothesis $\theta = \zeta$, it is natural to choose $A(\eta)$ to maximize $\beta(\zeta, \eta)$. Such a test is *most powerful (against the alternative $\theta = \zeta$)*. The following classical result characterizes the most powerful test in this situation.

The Neyman-Pearson Lemma: For fixed η, the acceptance region of the level α test that maximizes

$$\int_\Theta \beta(\zeta, \eta)\, \pi(d\zeta) \tag{15}$$

for an arbitrary measure π on Θ is

$$A_\pi(\eta) \equiv \{x \colon T_\pi(\eta, x) \geq c_\eta\}, \tag{16}$$

where

$$T_\pi(\eta, x) \equiv \frac{f(x|\eta)}{\int_\Theta f(x|\zeta)\pi(d\zeta)}, \tag{17}$$

with c_η chosen so that $\beta(\eta, \eta) = \alpha$.

The acceptance region $A_\pi(\eta)$ defined in equation 16 plays a crucial role in constructing optimal confidence sets. The set

$$C_A(x) = \{\eta \in \Theta : x \in A(\eta)\} \tag{18}$$

of all η that are accepted at significance level α is a $1 - \alpha$ confidence region for θ based on the observation

[1] This discussion assumes X is continuous. For X discrete, we could introduce an independent, uniformly distributed random variable U observed along with X. This is equivalent to considering randomized decision rules. See Evans et al. [2003] for more rigor.

$X = x$. We want to minimize the expected ν-measure of the confidence region $C_A(X)$ by choosing the acceptance regions $A(\eta)$ well. The measure ν on the parameter space Θ can be essentially arbitrary, but it needs to be defined on a broad enough class of subsets of Θ that $C_A(x)$ is ν-measurable for any value of x. In applications, ν is typically Euclidean volume.

The following theorem is due to Pratt [1961].

Pratt's Theorem:

$$\mathbf{E}_\zeta[\nu(C_A(x))] = \int_\Theta (1 - \beta(\zeta, \eta))\, \nu(d\eta), \tag{19}$$

where $\mathbf{E}_\zeta[\cdot]$ is expectation when $\theta = \zeta$, and $\beta(\cdot, \cdot)$ is the power function of the family of tests $\{A_\eta\}$ corresponding to the confidence set C_A.

Pratt's theorem links maximizing the power function β to minimizing the expected size of the confidence region $C_A(X)$. The following result combines the Neyman-Pearson Lemma and Pratt's Theorem.

Corollary: The confidence set C_A that minimizes

$$\int_\Theta \mathbf{E}_\zeta[\nu(C_A(X))]\, \pi(d\zeta) \tag{20}$$

is C_{A_π}.

What is the role of the measure π? The following is proved in great generality in Evans et al. [2003].

Theorem [Evans et al. 2003]: There exists a measure λ on Θ such that the acceptance regions A_λ give the confidence procedure that minimizes

$$\max_{\theta \in \Theta} \mathbf{E}_\theta[\nu(C_A(X))]\,. \tag{21}$$

This is MES, and λ is referred to as the *least favorable alternative* because the alternative defined by λ maximizes the Bayes risk (see section 5).

This result can be adapted to show that there is another measure μ on Θ for which C_{A_μ} is the minimax regret procedure. Determining these priors exactly is not computationally feasible except in simple cases. Section 5 sketches an efficient method to approximate λ and μ numerically.

4. CMB DATA ANALYSIS

The cosmic microwave background radiation (CMB) consists of redshifted photons that have travelled since the *time of last scattering*, approximately 300,000 years after the Big Bang, when the Universe had cooled enough to allow atoms to form and photons to travel freely. The small fluctuations in the temperature of the CMB are the signature of the primordial variability that led to the structure visible in the Universe today, such as galaxies and clusters of galaxies. Theoretical research connects unknown physical constants that characterize the Universe—such as the fraction of ordinary matter in the Universe, the fraction of dark matter in the Universe, Einstein's cosmological constant, Hubble's constant, the optical depth of the Universe, and the spectral index—to the angular distribution of the fluctuations. See chapter two of Longair [1998] for an introduction.

Estimating these cosmological parameters from observed CMB fluctuations is conceptually similar to the example given in section 2.2. The physically interesting parameters are the cosmological parameters, while the canonical parameter is the angular power spectrum of the CMB. The data are assumed to be a realization of a normally distributed vector with mean zero and covariance matrix

$$\mathbf{N} + \sum_\ell \left(\frac{2\ell + 1}{4\pi}\right) C_\ell(\theta) B_\ell^2\, \mathbf{P}_\ell, \tag{22}$$

where $\mathbf{N}$ is the measurement error covariance matrix (which is assumed to be known), $\{C_\ell(\theta)\}$ is the CMB power spectrum for the cosmological parameter vector θ, $\{B_\ell\}$ is the transfer function resulting from the beam pattern of the observing instrument, and $\mathbf{P}_\ell$ is a matrix whose (i, j) entry is the degree ℓ Legendre polynomial evaluated at the cosine of the angle between pixel i and pixel j. This representation is based on the spherical harmonic decomposition of a spherical, isotropic Gaussian process model for the CMB. The software package CMBFAST [Seljak and Zaldarriaga 1996] is the standard for calculating the spectrum from cosmological parameters; the nonlinearity of this mapping is a major complication in this problem.

Table IV lists the parameters we use and their *a priori* bounds, based on Abroe et al. [2002]. Figure 4 shows the data: the 5,972 observations in the MAXIMA-1 8 arcminute resolution data set [Hanany et al. 2000]. We compress the data to 2,000 linear combinations of the original observations, then form 95% MES and MR joint confidence regions for the parameters. Figure 5 shows the MES confidence set in the spectral domain. A total of 1,000 models were tested; 35 were accepted. (Generating spectra from the randomly selected parameter vectors is computationally expensive. These results are preliminary: We plan to test more models in the future.) Their spectra are the heavier curves in the figure. The lighter curves are spectra of 300 of the rejected models. The dark band is an approximate 95% confidence region for the angular power spectrum of CMB fluctuations.

The parameter values for each of the 1,000 tested spectra are known: Table V lists 15 the 35 accepted vectors along with the minimum and maximum accepted values for each parameter. For example, all the accepted values of the total energy density relative to the critical energy density, $\Omega = \Omega_m + \Omega_\Lambda$,

Table IV Cosmological parameters and their bounds, following Abroe et al. [2002]. The parameters also must satisfy $\Omega_b \le \Omega_m$ and $0.6 \le \Omega_m + \Omega_\Lambda \le 1.4$.

Parameter (Symbol)	**Lower**	**Upper**
Total Matter (Ω_m) †	0.05	1.00
Baryonic Matter (Ω_b) †	0.005	0.15
Cosmological Constant (Ω_Λ) †	0.0	1.0
Hubble Constant (H_0) (km s^{-1} Mpc^{-1})	40.0	90.0
Scalar Spectral Index (n_s)	0.6	1.5
Optical Depth (τ)	0.0	0.5

† Relative to critical density.

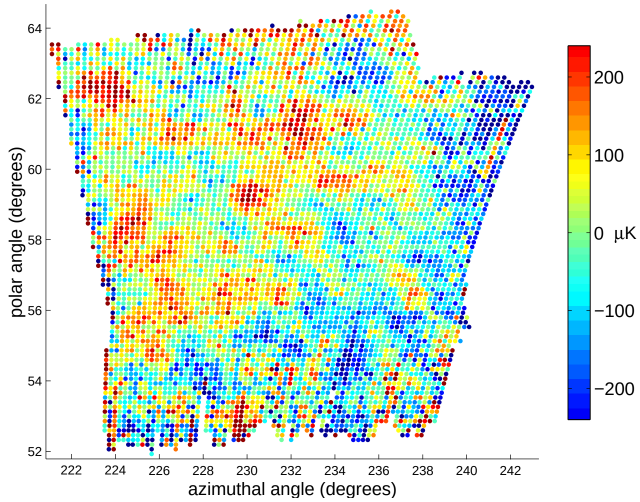


Figure 4: The MAXIMA-1 data set used in this analysis. There are 5,972 pixels at 8 arcminute resolution.

are between 0.915 and 1.334. The MAXIMA-1 experiment has much higher resolution than previous experiments, but it still does not constrain most of the parameters individually, owing partly to tradeoffs among the parameters. (Our data compression also might contribute to the uncertainty; we have not yet explored the sensitivity to the compression scheme.)

From a frequentist viewpoint, the fact that there is a parameter vector that accounts adequately for the data (that is accepted) and which has $\Omega = 1.334$ means that we cannot rule out the possibility that $\Omega = 1.334$ at significance level 0.05. Bayesian techniques make inferences starting with the marginal posterior distribution for each parameter by itself: Whether the posterior credible region includes $\Omega = 1.334$ depends on the posterior weight assigned to the set of *all* models with $\Omega = 1.334$. That weight, in turn, depends on the prior as well as the data.

Table V Fifteen of the 35 cosmological parameter vectors accepted by MES. The final two rows list the minimum and maximum accepted values of each parameter.

Ω_b	Ω_m	Ω_Λ	τ	H_0	n_s	Ω
0.042	0.674	0.241	0.317	77.00	1.117	0.915
0.078	0.368	0.632	0.161	69.71	0.834	1.000
0.088	0.786	0.214	0.445	68.65	1.151	1.000
0.131	0.860	0.176	0.417	67.07	1.027	1.036
0.081	0.540	0.526	0.000	77.15	0.809	1.066
0.079	0.321	0.773	0.364	69.35	1.002	1.094
0.134	0.940	0.161	0.466	66.83	1.038	1.101
0.101	0.699	0.482	0.000	44.68	0.833	1.181
0.089	0.425	0.771	0.217	77.85	0.896	1.196
0.130	0.591	0.635	0.364	43.03	0.944	1.226
0.085	0.994	0.243	0.315	76.79	1.081	1.237
0.096	0.555	0.693	0.260	81.28	0.923	1.248
0.093	0.708	0.551	0.000	76.96	0.855	1.259
0.139	0.667	0.623	0.269	61.15	0.954	1.290
0.133	0.692	0.642	0.068	41.47	0.846	1.334
0.011	0.058	0.082	0.000	41.47	0.729	0.915
0.139	0.994	0.988	0.466	89.24	1.151	1.334

Figure 5 also plots error bars given by Hanany et al. [2000], based on their analysis of the MAXIMA-1 data. The error bar at $\ell = 223$ extends far above all the accepted spectra. Close inspection shows that each accepted spectrum passes either through the bar at $\ell = 147$ or through the bar at $\ell = 300$. None of the 1,000 spectra (including the 665 spectra that are not plotted) passes through all three of these bars. This shows the fundamental problem with the "chi-by-eye" procedure for comparing spectra with error bars: It is not clear how well the spectra should fit the bars, especially when estimates at different frequencies are dependent, as they are here. MES allows more precise comparisons, and maximizes the power of the tests in the sense described in section 3. The MR results, shown in Figure 6, are similar but only 25 spectra are accepted. Figures 5 and 6 also show the best fitting model based on the recent WMAP experiment [Bennett et al. 2003], which has much higher resolution than MAXIMA-1. At low ℓ, the WMAP model is quite similar to the models accepted by MES and MR using the MAXIMA-1 data.

5. APPROXIMATING THE LFA

The LFA λ is the measure π on Θ that maximizes

$$\mathbf{B}(\pi) \equiv \int_\Theta \mathbf{E}_\zeta[\nu(C_{A_\pi}(X))]\,\pi(d\zeta)\,. \qquad (23)$$

This is an instance of Bayes/minimax duality: The

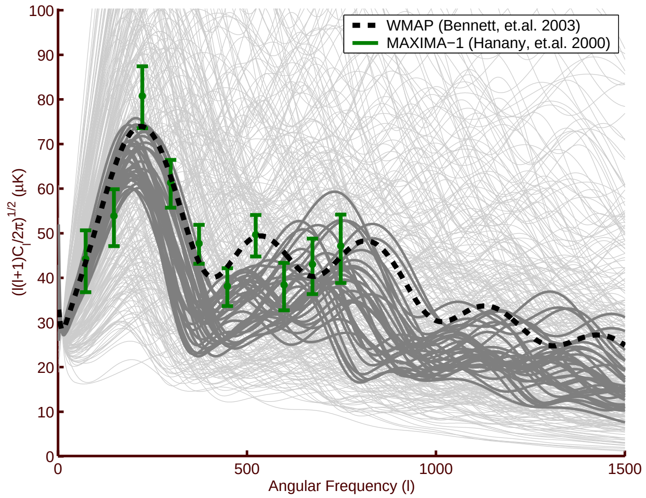


Figure 5: The 35 accepted spectra (dark curves) and 300 of the 965 rejected spectra (light curves) from the MES procedure applied to 8 arcminute MAXIMA-1 data. The vertical bars are Bayesian error bars based on the MAXIMA-1 data [Hanany et al. 2000]; the dashed curve is the best fitting model to the WMAP data [Bennett et al. 2003].

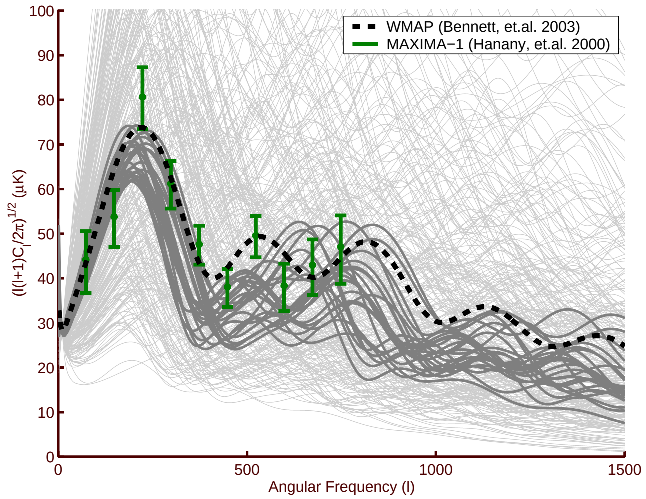


Figure 6: The 25 accepted spectra (dark curves) and 300 of the 975 rejected spectra (light curves) from the MR procedure applied to the 8 arcminute MAXIMA-1 data. The vertical bars are Bayesian credible intervals based on the MAXIMA-1 data [Hanany et al. 2000]; the dashed curve is the spectrum of the model that fits the WMAP data best [Bennett et al. 2003].

"worst" prior λ corresponds to the minimax procedure. It is computationally impractical to determine the LFA explicitly in all but the simplest situations. The main difficulty is the complicated relationship between π and A_π, which makes it hard to evaluate equation 23. Nelson [1966] and Kempthorne [1987] propose computational methods for determining least favorable priors in general situations, but they assume that calculating the Bayes risk (equation 23) is a solved problem: It is not part of their algorithms.

The approach we use here is described in greater detail in Schafer and Stark [2003]. It involves two levels of numerical approximation. First, the support of the prior is restricted to a finite set of points. Second, Monte Carlo methods are used to estimate equation 23 for any prior π supported on this discrete set. Schafer and Stark [2003] show that as the size of the Monte Carlo simulations increases, the estimate of $\mathbf{B}(\pi)$ converges uniformly in π to $\mathbf{B}(\pi)$.

Let $\{\theta_i\}_{i=1}^p$ be the support points of the prior. Let $\{\eta_j\}_{j=1}^q$ be parameter values selected at random from the compact parameter space Θ according to the measure ν. For each η_j, simulate a set of data $x_{j1}, x_{j2}, \ldots, x_{jn}$. Construct matrices $\mathbf{A}_j$, $j = 1, 2, \ldots, q$, with (i, k) entry

$$\mathbf{A}_j(i,k) \equiv \frac{f(x_{jk}|\theta_i)}{f(x_{jk}|\eta_j)}. \tag{24}$$

We need to estimate $\mathbf{B}(\pi)$, for a prior π supported on $\{\theta_i\}_{i=1}^p$. The empirical distribution of the vector $\mathbf{A}_j\pi$ is an approximation to the distribution of the test statistic under the null hypothesis $\theta = \eta_j$. It can be used to estimate the threshold to form the acceptance region $A_\pi(\eta_j)$.

Choosing a decision rule can be thought of as picking a strategy in a zero-sum two-person game: Statistician versus Nature. The Statistician chooses a set of decision vectors $\mathbf{d}_j$, $j = 1, 2, \ldots, q$. Nature chooses π, a distribution over possible true values of the parameter. The Statistician pays Nature the approximate risk

$$\tilde{\mathbf{R}}(\mathbf{d}, \pi) \equiv \sum_{j=1}^{q} \mathbf{d}_j^T \mathbf{A}_j \pi. \tag{25}$$

The Statistician can choose $\mathbf{d}$ to minimize $\tilde{\mathbf{R}}(\mathbf{d}, \pi)$ by setting the component d_{jk} to one if $x_{jk} \in A_\pi(\eta_j)$ and to zero otherwise. Let $\mathbf{d}^\pi$ be that optimal decision function. Define

$$\tilde{\mathbf{B}}(\pi) \equiv \tilde{\mathbf{R}}(\mathbf{d}^\pi, \pi). \tag{26}$$

Then $\tilde{\mathbf{B}}(\pi)$ is an approximation of the Bayes risk for prior π. Maximizing $\tilde{\mathbf{B}}$ over π amounts to finding the (approximate) optimal strategy for Nature, the prior that maximizes the payout by the Statistician. This is a matrix game, and finding an optimal strategy is a well-studied problem. A fictitious play algorithm proposed by Brown and Robinson [Robinson 1951] works well here because it can handle the constraint on the Statistician's strategies that ensures $1 - \alpha$ coverage. Solving this matrix game for large problems is computationally expensive; typically the most costly steps are to simulate data from a randomly chosen parameter vector and to evaluate the likelihood function. An implementation in Fortran-90, runnable on parallel computers, is available at the URL `www.stat.berkeley.edu/∼stark/Code/LFA_Search`. This subroutine also can find an approximate LRA for MR.

6. CONCLUSION

Expected size is a useful measure of the performance of a confidence estimator. It is directly related to the power of the procedure to reject false parameter values; this is a natural property to maximize. Generally there is no estimator that minimizes expected size for all parameter values simultaneously: Some tradeoff must be imposed. Minimax expected size (MES) and minimax regret expected size (MR) trade off expected sizes at the possible parameter values optimally—in different senses. MES and MR are alternatives to the likelihood ratio test approach to confidence sets proposed by Feldman and Cousins [1998]. MES and MR incorporate bounds on parameters by minimizing the maximum expected size only over the set of parameters that satisfy these bounds, and by including only parameters within the bounds. MR is less conservative than MES. These regions typically cannot be calculated analytically, but they can be approximated numerically, and we provide a Fortran-90 subroutine.

MES and MR can be used to estimate cosmological parameters from observations of the cosmic microwave background radiation, incorporating bounds on the parameters to produce confidence sets that are small in expectation. MES and MR use the subroutine CMBFAST to map cosmological parameters to power spectra. They do not involve the complicated relationship between the parameters of interest and the canonical parameters explicitly. MES and MR can test cosmological models formally, avoiding potentially misleading "chi-by-eye" comparisons between spectra and spectrum estimates.

Acknowledgments

MAXIMA-1 data are courtesy of the MAXIMA collaboration. Analysis of CMB data was performed at NERSC using resources made available by Julian Borrill. We had many helpful conversations with Andrew Jaffe regarding microwave cosmology, and with Poppy Crum regarding signal detection problems in psychophysics.

References

G. J. Feldman and R. D. Cousins, Phys. Rev. D **57**, 3873 (1998).

G. Backus, Proc. Natl. Acad. Sci. **84**, 8755 (1987).

G. Backus, Geophys. J. **94**, 249 (1988).

M. Schervish, *Theory of Statistics* (Springer-Verlag, New York, 1995).

S. Evans, B. Hansen, and P. Stark, Tech. Rep. 617, Univ. of California, Berkeley (2003).

P. Bickel and K. Doksum, *Mathematical Statistics: Basic Ideas and Selected Topics* (Holden Day, San Francisco, 1977).

E. Lehmann, *Testing Statistical Hypotheses* (John Wiley and Sons, New York, 1986), 2nd ed.

M. DeGroot, in *Encyclopedia of Statistical Science*, edited by S. Kotz, N. Johnson, and C. Read (John Wiley and Sons, New York, 1988), vol. 8, pp. 3–4.

J. Pratt, J. Am. Stat. Assoc. **56**, 549 (1961).

T. Affolder, H. Akimoto, A. Akopian, M. Albrow, P. Amaral, S. Amendolia, D. Amidei, J. Antos, G. Apollinari, T. Arisawa, et al., Phys. Rev. D **61**, 072005 (2000).

J. Miller, Perception and Psychophysics **58**, 65 (1996).

H. Kadlec, Psychological Methods **4**, 22 (1999).

M. Longair, *Galaxy Formation* (Springer-Verlag, New York, 1998).

U. Seljak and M. Zaldarriaga, Astrophys. J. **469**, 437 (1996).

M. Abroe, A. Balbi, J. Borrill, E. Bunn, P. Ferreira, S. Hanany, A. Jaffe, A. Lee, K. Olive, B. Rabii, et al., Month. Not. Royal Astron. Soc. **334**, 1 (2002).

S. Hanany, P. Ade, A. Balbi, J. Bock, J. Borrill, A. Boscaleri, P. de Bernardis, P. Ferreira, V. Hristov, A. Jaffe, et al., Astrophys. J. Lett. **545**, L5 (2000).

C. Bennett, M. Halpern, G. Hinshaw, N. Jarosik, A. Kogut, M. Limon, S. Meyer, L. Page, D. Spergel, G. Tucker, et al., Astrophys. J. Suppl. **148**, 1 (2003).

W. Nelson, Ann. Math. Stat. **37**, 1643 (1966).

P. Kempthorne, SIAM J. Sci. Stat. Comput. **8**, 171 (1987).

C. Schafer and P. Stark (2003), In preparation.

J. Robinson, Ann. Math. **54**, 296 (1951).

Measures of Significance in HEP and Astrophysics

James T. Linnemann
Michigan State University, E. Lansing, MI 48840, USA and
Los Alamos National Laboratory, Los Alamos, NM 87545, USA

I compare and discuss critically several measures of statistical significance in common use in astrophysics and in high energy physics. I also exhibit some relationships among them.

1. INTRODUCTION

Significance testing for a possible signal in counting experiments centers on the probability that an observed count in a signal region, or one more extreme, could have been produced solely by fluctuations of the background source(s) in that region. Statisticians refer to this probability as a p-value. The traditions for calculating signal significance differ between High Energy Physics (HEP) and High Energy Gamma Ray Astrophysics (GRA). Both fields often quote significance in terms of equivalent standard deviations of the normal distributions (statisticians sometimes refer to this as a Z-value).

I will present several of the commonly used methods in HEP and GRA, apply them to examples from the literature, then discuss the results. Here I will concentrate on observed significance, the significance of a particular observation, rather than predictions of significance for a given technique as a function of exposure. The prediction problem is slightly different, involving the power of the test, or the probability of making an observation at a given significance level.

GRA has emphasized simple, quickly-evaluated analytical formulae for calculating Z directly (choosing asymptotically normal variables), while HEP has typically calculated probabilities (p-values) and then translated into a Z-value by $p = P(s \geq \text{observed} \mid \text{assume only background})$;

$$Z = \Phi^{-1}(p); \quad \Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2}\, dt$$

This relation can be written[1] for large $Z > 1.5$ as

$$Z \approx \sqrt{u - Ln\ u}; \quad u = -2Ln(p\sqrt{2\pi})$$

giving a rough dependence of $Z \sim \sqrt{-Ln\ p}$. While more general than the search for a simple formula for Z-values, the HEP approach loses track of the analytic structure of the problem.

Observations in GRA typically consist of a count of gamma rays when pointing directly at a potential source, called an on-source count, N_{on}. The analogous quantity in HEP is the number of counts in a signal region. The background relevant to an observation of a source is typically estimated in GRA by an off-source observation. The relative exposure of the two observations is denoted by $\alpha = T_{on}/T_{off}$, often less than unity. Then the background count mean's estimate is $b\ = \alpha N_{off}$, its (Poisson) uncertainty $\delta b = \alpha\sqrt{N_{off}}$, and thus one derives

$$\alpha = (\delta b)^2/b \tag{1}$$

GRA expressions are couched in terms of α. I will also use $x = N_{on},\ y = N_{off},\ k = x + y$ for compactness.

In HEP, sometimes a side-band method of background estimation is used, rather like in a GRA measurement; or b may be estimated as a sum of contributions from Monte Carlo and data-based side-band estimates, so that often $b \pm \delta b$ is quoted, where δb is derived from adding uncertainties in quadrature. One can use Eq.1 to *define* α when comparing HEP results with GRA expressions. Non-integer values for effective N_{off} result, but usually cause no problems.

2. Z-VALUE VARIABLES

Many expressions for Z are of the form of a ratio of estimates of signal to its variance, where the signal is estimated by $s = N_{on} - b = x - \alpha y$. Then $Z = s/\sqrt{V}$, where V is a variance estimate for s. A standard GRA reference[2] gives as an example (their Equation 5) $V_5 = N_{on} + \alpha^2 N_{off}$. The authors note that this expression treats N_{on} and N_{off} as independent; this does not consistently calculate V under the null hypothesis, $\mu_{on} = \alpha\mu_{off}$ and in fact biases against signals for $\alpha < 1$ by overestimating V. I have derived a related formula, $V_{5'} = \alpha(1 + \alpha)N_{off}$, by using only the background to estimate the mean and variance: while not optimal, it at least is consistent with the null. They also provide $V_9 = \alpha(N_{on} + N_{off})$, which better implements the null hypothesis. However, their widely-used recommendation is likelihood ratio $L(\mu_s, \mu_b)/L(\mu_b)$,

$$Z_L = \sqrt{2}\ (\ x\ Ln\frac{x(1+\alpha)}{k\alpha} + y\ Ln\frac{y(1+\alpha)}{k}\ )^{\frac{1}{2}}.$$

Z_L derives from the standard likelihood ratio test for a composite hypothesis, and Wilks' Theorem, giving its asymptotic normal behavior. The numerator and denominator likelihoods are each separately maximized:

one for a signal + background model, the other for a background-only (null) model.

One may instead seek an asymptotically normal variable with nearly constant variance[3],

$$Z_0 = \frac{2}{\sqrt{1+\alpha}}(\sqrt{x+3/8} - \sqrt{\alpha(y+3/8)}\ ).$$

The 3/8 speeds convergence to normality from the underlying discreteness.

2.1. Other Frequentist Methods

One widely used form is $Z_{sb} = s/\sqrt{b}$ (sometimes[4] called the "signal to noise ratio"). This entirely ignores the uncertainty in the background estimate. It is often used for optimizing selection criteria, because of its simplicity. Slightly better is a Z_P calculated from the Poisson probability p-value:

$$p_P = P(\geq x|b) = \sum_{j=x}^{\infty} e^{-b}b^j/j! = \Gamma(x,0,b)/\Gamma(x).$$

here written[6] in terms of an incomplete Γ function. Z_P still ignores uncertainty in b. Occasionally one sees substitutions of $b \to b + \delta b$ as a feeble attempt to incorporate the uncertainty in b.

Finally, one may view a significance calculation directly as a p-value calculation which one could use as a test of the null hypothesis. Z_L use the standard (non-optimal) test of a composite hypothesis against a null. However, the relationship of the Poisson means, whether $\mu_{on} > \alpha\mu_{off}$, is a special case of a composite hypothesis test that admits a more optimal solution. There exists a Uniformly Most Powerful test among the class of Unbiased tests for this case, in the form of a binomial proportion test for the *ratio* of the two Poisson means[5]. The UMPU properties are, strictly speaking, derived only with an assumption of randomization, that is, hiding the underlying discreteness by adding a random number to the data. This test yields a binomial probability p-value (using $k = x + y$):

$$p_{Bi} = P_{Bi}(\geq x|\ w,k) = \sum_{j=x}^{k} \frac{k!}{j!(k-j)!} w^j (1-w)^{k-j},$$

where $w = \alpha/(1+\alpha)$ is the expected ratio of the Poisson means for x and $x+y$. After some manipulation, this can be written in terms of incomplete and complete beta functions[1, 6], which is convenient for numerical evaluation:

$$p_{Bi} = B(w, x, 1+y)/B(x, 1+y)$$

This test is conditional on $x + y$ fixed because of the existence of a nuisance parameter: there are two Poisson means, but the quantity of interest is their ratio. While this test is known to both the GRA[3] and HEP[7] communities, it is common practice in neither, and its optimality properties are not common knowledge.

Given the (restricted) optimality of the test, and the lack of a UMP test for this class of composite hypotheses, this test ought to be more frequently used to calculate significance, even though it is clearly a longer calculation than Z_L. For moderate x, y, closed forms in terms of special functions are available, while some care is required for larger n. For $Z_B < 3$, the Z-values reported may be somewhat too small[3, 8], but for typical applications one is more interested in $Z_B > 4$.

It is interesting to note that taking a normal approximation to the binomial test (that is, comparing the difference of binomial proportion from its expected value, to the square root of its normal-approximation variance) yields $(x/k - w)/\sqrt{w(1-w)/k}$, which can be shown to be identical to $Z_9 = s/V_9$.

A different approach attempts to move directly from likelihood to significance by using a 3rd-order expansion[9]. The mathematics is interesting, combining two first order estimates (which give significance to order $1/\sqrt{n}$) to yield a $1/\sqrt{n^3}$ result. Typically, the first-order estimates are of the form of a normal deviation, Z_t (like Z_9), and a likelihood ratio like Z_L; of these, the likelihood ratio is usually a better first-order estimate. The two are then combined into the third order estimate by a formula such as

$$Z_3 = Z_L + \frac{1}{Z_L} Ln(Z_t/Z_L).$$

Generically, $Z_t = \Delta/\sqrt{V}$ is a Student t-like variable, where Δ is the difference of the maximum likelihood value of θ (the parameter of interest) from its value under the null hypothesis, and V is a variance estimate derived from the Fisher Information $\partial^2 L/\partial^2\theta$. The attraction of the method is to achieve simple formulae with accurate results. However, the mathematics becomes more complex[10] when nuisance parameters are included, as is needed when the background is imperfectly known. Here I will only compare the approximate calculation for a perfectly known background to the corresponding exact calculation, p_P.

3. BAYESIAN METHODS

HEP common practice often involves Bayesian methods of incorporating "systematic" uncertainties for quantities such as efficiencies[11]. These methods are also used for calculating significance, particularly when the background b is a sum of several contributions, since the method naturally extends to complex situations where components of δb are correlated. The typical calculation represents the lack of knowledge of b by a posterior density function $p(b|y)$; it is referred to as a posterior density because it is posterior to the off-source measurement y. The usual way of proceeding is to calculate Poisson p-values $p_P = P(\geq x|b)$ as was done above, but this time taking into account the uncertainty in b by performing an average of p-values weighted by the Bayesian posterior $p(b|y)$, that is

$$p_{Ba} = \int p_P(\geq x|b)\; p(b|y)\; d\,b.$$

This can be evaluated by Monte Carlo integration, or by a mixture of analytical and numerical methods. I will pursue the latter course here. The most common usage in HEP is to represent $p(b|y)$ as a truncated normal distribution

$$p_N(b|y) = \frac{1}{\delta b\sqrt{2\pi}}\exp\frac{-(b-\alpha y)^2}{2(\delta b)^2},\ b > 0.$$

If b is a sum of many contributions, its distribution should asymptotically approach a normal. An alternative I have advocated in HEP[12], and which is also known to the GRA communitity[13], is to start from a flat prior for b and derive the $p(b|y)$ in the usual Bayesian fashion, leading to a Gamma posterior:

$$p_\Gamma(b|y) = \beta^y e^{-\beta}/y!\ ,\ \beta = b/\alpha.$$

This is most appropriate when a single contribution to b dominates and its uncertainty is actually due to counting statistics. I will refer to the Z-values which result from these two choices as Z_N for the normal posterior, and Z_Γ for the Gamma function posterior. Choosing to represent p_P as a sum, and performing the b integration first gives the p-value for the Gamma posterior[13]

$$p_\Gamma = \sum_{j=x}^{\infty}\frac{(y+j)!}{j!y!}\frac{\alpha^j}{(1+\alpha)^{1+y+j}}.$$

Despite appearances, p_Γ is identical to p_{Bi}. The Beta function representation of p_{Bi} is much more suitable for large values of $x,\ y$. The two expressions can be made somewhat closer by using $w = \alpha/(1+\alpha)$.

Bayesian practice typically focuses on direct comparison of specific hypotheses through the odds ratio. However, predictive inference[14] is commonly used in model checking (significance testing is just checking the background-only model). Predictive inference in our case is directly related to calculating $p(x|y)$, that is, averaging over the unknown parameter b.

$$p(j|y) = \int\ p(j|b)\ p(b|y)\ d\,b$$

Interestingly, some Bayesian practitioners go farther, and are willing to calculate a “Bayesian p-value”[14],

$$p_{Bayes} = \sum_{j=x}^{\infty} p(j|x)$$

which is precisely the p_{Ba} given above (there we summed before integrating).

4. COMPARISON OF RESULTS: RELATIVE PERFORMANCE

I have taken several interesting test cases from the HEP and GRA literature. The input values and Z-value calculation results are shown in Table 1. For the HEP cases, the values reported in the papers are N_{on}, b, and δb, while in the GRA case, the reported values are N_{on}, N_{off}, and α. I have also included a few artificial cases in order to sample the parameter space reasonably.

It is worth remarking that there are numerical issues to be faced in evaluation of the more complex methods. These remarks apply–at a minimum–to a Mathematica implementation. The Binomial is straightforward in its Beta function representation. The Bayes p-value methods may involve an infinite sum, and are touchy and slow for large n; [13] suggests approximating the summation by an integral. Fraser-Reid and the Bayes p-value summation results may be sensitive to whether integers are floating point values are used. An alternative attack is to leave the p_P as a Γ function ratio and trade an integration for the infinite sum. Doing so in the Bayes Gaussian case is less unstable than summing, but for large n requires hints on the location of the peak of the integrand.

For the purposes of the present section, I will take the Frequentist UMPU Binomial ratio test as a reference standard, because of its optimality properties. I will have more to say on this later.

None of these examples from the recent literature was published with a seriously wrong significance level. To me, the most striking result in the table is that the Bayes Gamma prior method produces results *identical* to the Binomial result (MSU graduate student Hyeong Kwan Kim has proven the identity).

The method most used in HEP, Bayes with a normal posterior for b, produces Z’s always larger than those from Bayes Gamma. Viewing the calculation as averaging the Poisson p-value $p_P(b)$ over the posterior for b, the shorter tails of the normal compared to the gamma place less weight on the larger probabilities (smaller p-values) obtained when the off-source measurement happens to underestimate the true value of b. The difference is most striking for large values of α, that is, when the background estimate is performed with less sensitivity than the signal estimate; in this case, results differing in significance by over .5 σ can occur. The most common method in GRA, the simple Log Likelihood ratio formula, produces comparable or slightly higher estimates of significance, but seems less vulnerable to problems at large α. It appears to claim the highest significance of these methods at small n. The variance stabilization method Z_0 presented in [3] does not appear to be in general use in GRA, but produces results of similar quality to the other two mainstay methods. All methods agree for $N > 500$, where the normal approximations are good, even out to 3-6 σ tails.

The "not recommended" methods all produce results off by more than .5 σ for several low-statistics cases. Z_9, which approximates Z_{Bi}, does best; Z_5 is indeed biased against real signals compared to other measures, and its alleged improvement $Z_{5'}$, while curing that problem, overestimates significance as the price for its less efficient use of information compared to Z_9.

As expected, ignoring the uncertainty in the background estimate leads to overestimates of the significance. $s/\sqrt{b}$ is much more over-optimistic than an exact Poisson calculation, particularly for small n, or $\alpha > 1$, where the background uncertainty is most important. The best that can be said for $s/\sqrt{b}$ is that it is mostly monotonic in the true significance, at least as it is typically used (for comparing two selection criteria with N varying by an order of magnitude at most). The 3rd order Fraser-Reid approximation is fast and accurate up to moderate n, suggesting it is worth pursuing the full nuisance parameter case. However, the approximation fails for one large Z, and is very slow for the largest n.

Of the ad-hoc corrections for signal uncertainty, none are reliable; the "corrected" Poisson calculation is less biased than the un-corrected, but still widely overestimates significance for $\alpha > 1$, and can't be used for serious work. The $s/\sqrt{b + \delta b}$ isn't much better than its "un-corrected" version.

To summarize, most bad formulae overestimate significance (the only exceptions are Z_5 for $\alpha < 1$ and Poisson with $b \to b + \delta b$). Thus, prudence demands using a formula with good properties. The Binomial test seems best for simple Poisson backgrounds. For backgrounds with several components, compare Bayes MC with Γ or Normal posteriors.

5. CALIBRATION OF ABSOLUTE SIGNIFICANCE: MONTE CARLO

In the previous section, results of significance calculations were compared to a reference calculation, the UMPU Binomial Test. That method produces the lowest reported significance among the methods with a sound theoretical basis. This alone could justify its use (on grounds of conservatism)[3], but would beg the question of whether the Binomial test is actually "correct." This has been studied by Monte Carlo simulation[1] in [3].

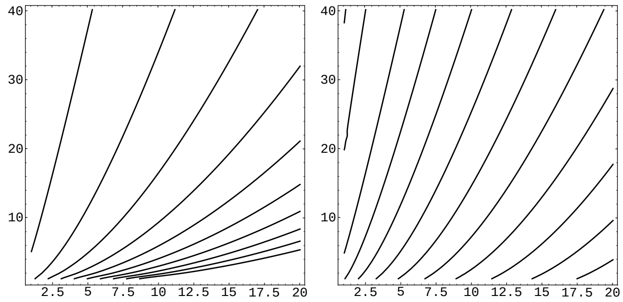


Figure 1: Contours of equal Z, case [18], for Z_{SB} (left) and Z_L (right).

A few observations on MC testing are useful. One might imagine simply generating instances of Poisson variables x, y with means μ, μ/α, and calculating Z^{MC} from p^{MC} = the fraction of events "more signal-like" than (N_{on}, N_{off}). Instead, [2, 3] a separate MC is done *for each individual measure*, because there is no unique "correct" Z-value for a given observation. The best that can be done is to ask that a method produce a Z value consistent with MC probabilities when the observation is analyzed by that method. The problem is that there is no unique definition of "more signal-like". One is essentially trying to find a unique ordering of points on the x, y plane to define those which are similarly far from the observed point N_{on}, N_{off}.

Each variable introduces its own metric, and contours of equal Z do not coincide for different Z variables, as seen in Figure 1.

The p-value for an observation (x_0, y_0) depends on these contours:

$$p^{MC}(x_0, y_0) = \int_{Z>Z_0} p(x, y)\ dx\ dy$$

where the integration is over the region beyond the contour line Z_0 passing through the observation: $Z(x, y) > Z_0(x_0, y_0)$.

For small n, the contours are markedly different, so that two *different* Z-values could both be correct if each agreed with their respective Z^{MC}. Still, the situation is not catastrophic, as values of Z are not wildly different, and presumably the Z^{MC} differ somewhat less than the reported values in Table 1. For larger n, the contours become straighter and more similar, and more importantly, the probability becomes more peaked, so that a smaller region contributes. Thus, the central limit forces convergence to a unique Z value for large n.

[1] There may have been typographical errors in the results for Z_{Bi}, identical to Z_9, but described as having different deviations from the true MC result. If the Z's were, by coincidence, identical, this might be an instance of the measure-dependence described below. Alas, the paper was published without the MC comparisons figures.

Reference	[15]	[16]	[17]	[18]	[19]	[19]	[20]	[21]	[22]	[23]	[22]	RMS
Non = x	4	6	9	17	50	67	200	523	167589	498426	2119449	
Noff = y	5	18.78	17.83	40.11	55	15	10	2327	1864910	493434	23671193	
α	0.2	0.0692	0.2132	0.0947	0.5	2.0	10.0	0.167	0.0891	1.000	0.0891	
$b = \alpha y$	1.0	1.3	3.8	3.8	27.5	30.0	100.0	388.6	166213	493434	2109732	
s = Non - b	3.0	4.7	5.2	13.2	22.5	37	100	134.4	1376	4992	9717	
δb	0.45	0.3	0.9	0.6	3.71	7.75	31.6	8.1	121.7	702.4	433.6	
δb/b	0.447	0.231	0.237	0.158	0.135	0.258	0.316	0.0207	0.000732	0.00142	0.000206	
Reported p		.0030	.027	2.0E-06								
Reported Z		2.7	1.9	4.6	3.0	3.0		5.9	3.2	5.0	6.4	
Recommended:												
Z_{Bi} Binomial	**1.66**	**2.63**	**1.82**	**4.46**	**2.93**	**2.89**	**2.20**	**5.93**	**3.23**	**5.01**	**6.40**	0
Z_Γ Bayes Gamma	**1.66**	**2.63**	**1.82**	**4.46**	**2.93**	**2.89**	**2.20**	**5.93**	*	*	*	0
Reasonable:												
Z_N Bayes Gauss (HEP)	1.88	2.71	1.94	4.55	3.08	*3.44*	*2.90*	**5.93**	**3.23**	**5.02**	**6.40**	.28
Z_0 $\sqrt{}$ + 3/8	1.93	2.66	1.98	4.22	3.00	3.07	2.39	5.86	**3.23**	**5.01**	**6.40**	.15
Z_L L Ratio (GRA)	1.95	2.81	1.99	4.57	3.02	3.04	2.38	**5.93**	**3.23**	**5.01**	**6.41**	.14
Not Recommended:												
$Z_9 = s/\sqrt{\alpha(N_{on} + N_{off})}$	*2.24*	*3.59*	2.17	*5.67*	3.11	**2.89**	**2.18**	6.16	**3.23**	**5.01**	**6.41**	.52
$Z_5 = s/\sqrt{N_{on} + \alpha^2 N_{off}}$	1.46	*1.90*	1.66	*3.17*	2.82	3.28	*2.89*	*5.54*	**3.22**	**5.01**	**6.40**	.93
$Z_{5'} = s/\sqrt{\alpha(1+\alpha)N_{off}}$	*2.74*	*3.99*	*2.42*	*6.47*	*3.50*	*3.90*	*3.02*	6.31	**3.23**	**5.03**	**6.41**	.53
Ignore δb:												
Z_P Poisson: ignore δb	2.08	2.84	2.14	4.87	*3.80*	*5.76*	*7.72*	*6.44*	3.37	*7.09*	6.69	1.9
Z_3 Fraser-Reid $\approx Z_P$	2.07	2.84	2.14	4.87	*3.80*	*5.76*	*(8.95)*	*6.44*	3.37	6.09	6.69	2.2
$Z_{sb} = s/\sqrt{b}$	*3.00*	*4.12*	*2.67*	*6.77*	*4.29*	*6.76*	*10.00*	*6.82*	3.38	*7.11*	6.69	2.9
Unsuccessful Hacks:												
Poisson: Nb → b + δb	1.56	2.46	1.64	**4.47**	3.04	*4.24*	*5.51*	6.01	3.07	*6.09*	**6.39**	1.1
s / $\sqrt{b + \delta b}$	*2.49*	*3.72*	*2.40*	*6.29*	*4.03*	*6.02*	*8.72*	*6.75*	3.37	*7.10*	6.69	2.4

Table I : Test Cases and Significance Results: Inputs are at top; α deduced from Eq.1 for HEP examples. The test cases are ordered in data counts; [19]; [20], and [23] have large values of α, troublesome for some methods. Z-values in **bold** are nearly equal the Binomial values; Z-values in *italics* differ by more than .5 . * indicates convergence failure. The last column gives the un-weighted RMS difference of the Z-values from to the Binomial values.

Although Monte Carlo studies can never explore the entire parameter space, the general conclusion of [3] is that Z_{Bi} is the best of the alternatives. Z_{Bi} is only slightly conservative for $Z > 3$. There, p_{Bi} is a bit larger than p^{MC} and thus $Z_{Bi} < Z^{MC}$ by 3% or less on the Z scale when $\min(N_{on}, N_{off}) < 20$, and Z_{Bi} performs even better for larger n. They found the deviations of other methods from Z^{MC} are typically larger. They also cite work[8] which finds larger fractional deviations[2] for Z_{Bi} for smaller Z. Since $Z > 3$ is the lower edge of the region where claims are liable to be made, and the degree of conservatism is small, this would also justify accepting Z_{Bi} as the reference standard, and as the recommended method of evaluating significance when there is any concern about the validity of other methods–at least when a single counting uncertainty dominates the knowledge of the background.

Acknowledgments

The author wish to thank LANL for hospitality and financial support during his sabbatical, Milagro and D0 colleagues for information and references, and Tom Loredo for reference [3]. This work was supported in part by NSF contract NSF0140106. The calculations were performed with the assistance (mostly) of *Mathematica.*

[2] It is not clear whether these limitations (originally studied in the purely-binomial setting) are due to discreteness; or whether the conditioning on $N_{on} + N_{off}$ causes the differences from Monte Carlo.

References

[1] Abramowitz & Segun, Handbook of Mathematical Functions, Dover (1968)

[2] Li & Ma, Astroph. J. 272 (1983) 314-324

[3] Zhang & Ramsden, Exp. Astro. 1 (1990) 145-163

[4] Babu & Feigelson, Astrostatistics (1996), Chapman & Hall

[5] Lehman, Testing Statistical Hypotheses, 2nd edition, Wiley (1986)

[6] Stuart & Ord, Kendall's Advanced Theory of Statistics, Vol 1 & 2

[7] James & Roos, Nuc Phys B172 (1980) 475-480

[8] D'Agostino et. al, Am. Statistician (1988) 198

[9] Fraser, "Statistical Inference: Likelihood to Significance", JASA 86 (1990) 258-65

[10] Fraser, Reid, & Wu, ftp://utstat.toronto.edu/pub/reid/research/general/196rev3.ps.Z

[11] Cousins & Highland, NIM A320 (1992) 331

[12] conferences.fnal.gov/cl2k/copies/linnemann1.pdf

[13] Alexandreas et. al., NIM A328 (1993) 570-577

[14] Gelman, Carlin, Stern, & Rubin, Bayesian Data Analysis, Chapman & Hall (1998)

[15] Artificial case suggested by an example in [13]

[16] Abe et. al., PRL 74 (1995) 2626-31: Top quark: (HEP CDF collab.; I chose one of many results)

[17] Abachi et. al., PRL 74 (1995) 2422-6: Top quark: (HEP D0 collab.)

[18] Abachi et. al., PRL 74 (1995) 2632-7: Top quark: (HEP D0 collab.)

[19] Two artificial examples from [3]

[20] An artificial example with large α

[21] icrr.u-tokyo.ac.jp/can/Symp2002/Presentations/S08-Rowell.pdf: Cyg. OB2 (GRA: Hegra collab.)

[22] Atkins, et al. (2003), Astroph. J., 595, 803-811: Crab Pulsar (GRA: Milagro collab)

[23] Reynolds et. al., Astroph. J. 404 (1993) 206-218: Crab Pulsar (GRA: Whipple collab)

How to Claim a Discovery

W.A. Rolke and A.M. López
University of Puerto Rico - Mayaguez

We describe a statistical hypothesis test for the presence of a signal. The test allows the researcher to fix the signal location and/or width a priori, or perform a search to find the signal region that maximizes the signal. The background rate and/or distribution can be known or might be estimated from the data. Cuts can be used to bring out the signal.

1. INTRODUCTION

Setting limits for new particles or decay modes has been an active research area for many years. In high energy physics it received renewed interest with the unified method by Feldman and Cousins [1]. Giunti [2] and Roe and Woodroofe [3] gave variations of the unified method, trying to resolve an apparent anomaly when there are fewer events in the signal region than expected. They all discuss the problem of setting limits for the case of a known background rate. The case of an unknown background rate was discussed in a conference talk by Feldman [4] and a method for handling this case was developed by Rolke and López [5]. Little work has been done though on the question of claiming a discovery. This problem could be handled by finding a confidence interval and claiming a discovery if the lower limit is positive. Instead the question of a discovery should be done separately, by performing a hypothesis test with the null hypothesis H_o:"There is no signal present". Rejecting this hypothesis will then lead to a claim for a new discovery. In carrying out a hypothesis test one needs to decide on the type I error probability α, the probability of falsely rejecting the null hypothesis. This is of course equivalent to the major mistake to be guarded against, namely that of falsely claiming a discovery.

In practice a hypothesis test is often carried out by finding the p-value. This is the probability that an identical experiment will yield a result as extreme (with respect to the null hypothesis) or even more so given that the null hypothesis is true. Then if $p < \alpha$ we reject H_0; otherwise we fail to do so. For the test discussed here it is not possible to compute the p-value analytically, and therefore we will find the p-value via Monte Carlo.

Maybe the most important decision in carrying out a hypothesis test is the choice of α, or what we might call the discovery threshold. As we shall see, this decision is made much easier by the method described here because we will need only one threshold, regardless of how the analysis was done. What a proper discovery threshold should be in high energy physics is a question outside the scope of this paper, although we might suggest $\alpha = 0.001$ (roughly equivalent to 3σ). Sinervo [6] argues for a much stricter standard of 5σ, or $\alpha = 2.9 * 10^{-7}$. We believe that such extreme values were used in the past because it was felt that the calculated p values were biased downward by the analysis process, and a small α was needed in order to compensate for any unwittingly introduced biases. If we were to trust that our p-value is in fact correct, a 1 in 1000 error rate should to be acceptable.

A general introduction to hypothesis testing with applications to high energy physics is given in Sinervo [6]. A classic reference for the theory of hypothesis testing is Lehmann [7].

2. THE SIGNAL TEST

Our test uses $T = x - b$ or $T = x - y/\tau$ as the test statistic, depending on whether the background rate b is assumed to be known or not. Here x is the number of observations in the signal region, y is the number of observations in the background region and τ is the probability that a background event falls into the background region divided by the probability that it falls into the signal region. Therefore y/τ is the estimated background in the signal region and $x - y/\tau$ is an estimate for the signal rate λ. T is the maximum likelihood estimator of λ, and it is the quantity used in Feldman and Cousins [1] without being set to 0 when $x - y/\tau$ is negative. This is not necessary here because a negative value of $x - y/\tau$ will clearly lead to a failure to reject H_0.

Other choices for the test statistic are of course possible. For example, a measure for the size of a signal that is often used in high energy physics is $S/\sqrt{b}$. Under the null hypothesis this statistic is approximately Gaussian, at least if there is sufficient data. Unfortunately the approximation is not sufficiently good in the extreme tails where a new discovery is made, leading to p-values that are much smaller than is warranted. Even when using Monte Carlo to compute the true p-value, this test statistic can be shown to be inferior to the one proposed in our method because it has consistently lower power, that is its probability of detecting a real signal is smaller.

In order to find the p-value of the test we need to know the null distribution. In the simplest case of a known background rate and everything else fixed this is given by the Poisson distribution, but in all other cases it is not possible to compute the null distribution analytically, and we will therefore find it via Monte Carlo. As an illustration consider the following case shown in figure 1: here we have 100 events on the interval $[0, 1]$, with the signal region a priori set to be $[0.44, 0.56]$. There are 25 events in the signal region, and the background distribution is known to be flat. Therefore we find $x = 25$, $y = 75$, $\tau = 7.33$ and $T = 14.77$. Because we know that the background is flat on $[0, 1]$, and because under the null hypothesis all 100 events are background we can simulate this experiment by drawing 100 observations from a uniform distribution on $[0, 1]$ and computing T for this Monte Carlo data set. Repeating this 150000 times we find the histogram of Monte Carlo T values shown in figure 2, case 1. In this simulation 8 of the 150000 simulation runs had a value of T greater than 14.77, or $p = 0.000053$. Using $\alpha = 0.0001$ we would therefore reject the null hypothesis and claim a discovery. Note that in addition to rejecting the null hypothesis we can also turn the p-value into a significance by using the Gaussian distribution and claim that this signal is a 3.87σ effect.

How would things change if the signal region had not been fixed a priori but instead was found by searching through all signal regions centered at 0.5 and we would have accepted any signal with a width between 0.01 and 0.2? That is if we had kept the signal location fixed but find the signal width that maximizes T, the estimate of the number of signal events? Again we can find the null distribution via Monte Carlo, repeating the exact analysis for each simulation run individually. The histogram of T values for this case is shown in figure 2, case 2. Here we find a value of T larger than 14.77 in 570 of the 150000 runs for a p-value of 0.0038 or 2.67σ. At a discovery threshold of $\alpha = 0.001$ we would therefore not find this signal significant anymore.

Even more, what if we also let the signal location vary, say anywhere in $[0.2, 0.8]$? That is for any pair of values (L, H) we define $[L, H]$ as the signal region and $[0, L), (H, 1]$ as the background region, compute $T_{L,H}$ for this pair and then maximize over all possible values of L and H. Note that because $T_{L,H}$ is monotonically increasing in τ as long as all the observations stay either in the signal or in the background region, we can find the maximum fairly quickly by letting L and H be the actual observations. The histogram of $T_{L,H}$ values for this case is shown in figure 2, case 3. We find a value of T larger than 14.77 in 9750 of the 150000 runs for a p-value of 0.065 or 1.51σ, clearly not significant.

It was necessary in the second and third cases above to limit the search region somewhat, to the interval $[0.2, 0.8]$ and to signals at least 0.01 and at most 0.2 wide, because otherwise the largest value of T is almost always found for a very wide signal region, even when a clear narrow signal is present. This restriction will not induce a bias as long as the decision on where to search are made a priori.

In the general situation where the background is not flat on $[0, 1]$ we can make use of the probability integral transform. Of course this requires knowledge of the background distribution F, but if it is not known we can estimate it from the data, either using a parametric function fitted to the data or even using a nonparametric density estimator. Again all calculations are done under the null hypothesis so we do not need to worry about the signal or its distribution.

As long as we copy exactly for the Monte Carlo events what was done for the real data we will find the correct p-value. This includes using cuts used to improve the signal to noise ratio, but it then requires the ability to correctly Monte Carlo all the variables used for cutting, including their correlations.

3. PERFORMANCE OF THE METHOD

As an illustration for the performance of the signal test consider the following experiment: we generate 100 events from a linear background on $[3, 5]$ and (if present) a Gaussian signal at 3.9 with a width of 0.05. Then we find the signal through a variety of situations, from the one extreme where everything is fixed a priori to the other where the largest signal of any width is found. The background density is found by fitting and the background rate is estimated. The power curves are shown in figure 3. No matter what combination of items were fixed a priori or were used to maximize the test statistic, and with it the signal to noise ratio, all cases achieved the desired type I error probability, $\alpha = 0.05$. Not surprisingly the more items are fixed a priori, the better the power of the test.

4. CONCLUSION

We have described a statistical hypothesis test for the presence of a signal. Our test is conceptually simple and very flexible, allowing the researcher a wide variety of choices during the analysis stage. It will yield the correct type I error probability as long as the Monte Carlo used to find the null distribution exactly mirrors the steps taken for the data. Monte Carlo studies have shown that this method has satisfactory power.

Acknowledgments

This work was partially supported by the Division of High Energy Physics of the US Department of Energy.

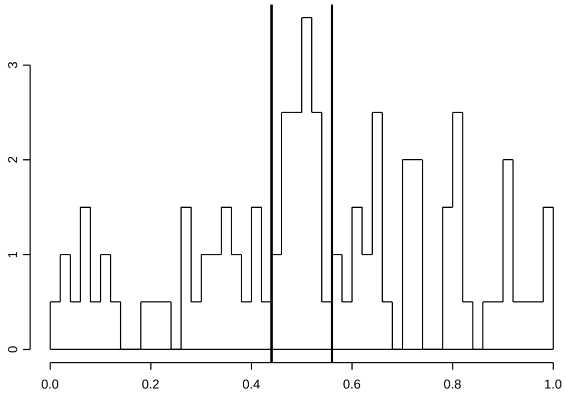
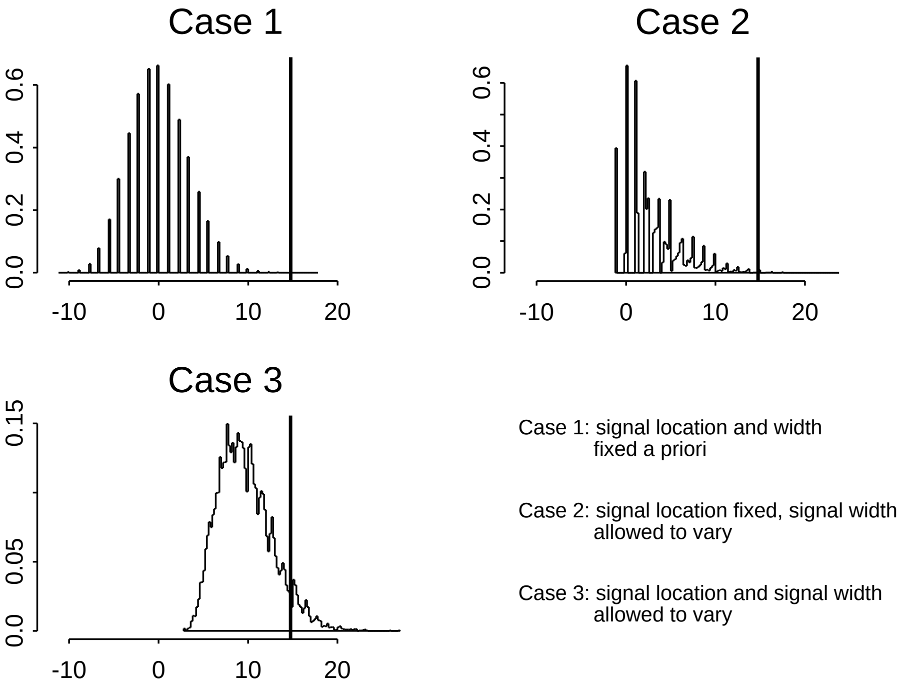


Figure 1: 100 Events on [0,1], with the signal region a priori set to be [0.44, 0.56]. There are 25 events in the signal region, and the background distribution is assumed to flat.

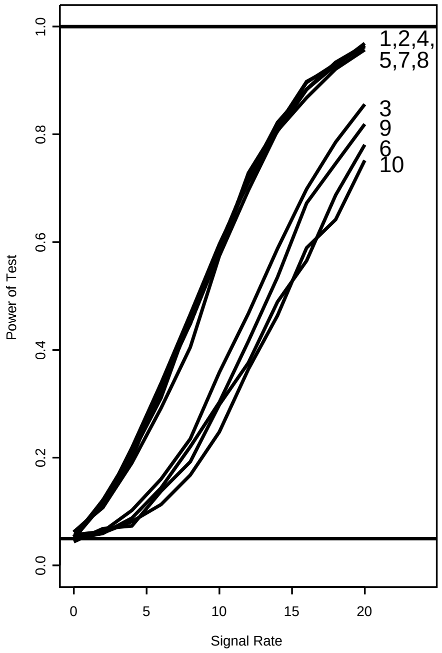


Case 1: signal location and width fixed a priori

Case 2: signal location fixed, signal width allowed to vary

Case 3: signal location and signal width allowed to vary

Figure 2: Histograms of T values of Monte Carlo simulation.



Peak	Width	Param.	Rate
1: fixed	fixed	fixed	estimated
2: fixed	fitted	fixed	estimated
3: fitted	fitted	fixed	estimated
4: fixed	fixed	fixed	exact
5: fixed	fitted	fixed	exact
6: fitted	fitted	fixed	exact
7: fixed	fixed	fitted	estimated
8: fixed	fitted	fitted	estimated
9: fitted	fitted	fitted	estimated
10:fitted	fitted	fitted	exact

Linear background on [3,5]

Gaussian Peak at 3.9, s=0.1

Sample Size :100

Figure 3: Power curves for 10 different cases such as signal location fixed a priori or not, same for signal width, background estimated or ect. alpha=0.05 is used.

References

[1] R.D. Cousins, G.J. Feldman, "A Unified Approach to the Classical Statistical Analysis of Small Signals", Phys. Rev, D57, (1998)

[2] 3873.C. Giunti, "A new ordering principle for the classical statistical analysis of Poisson processes with background" , Phys. Rev D59, 053001 (1999).

[3] B.P. Roe, M.B. Woodroofe, "Improved Probability Method for Estimating Signal in the Presence of Background", Phys. Rev D60 053009 (1999)

[4] G. Feldman, "Multiple measurements and parameters in the unified approach", talk at Fermilab Workshop on Confidence Limits 27-28 March, 2000, http://conferences.fnal.gov/cl2k/, p. 10-14.

[5] W.A. Rolke, A.M. López, "Confidence Intervals and Upper Bounds for Small Signals in the Presence of Background Noise", Nucl. Inst. and Methods A458 (2001) 745-758

[6] P.K. Sinervo, "Signal Significance in Particle Physics", Proceedings of the Conference: Advanced Statistical Techniques in Particle Physics, Institute for Particle Physics Phenomenology, University of Durham, UK, 2002, 64-76.

[7] E.L. Lehmann "Testing Statistical Hypotheses", 2nd Ed. (1986) Wiley, New York.

Scan Statistics in High Energy Physics

F. Terranova
I.N.F.N., Laboratori Nazionali di Frascati, Frascati (Rome), Italy

Scan Statistics are useful tools to signal a departure from the underlying probability model that describes the experimental data. They are commonly used in many research areas such as bioinformatics, control theory and medicine [1]; applications to astrophysics have also been suggested [2] . We consider, here, possible applications to high energy physics (HEP). It is shown that local perturbations ("bumps" of events or unexpected narrow resonances) are better dealt within this framework and, in general, tests based on these statistics provide a powerful and unbiased alternative to the traditional techniques related with the χ^2 and Kolmogorov distributions. We also focus on the extensions needed to meet the challenges of particle physics problems and detail the differences between HEP and non-physics applications.

1. INTRODUCTION

Local perturbations of the expected distribution in a given kinematic variable can signal a departure from the underlying model used to describe the experimental data ("null hypothesis", NH). In high energy physics this phenomenon is related, for instance, to the appearance of an unexpected resonance and the determination of the statistical significance of the excess is a delicate task especially if aimed at claiming a discovery or planning a confirmatory data taking. Global distortions are commonly dealt with by the (unbinned) Kolmogorov-Smirnov test statistics or their extensions. However, the power[1] of these tests is drastically reduced for local perturbations. Conversely, the Pearson χ^2 test performs a binning of the range and compares the content of each bin (k_i) with its expectation under NH (b_i). If the parameters defining NH are known, the Pearson test statistic T

$$T \equiv \sum_{i=1}^{N_{\rm bin}} \frac{(k_i - b_i)^2}{b_i} \qquad (1)$$

behaves as $\chi^2(N_{\rm bin})$ in the asymptotic limit. This test is better suited for local perturbations. Note, however, that the test is unbiased only if the binning grid is fixed *a priori* and the power depends strongly on the peak position. In general, local perturbations shared among different bins are tagged less effectively than peaks where the data cluster around one bin. In this respect, scanning the distribution with a running window of fixed length would be more appropriate and would avoid *a priori* fixing of the binning grid. This technique is sometimes employed in particle physics [3] but no quantitative estimates of the p-value for the null hypothesis are provided due to the strong correlation of the contents of nearby bins. In fact, the problem of the correlations can be solved analytically in the framework of Scan Statistics. Given N events distributed along the $[A, B]$ range, we call $S(w)$ ("scan statistic") the largest number of events in a window of fixed length w. If the distribution of $S(w)$ is known, it is possible to compute the probability $P(S(w) \geq k)$ for the null hypothesis to produce a cluster $S(w)$ greater or equal than the one actually observed. Hence, the p-value of the null hypothesis can be assessed. In this context, an *a priori* binning similar to the one of the Pearson χ^2 test is no more needed. Moreover, the test statistic $S(w)$ is not affected by the drawbacks of the Kolmogorov-Smirnov (KS) tests.

2. THE PROPERTIES OF THE SCAN STATISTIC

For very simple cases, the computation of $P(S(w) \geq k)$ can be carried out through direct integration of the probability density function (p.d.f.). This is the case, for instance, of $P(S(w) \geq 2)$ when the events are $X_1 \ldots X_N$ independent and identically distributed random variables with common density $f(x) = 1$ for $x \in (0, 1)$ and zero elsewhere [1]. Here, the problem can be solved noting that $P(S(w) \geq 2) = P(W_2 \leq w)$, W_2 being the size of the smallest interval that contains 2 events. In fact

$$W_k = \min_{1 \leq i \leq N-1} \{X_{(i+1)} - X_{(i)}\} \qquad (2)$$

where $X_{(i)}$ denote the *ordered* value of the X's and

$$P(W_2 > w) = P\left\{ \bigcap_{i=1}^{N-1} \left[X_{(i+1)} - X_{(i)} > w\right] \right\} \qquad (3)$$

[1] The power of an hypothesis test against a specific alternative hypothesis is the chance that the test correctly rejects the null hypothesis when that alternative hypothesis is true; that is, the power is 100% minus the chance of a Type II error when that alternative hypothesis is true. On the other hand, the significance level of a hypothesis test is a fixed probability of wrongly rejecting the null hypothesis, if it is in fact true.

Thus, we are interested on the simultaneous occurrence of the conditions $X_{(i+1)} - X_{(i)} > w$ for all $i = 1, \ldots N-1$. $P(W_2 > w)$ results from the integration of the p.d.f. of the ordered distribution once the boundary of the integration is chosen in a way to fulfill $X_{(i+1)} - X_{(i)} > w$. If $w < 1/(N-1)$ we end up with the following integral:

$$P(W_2 > w) = N! \int_{(N-1)w}^{1} dx_N \int_{(N-2)w}^{x_N - w} dx_{N-1} \ldots \int_{w}^{x_3 - w} dx_2 \int_{0}^{x_2 - w} dx_1 = (1-(N-1)w)^N \qquad (4)$$

otherwise $P(W_2 > w) = 0$. Note that the joint p.d.f. of $X_{(1)} \ldots X_{(N)}$ is $N!$ times the original joint p.d.f. for $X_1 \ldots X_N$. This follows from the fact that the ordering function maps $X_1 \ldots X_N$ into $X_{(1)} \ldots X_{(N)}$, i.e. it represents a $N!$ to 1 transformation. The p.d.f. of the transformed variable is the sum of the contributions from each of these component mappings, hence it is $N!$ denser. Unfortunately, for values of k greater than 2 or 3 and smaller than N or $N-1$ this direct integration approach becomes overly complicated and a combinatorial approach based on the Karlin-McGregor theorem turns out to be more appropriate.

In fact, in the vast majority of HEP applications, events are produced randomly through a Poisson process and are characterized by a kinematic variable that is randomly distributed over an interval. Let us consider an interval $[A, B]$ of a continuous variable x and a Poisson process ("background") whose mean value per unit interval is denoted with λ. Hence, the probability of finding $Y_x(w)$ events in an interval $[x, x+w]$ is $P(Y_x(w) = k) = e^{-\lambda w}(\lambda w)^k/k!$ The number of events in any disjoint non-overlapping intervals are independently distributed. Again, the scan statistic is the largest number of events to be found in any subinterval of $[A, B]$ of length w. The formulas for fixed N can be extended to the Poisson case. Moreover in most of particle physics analysis a simple approximation by Naus [4] can be implemented. In this case $P(S(w) < k) \simeq Q_2 \left[Q_3/Q_2\right]^{\frac{\Delta}{w}-2}$ where $\Delta \equiv B - A$; Q_2 and Q_3 are functions of the Poisson probability $P(k, \lambda w)$ and their cumulative. Full analytic formulas for $P(S(w) < k)$, Q_2 and Q_3 can be found in [1, 5].

3. SEARCH FOR NARROW RESONANCES

From the discussion of Sec. 1, it is clear that the ideal situation to employ a goodness-of-fit test based on Scan Statistics (SS) is the search for narrow resonances along a 1-dim distribution of kinematic variables [2]. In this case the scanning width is fixed *a priori* and it corresponds to the expected width due to the finite detector resolution. To determine the power of the SS-based test we considered as alternative hypotheses local perturbations of the uniform distribution which leads to the appearance of a "excess" of events. The alternative functions are Poisson processes of mean S. The signal events are spread along $[A, B]$ according to a normal distribution of mean x_S and sigma σ_S. The significance of the test is at 95%. Fig. 1 shows the power of the KS, SS and χ^2 tests as a function of the signal position x_S. Here, $[A, B] = [0, 1]$, $\sigma_S = 0.05$, $w = 4\sigma_S$, $\lambda\Delta = 100$ and $S = 20$. The optimal bin size for the χ^2 test has been computed following the prescription [6] $N_{\text{bin}} = 2(\lambda\Delta)^{2/5}$, where $\lambda\Delta$ is the expected sample size in case of null hypothesis. Other choices of the binning for the χ^2 test, based on the knowledge of σ_S, have been tested by Monte Carlo experimentation. The corresponding powers do not exceed the one shown in Fig. 1. Signal events generated beyond the interval $[0, 1]$ are ignored (out of the sensitivity region $[A, B]$).

As anticipated in Sec. 1 the KS test is not appropriate for local perturbations. The power is limited compared to other statistics and depends on the peak position, having the highest sensitivity at the border of the distribution. The Pearson χ^2 test has a much higher power but in general the peak detection efficiency is reduced when the peak is shared between two adjacent bins. On average the χ^2 test underperforms with respect to SS since the correlations among the bins are ignored. However, the bin prescription for χ^2 is independent of the *a priori* knowledge of σ_S while SS makes use of this additional information. This is a drawback for SS if the actual width σ_S is not the same as the expected instrumental resolution sigma, because the scanning window is no more optimized. However, it can be shown [5] that within a large range of mismatch (up to a factor three) between

[2] E.g. the case of [3] where the resonance was sought for through the distribution of the dijet invariant mass sum.

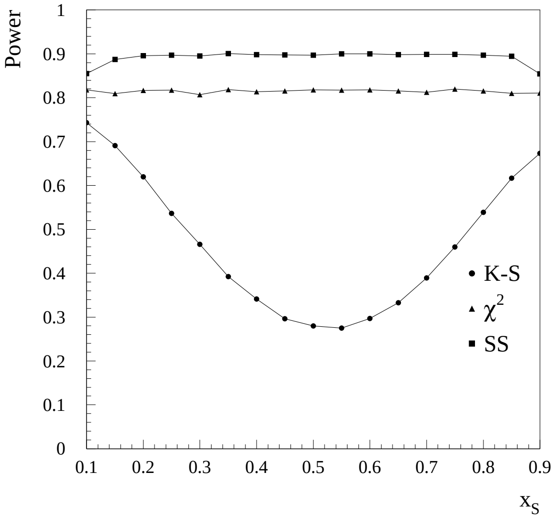


Figure 1: The power of the test statistic versus the peak positions for $S = 20$, $B = 100$, $\sigma_S = 0.05$ and $w = 4\sigma_S$.

the resonance width and the scanning window w the SS-based test still has higher power than Pearson χ^2. Moreover, the SS p-value tables turn out to be correct [5] even when very few events are observed. This is due to the fact that $P(S(w) < k)$ is derived invoking neither the Central Limit Theorem nor any normality approximation. Conversely, the corresponding tables for χ^2 hold in the asymptotic limit only and need to be recomputed by Monte Carlo experimentation if the number of events per bin is not sufficiently high. A more detailed comparison among the various tests can be found in [5].

4. NON UNIFORM BACKGROUND AND PARAMETER ESTIMATION

The assumption of uniform background in $[A, B]$ is rarely fulfilled in HEP. In most of the cases the expected events per unit interval is a function of the position along $[A, B]$: $\lambda = \lambda(x)$. The most straightforward extension of SS is due to Weinstock [7] and allows for stretching the window during the scan: the scan statistic is computed with a x-dependent window of variable width $w'(x)$ that always contains $w/(B - A)$ percent of the expected events under the null hypothesis. The advantage of this approach is that we expect the p-value tables for NH to be identical to the ones of the standard SS with width w and uniform background. On the other hand, the optimal pulse alternative for this test is no more the optimal pulse of $S(w)$ but the corresponding pulse after stretching of its domain. In general, this causes a non negligible loss of power only for strongly varying background (e.g. exponential decays). In those cases, more powerful generalizations of $S(w)$ exist [8] but the p-value tables have to be computed numerically.

The null hypothesis is often specified up to a set of free parameters which have to be extracted from the data. The concept of degree-of-freedom allows to correct the p-value of the Pearson χ^2 for NH keeping into account the fact that NH is defined using part of the experimental information. If b_i of Eq. 1 are functions of a set of M unknown parameters $\underline{\theta}$, the test statistic $T(k_i, b_i(\hat{\underline{\theta}}))$ behaves asymptotically as $\chi^2(N_{bin} - M)$, or $\chi^2(N_{bin} - M - 1)$ if the normalization is fixed. $\hat{\underline{\theta}}$ are the estimated parameters from the data through χ^2 minimization. Unfortunately, this concept cannot be implemented in a straightforward manner to SS. Here, SS should be extended to devise the optimal estimate of the underlying background density that is unbiased and consistent under both the null hypothesis and the occurrence of a local excess of width σ_S. This problem is still unsolved for a generic function. Unbiased estimators have been obtained for simple functional dependencies as in the case of the linear regression [1]. In general, this implies an optimal splitting of the range $[A, B]$ between a “minimization region” used to draw $\hat{\underline{\theta}}$ and a complementary “scanning region” where the resonance is searched for. This technique allows to retain high power at least in the scanning region. Clearly, it is always possible to estimate $\hat{\underline{\theta}}$ using the whole range, determine naively $S(w', \hat{\underline{\theta}})$ instead of $S(w', \underline{\theta}_{true})$ and re-compute numerically the p-value tables (e.g. by MC). However, in general this approach results in a significant power loss.

5. CONCLUSIONS

Tests based on Scan Statistics are currently applied in several research areas to investigate local anomalies in time series of events. High energy physics offer many case studies where SS tests are expected to be powerful and easy to implement. The search for narrow resonances and the determination of the statistical relevance of local distortions in the distribution of the experimental data are two of them.

The Scan Statistics can be adapted to cope with common situations in particle physics such as the occurrence of non-uniform background (see Sec. 4) and scan over periodical angular variables [5]. On the other hand, their use become less straightforward when the scanning width is unknown *a priori* or the parameters of the null hypothesis must be estimated from the data. These cases were discussed in Sec. 3,4 and even in these conditions SS provide significant improvements with respect to traditional goodness-of-fit tests.

References

[1] J. Glaz, J. Naus and S. Wallenstein, “Scan Statistics”, Springer, New York, 2001.

[2] K.J. Orford, J. Phys. G26 (2000) R1.

[3] K. Ackerstaff et al. [OPAL Collaboration], Phys. Lett. B429 (1998) 399.

[4] J.I. Naus, J. Amer. Stat. Ass. 77 (1982) 177.

[5] F. Terranova, “Peak finding through Scan Statistics”, to appear in Nucl. Instr. Meth., arXiv:physics/0311020.

[6] D. Moore, in Goodness of fit techniques, R.B. D’Agostino and M.A. Stephens eds., Dekker, New York, 1986.

[7] M. Weinstock, Int. J. Epidem. 10 (1981) 289.

[8] M. Kulldorff et al., Stat. Med. 14 (1995) 799.

Supernova 1987A Neutrino Signal: Statistical Power of a Small Sample

Alfred Scharff Goldhaber
C.N.Yang Institute for Theoretical Physics, SUNY Stony Brook, NY 11794-3840 USA

Of the 20 neutrino events associated with SN1987A, all eight from IMB and four from Kamiokande II have measured energies lying at or above 20 MeV. Unlike the eight lower-energy events, this sample shows strong forward peaking, so that the *a priori* probability of obtaining a distribution as forward as this from an isotropic mechanism is less than one part in a million. Standard theory expects a nearly isotropic distribution. Previous analysis of these events are reviewed. More sophisticated statistical approaches (which should give a higher probability than quoted above, but lower than in previous analysis) are invited.

1. INTRODUCTION

At the beginning of 1987 there appeared by far the nearest supernova in living memory. Without very much conscious planning, it happened that there were not one but two underground Čerenkov water detectors in operation, able to detect neutrinos emitted by the supernova [1–4]. The great fact about these observations is that, within an order of magnitude, they confirmed the evolving picture of supernova development, which implies that most of the energy gained from gravity during the collapse phase should be radiated in the form of neutrinos.

However, from the very beginning it was clear that at the next stage of detail these events present a puzzle: For events with more than 20 MeV visible electron or positron energy, an expected nearly isotropic distribution is observed to be rather forward peaked (see Figure 1). Except for small impurities [5], there are only three types of target in the water, electrons, protons, and ^{16}O nuclei. For energetic (20 MeV or more) incident neutrinos, center of mass motion guarantees an extremely forward-peaked ($\cos\theta > 0.98$), angular distribution on the electron target for final electrons or positrons, and the cross section is quite small because of the small target mass. For proton targets, the interaction due to the charged weak current is very well understood, being close to isotropic for incident electron antineutrinos, and vanishing for all other types in the energy range of interest.

The oxygen nuclei have a substantial threshold for charged current interactions by electron neutrinos or antineutrinos, and standard nuclear wave functions imply if anything a backward bias in the angular distribution. Thus the perfectly known cross sections are either too small and much too forward peaked (electrons), or too nearly isotropic to fit well with the data. Even the oxygen nuclei do not seem to give much room for understanding the results. One concludes that either there is a substantial statistical fluctuation or there is some physics that has not yet been identified.

The aim of the present work is to present a case, using a simple probabilistic analysis, that the significance of this discrepancy is large enough to justify further attention, even if it is beyond our control to repeat the experiment on demand by ordering up another nearby supernova!

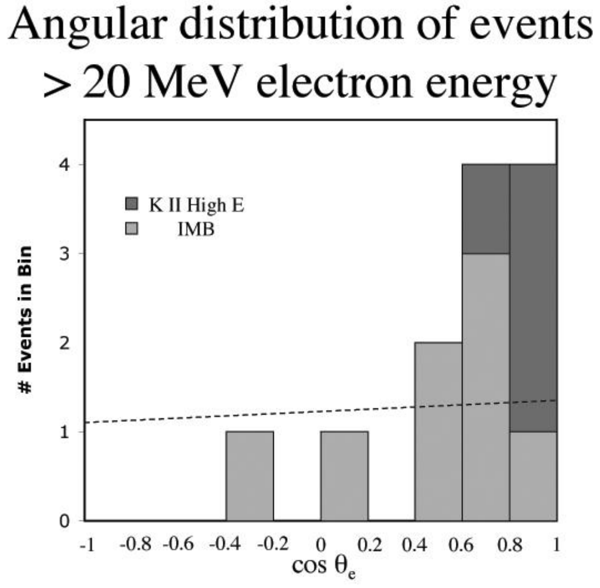


Figure 1: This histogram shows the distribution in direction cosine with respect to a line from SN 1987A to the detector for the eight events of IMB [3], all at or above 20 MeV in energy, and the four events in that energy range out of 12 events altogether of Kamiokande II [K II] [4]. The dashed line illustrates the expected angular distribution for the final positrons in the reaction $\bar{\nu}_e p \to e^+ n$. Evidently there is a substantial forward fluctuation in the samples from both detectors.

2. PROBABILITY CALCULATION

Let us start by focusing on the four events at or above 20 MeV in the data from K II [1]. The 8 events at or below 13 MeV arguably may be excluded on the basis of two considerations having nothing to do with angular distribution. First, 20 MeV was the threshold for detection at IMB, and thus this sample is directly comparable to the entire sample of 8 events from IMB. Secondly, there is a notable gap of about 7 MeV between the lower- and higher-energy events at K II, giving at least the possibility that different mechanisms are responsible for the signals in the two energy ranges. If one now does look at the angular distribution of the lower-energy events, it is quite consistent with isotropy, appropriate to absorption of electron antineutrinos on protons [4].

Of the four higher-energy events at K II, three are in the most forward decile in $\cos\theta$, and the remaining one is in the next most forward decile. If we ask the *a priori* probability that this distribution or a more forward one would have resulted by chance from an isotropic distribution, it is easy to compute the answer. The most forward distribution would be all four events in the top decile, with probability 10^{-4}. The next most forward is the observed signal, with exactly one of the four events in the second decile. This from simple combinatorics has probability 4×10^{-4}, yielding finally a total probability 5×10^{-4}.

The general method of computation exemplified by the above exploits the probability of any particular distribution of events among bins. If there are N events altogether, this probability is

$$P(\{n_i\}) = N! \prod p_i^{n_i}/n_i! \quad , \qquad (1)$$

where p_i is the expected probability for an event to appear in bin i. The total *a priori* probability P_T for a distribution at least as forward as a given distribution is obtained by summing the above probabilities over all distributions in which any number of events are moved forward by any allowed number of bins from their places in the given distribution.

The result for the 8 events seen at IMB, again assuming an isotropic mechanism, is approximately $P_{IMB} = 1.75 \times 10^{-3}$, so that the product of the two *a priori* probabilities is smaller than one part in a million. This seems more than enough to justify serious attention to the possibility of a genuine physical mechanism for the forward distribution, but we still need to look at this calculation more closely.

3. CRITIQUE AND ALTERNATE APPROACHES

There are a number of considerations needed to obtain a full perspective on the likelihood that the observed distribution could have arisen by chance.

3.1. Bin Size

An attractive feature of the probability defined above is that it decreases monotonically as bins are subdivided, going to a nonzero asymptotic value in the limit of infinite subdivision, with little change once the bin size becomes small compared to event spacings. Thus the binning adopted here should be quite close to optimum for efficiency and accuracy together, especially in view of the angular uncertainties of the observations.

3.2. Anisotropy in Prediction

Although the angular distribution for $\bar{\nu}_e p \to e^+ n$ in the relevant energy range is roughly isotropic, there is a mild forward distortion linear in $\cos\theta$ [6], with amplitude about 0.1 in the energy range 20-40 MeV. Evidently this increases somewhat the probability of forward-peaked distributions for higher-energy events as observed in the two experiments. Further, for IMB, the efficiency in the most backward bin is reduced relative to the average by about 10 %, a small additional contribution to expectation of a forward tendency in the data [3].

3.3. General Fluctuation versus Forward Fluctuation

The conventional method of estimating the probability of a large fluctuation from isotropy is to allow that fluctuation to be in any direction, not just forward along the beam direction. Clearly this would give a substantially larger probability for the fluctuations of each of the two distributions. A Monte Carlo estimate of this probability using the Smirnov-Cramer-von Mises statistic was given for the IMB events [3], and not surprisingly was almost one order of magnitude bigger than the estimate above. Clearly a similar statement would apply for K II. However, multiplying these two numbers together surely is too conservative, because the fluctuations in both samples actually are forward with respect to the same (beam) direction.

3.4. Towards a More Appropriate Measure of Probability

To make a precise statement about the relevant probability in view of the above considerations could require a more sophisticated analysis, but the following might be a reasonable first approximation: Use the IMB calculation for the probability of such a large fluctuation in their data. Then keep the ‘as forward as’ calculation for K II, because that happens to be the direction of the actual IMB fluctuation. This would imply a probability of the combined fluctuations in the two samples in the range of a part in 100 thousand. That still is enough according to usual thinking to impel careful searching for some physics underlying the observation. A more conventional alternative worth pursuing would be to repeat the Monte Carlo estimation approach for the combined sample, which also might give a smaller probability than the product of separate probabilities for the two samples.

3.5. Keeping the Low-Energy Events

Another estimation was given by Vogel and Beacom [VB] [7], who lumped together the high- and low-energy data from K II (consistent with the hypothesis that all events are on a proton target). Evidently this assumption dilutes the signal as portrayed in Figure 1. To estimate the probability of such a fluctuation, they considered only values of the first and second moments of the $\cos\theta$ distribution. A glance at the figure makes it obvious that with a fairly uniform distribution added in for the eight low-energy events one needs more than these two moments to characterize the fluctuation. Thus both the high-energy and the inclusive data selections clearly call for more sophisticated statistical treatment.

3.6. Electron Target Hypothesis

As already mentioned, the forward events (actually with one possible exception) are not forward enough to be associated easily with electron targets, given the claimed angular resolution [3, 4]. Kiełczewska [8] investigated the flux of μ and τ neutrinos needed to account for the forward events (ignoring the implications of the stated angular resolution). [The restriction to these flavors was to avoid feeding the reaction on proton targets.] She found that rather extreme assumptions, including a much harder spectrum for these neutrinos than for e neutrinos, were needed to make the total radiated neutrino energy less than the energy available from gravitational collapse. Taking into account mixing of neutrino species, something not established at that time, would require revision of that analysis, using considerations found, e.g., in [9].

4. ANOTHER PUZZLE

So far we have been exploring an anomaly in observation of a single astrophysical event. However, there is a generic anomaly in the theory of supernova formation. Even the most advanced computer simulations, including as best is feasible all recognized physics, fail to predict supernova explosions [10]. A possible reason for this failure may lie in what has not yet been possible to include in the calculations, in particular full three-dimensionality for the spatial evolution, which up to now has been mimicked using only two space dimensions. However, it is at least conceivable that some physics of neutrino interactions with matter has yet to be included. [It is generally accepted that momentum transfer from outgoing neutrinos to infalling matter is crucial to detonation.] If so, then one has the appealing possibility that a single mechanism could explain the angular anomaly in the SN1987A data, and also supply the additional 'punch' necessary to detonate supernovae. As explained already, neither the electron nor the proton targets are promising; neutrino interactions with these targets are well determined by existing knowledge. While nuclear targets are sufficiently complex to make some new mechanism conceivable, there is as yet no explicit demonstration of such a mechanism either theoretically or by experiments other than the SN1987A observations. Given extensive analysis using reliable nuclear wave functions, and including, e.g., virtual pion coupling simultaneously to two nucleons, a new mechanism most likely would involve a collective interaction with the nucleus as a whole.

Acknowledgments

I thank Maurice Goldhaber for drawing my attention to the SN1987A puzzle, Anna Hayes for extensive calculations and discussions about current knowledge on neutrino-nucleus interactions, and William Murray for emphasizing that conventional approaches in terms of probability of large fluctuations must give a larger probability than my 'as forward as observed' approach. Much of this work evolved during extended visits to Los Alamos National Laboratory, Stanford Linear Accelerator Center, and Massachusetts Institute of Technology. The work was supported in part under NSF grant PHY-0140192.

References

[1] K. Hirata *et al.*, Phys. Rev. Lett. **58**, 1490 (1987).

[2] R.M. Bionta, *et al.*, Phys. Rev. Lett. **58**, 1494 (1987).

[3] C.B. Bratton *et al.*, Phys. Rev. D **37**, 3361 (1988).

[4] K.S. Hirata *et al.*, Phys. Rev. D **38**, 448 (1988).

[5] W.C. Haxton, Phys. Rev. D **36**, 2283 (1987), as part of a thorough discussion of various neutrino reactions on nuclei, observes that the very small ^{18}O contamination can give substantial reaction rate, but still too small to influence the SN1987A observations.

[6] C.H. Llewellyn Smith, Phys. Rept. **3**, 261 (1972).

[7] P. Vogel and J.F. Beacom, Phys. Rev. D **60**, 053003 (1999).

[8] D. Kiełczewska, Phys. Rev. D **41**, 2967 (1990).

[9] E.K. Akhmedov, C. Lunardini, and A.Yu. Smirnov, Nucl. Phys. **B643**, 339 (2002).

[10] R. Buras, M. Rampp, H.-Th. Janka, and K. Kifonidis, Phys. Rev. Lett. **90**, 241101 (2003).

Pitfalls of Goodness-of-Fit from Likelihood

Joel Heinrich
University of Pennsylvania, Philadelphia, PA 19104, USA

The value of the likelihood is occasionally used by high energy physicists as a statistic to measure goodness-of-fit in unbinned maximum likelihood fits. Simple examples are presented that illustrate why this (seemingly intuitive) method fails in practice to achieve the desired goal.

1. INTRODUCTION

> For every complex problem, there is a solution that is simple, neat, and wrong. *H.L. Mencken*

The complex problem considered here is goodness-of-fit (g.o.f.) for unbinned maximum likelihood fits in cases when binned g.o.f. methods and Kolmogorov-Smirnov are not well suited:

A physicist, having fit a complicated model to his multi dimensional data to obtain estimates of the values of certain parameters, is also expected to check how well the data match his model. In the sections that follow, we discuss a g.o.f. method, still occasionally used in high energy physics (HEP), that is simple, neat, and wrong.

2. THE SNW[1] METHOD

We start with a brief description of the method. (A true derivation, for obvious reasons, is not available.)

observation: Maximum likelihood fits are performed by maximizing the likelihood $L(\vec{\theta}, \vec{x})$ with respect to the (unknown) parameters $\vec{\theta}$ for fixed data $\vec{x}$.

faulty intuition: Thus, the value of the likelihood provides the g.o.f. between the data and the probability density function (p.d.f.): The value of the likelihood at the maximum,

$$L_{\max} = L(\hat{\vec{\theta}}, \vec{x})$$

corresponds to the best fit—the smaller the likelihood, the worse the g.o.f., ...

obstacle: To calculate this "g.o.f." P-value, we need the distribution of $L_{\max}$ for an ensemble of random $\vec{x}$ deviates from the p.d.f. using the true (but unknown) parameters $\vec{\Theta}$.

faulty resolution: We approximate this by replacing $\vec{\Theta}$ with the parameter estimate obtained from the fit to the actual data.

This method has a long history of use in high energy physics. It's recommended by several excellent statistical data analysis texts written by (and for) high energy particle physicists. Consequently, and because the method is "obvious", it's still being used in (some) HEP analyses.

Reference [1], written by a statistician and four physicists, describes the method, but criticizes:

> The likelihood of the data would appear to be a good [g.o.f.] candidate at first sight. Unfortunately, this carries little information as a test statistic, as we shall see...

Since this was ignored, maybe its warning was not strong enough. I have found no mention of the method in texts written (solely) by statisticians.

3. A SIMPLE TEST OF THE METHOD

> Always test your general reasoning against simple models. *John S. Bell*

Reference [2], following the above advice, tests the method against the p.d.f.

$$\frac{1}{\tau}e^{-t/\tau} \qquad (t \geq 0)$$

where t (we have in mind the decay-time of a particle) follows an exponential distribution, and τ (the mean lifetime) is a parameter whose value, being unknown, is estimated from data. The likelihood for N observations t_i is given by

$$-\ln L = \sum_{i=1}^{N} \left[\ln \tau + \frac{t_i}{\tau} \right]$$

The value ($\hat{\tau}$) of τ that maximizes the likelihood, and the value ($L_{\max}$) of the likelihood at its maximum, are given by

$$\hat{\tau} = \frac{1}{N} \sum_{i=1}^{N} t_i \qquad -\ln L_{\max} = N(1 + \ln \hat{\tau})$$

[1] Simple, Neat, Wrong.

3.1. The First Surprise

The value of the likelihood at its maximum (in this test case) is just a simple function of $\hat{\tau}$—all samples with the same mean obtain the same "g.o.f." value. This is a disaster for g.o.f. Even if the true value of τ—call it $\mathcal{T}$—were known in advance, so that we could calculate the P-value associated with the observed $\hat{\tau}$, merely comparing the $\hat{\tau}$ of the data with $\mathcal{T}$ is not sufficient to show that the observed data are modeled well by the exponential distribution.

3.2. The Second Surprise

Since under this method, our P-value ensemble is actually based on the value of $\hat{\tau}$ computed from the data (not knowing the true value $\mathcal{T}$), we *always* obtain a P-value of about 50%, *for any data whatsoever.* This is a second disaster for g.o.f. By construction, the distribution of $L_{\max}$ from our ensemble of N-event pseudo experiments tracks the $L_{\max}$ observed from the data.

The fact that the method yields "reasonable" P-values has undoubtedly contributed to its longevity in practice: P-values very near 0 or 100% would have triggered further investigation.

3.3. Lessons Learned

In this example, g.o.f. is equivalent to testing the single hypothesis: "The data are from an exponential distribution of unspecified mean." $L_{\max}$ provided no information with respect to this hypothesis.

What went wrong? In our test case, the likelihood could be expressed as a function of just the parameter and its maximum likelihood estimator (m.l.e.): $L(\tau; \hat{\tau})$. *All* data samples with the same m.l.e. gave the same "g.o.f."

Exactly the same thing happens in the Gaussian (normal) case—the likelihood can be written using solely the 2 parameters and their estimators: $L(\mu, \sigma; \hat{\mu}, \hat{\sigma})$.

Other "textbook" distributions—scaled gamma, beta, log-normal, geometric—also fail in the same way. Geometric is a discrete distribution, so the problem is not restricted to the continuous case.

4. MORE TROUBLE: NON INVARIANCE

Returning to our exponential example, suppose we make the substitution $t = x^2$. The p.d.f. transforms as

$$\frac{1}{\tau} e^{-t/\tau} dt = \frac{2x}{\tau} e^{-x^2/\tau} dx$$

and the g.o.f. statistic is now calculated as

$$-\ln L = \sum_{i=1}^{N} \left[\ln \tau + \frac{x_i^2}{\tau} - \ln(2x_i) \right] \qquad \hat{\tau} = \frac{1}{N} \sum_{i=1}^{N} x_i^2$$

$$\begin{aligned} -\ln L_{\max} &= N(1 + \ln \hat{\tau}) - \sum_{i=1}^{N} \ln(2x_i) \\ &= N(1 + \ln \hat{\tau}) - \sum_{i=1}^{N} \ln(2\sqrt{t_i}) \end{aligned}$$

That is, the "g.o.f." statistic is not invariant under change of variable in the continuous p.d.f. case. (The value of the m.l.e. is, of course, invariant.)

Under change of variable, the "g.o.f." statistic picks up an extra term from the Jacobian—an extra function of the data. We're free to choose any transformation, so we can make the "g.o.f." statistic more or less anything at all—a serious pathology.

At this point, experts point out that *ratios* of likelihoods have the desired invariance under change of variable, but, while the likelihood ratio is a useful test statistic in certain special cases, it is not at all clear how to obtain a useful g.o.f. statistic from the likelihood ratio in the general, unbinned, case.

5. A REPLACEMENT MODEL

Since we now lack an intuitive understanding, we need a replacement intuition for what is going on. I propose this model:

Denote by H_0 the hypothesis that the data are from the p.d.f. in question. Specify an alternative hypothesis H_1 that the data are from a uniform p.d.f. (flat in the variables that we happen to have chosen). At least, the H_1 p.d.f. is flat over the region where we have data—outside that region it can be cut off.

Performing a classic Neyman-Pearson hypothesis test of H_0 vs H_1, we use the ratio of their likelihoods as our test statistic:

$$\lambda(\vec{x}) = \frac{L(\vec{x}|H_0)}{L(\vec{x}|H_1)} = \frac{L(\vec{x})}{\text{constant}}$$

So, the "g.o.f." statistic can be re-interpreted as suitable for a hypothesis test that indicates which of H_0 (our p.d.f.) and H_1 (a flat p.d.f.) is more favored by the data—a well established statistical practice.

The benefit of the new interpretation is that it explains behaviors that were baffling under the g.o.f. interpretation: Neyman-Pearson hypothesis tests and g.o.f. tests behave quite differently.

For example, a reasonable g.o.f. statistic should be at least approximately distribution independent, but $\lambda(\vec{x})$ is often highly correlated with the m.l.e.'s

(100% in our exponential case). This high correlation was confirmed in the example contributed by K. Kinoshita[3] to the 2002 Durham Conference. Not knowing the true value of the parameters then makes it difficult, or impossible, to use $\lambda(\vec{x})$ as g.o.f., since we don't know what $\lambda(\vec{x})$ *should* be.[2] The behavior of these correlations is natural and obvious in the hypothesis test picture: changing the parameters changes the "flatness" of the H_0 p.d.f., and $\lambda(\vec{x})$ reflects this.

Reference [1] pointed out that, with no unknown parameters, one can always transform the p.d.f. to a flat distribution. Then $\lambda(\vec{x})$ becomes constant independent of the data—bad news for g.o.f. In the hypothesis test picture, this becomes a comparison between two identical hypotheses, and the result is what we would expect.

6. TEST BIAS

Take the H_0 p.d.f. to be

$$e^{-t} \qquad (t \geq 0)$$

This distribution is fully specified—no unknown parameters. Our "g.o.f." statistic is then

$$-\ln L = N\hat{t}$$

whose mean is $\langle -\ln L \rangle = N$, and variance is $\mathrm{Var}(-\ln L) = N$, for an ensemble of data sets from the H_0 p.d.f. A data set with $\hat{t}$ close enough to 1 will be claimed to be a good fit to the H_0 p.d.f.

But say, unknown to us, the data are really from a triangular p.d.f.:

$$1 - |t-1| \qquad (0 \leq t \leq 2)$$

The mean and variance of $N\hat{t}$ will be N and $N/6$ respectively, for data from the triangular distribution. So, although the exponential and triangular p.d.f.'s are quite different, the triangular data will be more likely to pass the g.o.f. test than exponential data for which it was intended. Statisticians refer to this situation as a case of "test bias".

We conclude that, even with no free parameters, the "g.o.f." test is biased: there exist "impostor" p.d.f.'s that should produce bad fits, but instead pass the "g.o.f." test with greater probability then the p.d.f. for which the test was designed. Reference [4] gives additional examples of this behavior.

From the hypothesis test point of view, this behavior makes sense. The exponential and triangular data have the same "distance" from the flat distribution, on the average, with the triangular data being less susceptible to fluctuations. The hypothesis test doesn't tell us when the data are inconsistent with both H_0 and H_1.

7. ANOTHER EXAMPLE

Here we try to find an example p.d.f. (with a free parameter) that the method in question can handle well. We use the insight provided by the hypothesis test picture. We want to keep the correlation between the free parameter and the g.o.f. statistic $L_{\max}$ to a minimum. In the hypothesis test picture, this is achieved when the "flatness" of the p.d.f. is independent of the parameter. A location parameter has this property. Additionally, we want the p.d.f. to be easily distinguishable from a flat p.d.f. So we choose the Gaussian

$$\frac{1}{\sqrt{2\pi}\sigma} e^{-0.5(x-\mu)^2/\sigma^2}$$

where μ is unknown, but σ is specified in advance. The likelihood is given by

$$-\ln L = \sum_{i=1}^{N} \left[\ln\sqrt{2\pi} + \ln\sigma + \frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2 \right]$$

When μ and σ are both unknown, their m.l.e.'s are

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad \hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i-\hat{\mu})^2$$

Using these expressions, we can rewrite the likelihood in the form $L(\mu,\sigma;\hat{\mu},\hat{\sigma})$:

$$-\ln L = \frac{N}{2}\left[\ln(2\pi) + \ln(\sigma^2) + \frac{\hat{\sigma}^2 + (\hat{\mu}-\mu)^2}{\sigma^2}\right]$$

When only μ is unspecified, its m.l.e. is $\hat{\mu}$ as above, and the value of the maximized likelihood is

$$-\ln L_{\max} = \frac{N}{2}\left[\ln(2\pi) + \ln(\sigma^2) + \frac{\hat{\sigma}^2}{\sigma^2}\right]$$

Our victory is that $L_{\max}$ only depends on $\hat{\sigma}$, which is an ancillary statistic for μ. That is, we don't need to know the true value of μ in order to calculate the distribution of our g.o.f. statistic in this carefully chosen example. In fact, a convenient form for the g.o.f. statistic is

$$N\frac{\hat{\sigma}^2}{\sigma^2} = \sum_{i=1}^{N}\left(\frac{x_i-\hat{\mu}}{\sigma}\right)^2$$

[2] Small correlations are not fatal. For example, if the P-value of g.o.f. for the observed data in a particular case ranged only between, say, 20% and 30%, for different true values within $\pm 3\sigma$ of the estimated value of a parameter, one would be justified in concluding "good fit" (assuming the g.o.f. statistic used had the right properties in other respects).

which is well known to have the distribution (under the null hypothesis) of a χ^2 with $N-1$ degrees of freedom.

7.1. The Bad News

Before we declare that the method performs well in this example, there are several ugly facts to consider:

- Data that match the null hypothesis well yield $N\hat{\sigma}^2/\sigma^2 \simeq N$. Much larger or much smaller values of the g.o.f. statistic imply poor g.o.f. This is in contrast to Pearson's χ^2 (binned χ^2), for example, where smaller χ^2 is always better g.o.f. So we must interpret this statistic differently than how we are used to.
- The g.o.f. in this example simply reduces to a comparison between the sample variance and σ^2. Any distribution with variance approximately equal to σ^2 will usually generate data that "pass the test", even distributions that look nothing like a Gaussian. This is the same kind of problem that we first saw in section 3.1.
- A construction similar to that of section 6 will produce "impostor" p.d.f.'s that pass the "g.o.f." test with greater frequency than the null hypothesis. So, we have not eliminated the test bias problem.

In this example, the g.o.f. method in question will be able to flag some, but not all, of data samples that poorly match the null hypothesis. In answer to the question "Are the data from a Gaussian with unspecified mean, and variance equal to σ^2?", this g.o.f. method can only answer "No" or "Maybe": it checks the variance part of the question, but does nothing to check the Gaussian part.

8. CONCLUSIONS

- This "g.o.f." method is fatally flawed in the unbinned case. Don't use it. Complain when you see it used.
- With fixed p.d.f.'s, the method suffers from test bias, and is not invariant with respect to change of variables. These problems persist when there are floating parameters.
- With floating parameters, the method is often circular: "g.o.f." becomes a comparison between the measured values and the true (but unknown) values of the parameters. . .
- The misbehavior of this "g.o.f." statistic is understandable when reinterpreted as the ratio between the likelihood in question and a uniform likelihood, and used to distinguish between these two specific hypotheses. Dual-hypothesis tests are not g.o.f. tests.

Acknowledgments

I would like to thank Louis Lyons for several helpful discussions of the points raised here, and the organizers of the PHYSTAT2003 Conference for arranging a superb program.

References

[1] W.T. Eadie, D. Drijard, F.E. James, M. Roos, and B. Sadoulet, *Statistical Methods in Experimental Physics*, chapter 11, pages 268, 271, (North-Holland Publishing Co, Amsterdam, 1971).

[2] J.G. Heinrich, "Can the likelihood function be used to measure goodness-of-fit?", CDF Internal Note 5639, (2001). `www-cdf.fnal.gov/publications/cdf5639_goodnessoffitv2.ps.gz`

[3] K. Kinoshita, "Evaluating quality of fit in unbinned maximum likelihood fitting", in *Proceedings of the Conference on Advanced Techniques in Particle Physics*, edited by M. Whalley and L. Lyons, p 176, (2002). `www.ippp.dur.ac.uk/Workshops/02/statistics/proceedings/kinoshita.ps`

[4] J.G. Heinrich, "Unbinned likelihood as goodness-of-fit for fixed distributions: A critical review", CDF Internal Note 6123, (2002). `www-cdf.fnal.gov/publications/cdf6123_gof_like_fixed.ps`

An Unbinned Goodness-of-Fit Test Based on the Random Walk

K. Kinoshita
University of Cincinnati, Cincinnati, OH 45221 USA

We describe a test statistic for unbinned goodness-of-fit of data in one dimension. The statistic is based on the two-dimensional Random Walk. The rejection power of this test is explored both for simple and compound hypotheses and, for the examples explored, it is found to be comparable to that for the χ^2 test. We discuss briefly how it may be possible to extend this test to multi-dimensional data.

1. INTRODUCTION

This search for an unbinned goodness-of-fit test has been motivated by the widespread use of unbinned maximum likelihood fitting for determining CP-violating parameters at Belle. While there are many cross-checks to insure that there are no spurious signals and biases, the fits tend to be complicated and not very transparent. They often involve probability density functions (PDF's) that differ with every event, based on measured quantities that add dimensions to the data that are not explicit in the fits. As there is no widely accepted unbinned goodness-of-fit test that applies to such fits, testing for statistical consistency of results has been uneven. The tests that have been done, resorting to binned χ^2 or toy Monte Carlo, have their place but have not been entirely satisfactory in addressing the question.

A common technique of unbinned tests involves first transforming the measured quantities to a variable in which the null hypothesis has a uniform distribution, where the PDF is flat, and then to test this "flattened" distribution for consistency with uniformity. There exists a variety of tests for uniformity, but most are not readily extended to multidimensional data, and they do not address compound hypotheses. A review of methods is given in [1].

In this report, we explore a test statistic that is based on the two-dimensional Random Walk. To begin, its distribution in the case of a flat PDF is discussed. The ensemble distribution is then found for several alternate hypotheses, and the rejection power is calculated for comparison with other goodness-of-fit tests. As the aim of a goodness-of-fit test as it would be applied at Belle is to test the validity of the parametrization used in fitting, it is also important to examine how the test is modified under compound hypotheses. The discussion is thus expanded to include data which are fitted to determine one or more parameters. Finally, we discuss the possibility of extending to multidimensional data.

2. RANDOM WALK AS A TEST OF FLATNESS

A data set consisting of N measurements of the one-dimensional quantity x lying in the interval $[0, 1]$ may be mapped trivially to points on a unit circle with polar angle ϕ on the interval $[0, 2\pi]$, so that each point is considered to be a unit vector with direction defined by ϕ. If the PDF in x is flat, the vector sum of the corresponding unit vectors in two dimensions corresponds to the net displacement, D, after a two-dimensional Random Walk of N steps with unit step size. For sufficiently large N, this distribution converges to a well-known form (Rayleigh, 1888) and the distribution in D^2 is an exponential decay with mean equal to N. We take D^2/N as the test statistic. A deviation of the root distribution from the hypothesis will result in a bias of the ensemble distribution of this test statistic away from the origin. This statistic is mathematically equivalent to the first order term in the Fourier series that describes the distribution of the data:

$$\begin{aligned}\mathcal{F}(k=1) &= \int_0^{2\pi} d\phi \sum_{j=1}^{N} e^{ik\phi}\delta(\phi-\phi_j) \qquad (1)\\ &= \sum_{j=1}^{N} e^{i\phi_j}\end{aligned}$$

where one can see that $D^2 \propto |\mathcal{F}(1)|^2$. One would expect this distribution to be most sensitive to an overall imbalance of the PDF in generally opposite ϕ directions. To obtain sensitivity to higher order differences, one could thus take successively higher order terms in the series, for $k = 2, ...$ In practice it may not be useful to examine terms above $k = 3$. In this study we look at $k = 1$ ($d = 1$) and define $K_k \equiv \frac{|\mathcal{F}(k)|^2}{N}$. What we have defined as K_1 appears in the review of D'Agostino and Stephens[2] as R in the context of the Von Mises test, a test for uniformity on a circle.

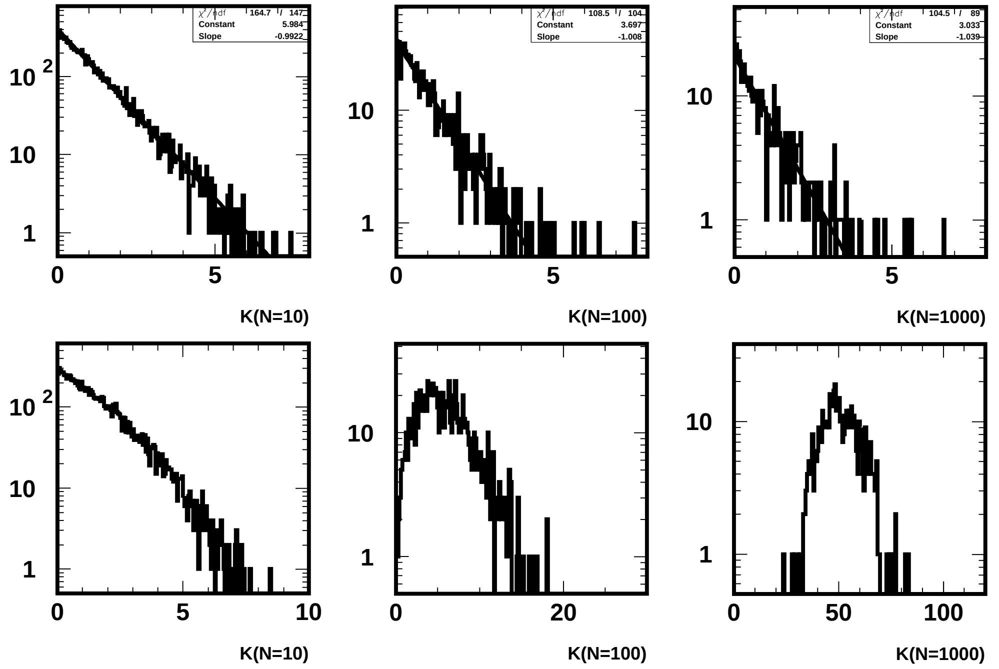


Figure 1: (top row) Distributions in K_1 for flat PDF: experiments with $N = 10$, $N = 100$, and $N = 1000$, shown with fits to an exponential form. (bottom row) Distributions in K_1 for PDF with the form $0.3 + 1.4X$ with $N = 10$, $N = 100$, and $N = 1000$.

Table I Rejection power for functions $\mathcal{A}_1$, $\mathcal{A}_2$, and $\mathcal{A}_3$ with a flat null hypothesis.

Function	Rejection Power		
	$N = 10$	$N = 100$	$N = 1000$
$\mathcal{A}_1$ (Linear)	0.117	0.824	1.00
$\mathcal{A}_2$ (Wide Gaussian)	0.152	0.910	1.00
$\mathcal{A}_3$ (Narrow Gaussian)	0.102	0.672	1.00

3. FLAT PDF

As mentioned above, the K_1 distribution for a flat PDF converges rapidly to an exponential with a decay constant of unity. Figure 1 (top row) shows the distributions in K_1 for ensembles of randomly generated experiments containing $N = 10$, 100, and 1000 events. Each of the three distributions is fitted via binned maximum likelihood to an exponential form. The fitted inverse decay constants ("slopes") are 0.992 ± 0.010, 1.008 ± 0.033, and 1.039 ± 0.049, respectively, in excellent agreement with the expectation.

To evaluate rejection power, these distributions may be compared with those obtained for PDF's that are not flat. The alternative hypotheses used in a study by Aslan and Zech [3] provide a convenient range of function types and allow for a direct comparson with the range of tests reviewed in their work. In that paper the rejection power of the alternative hypothesis was defined as one minus the probability for an error of the second kind, given a criterion that yields a 5% significance for the null hypothesis. Since in this case the null hypothesis gives an exponential distribution with unit decay constant, the 5% criterion is $K_1 > 3.0$. Ensembles of experiments were generated for each of three functions used in Ref. [3]:

$$\mathcal{A}_1(X) = 0.3 + 1.4X \qquad (2)$$

$$\mathcal{A}_2(X) = 0.7 + 0.3[n_2 e^{-64(X-0.5)^2}] \qquad (3)$$

$$\mathcal{A}_3(X) = 0.8 + 0.2[n_3 e^{-256(X-0.5)^2}] \qquad (4)$$

where the n_i are normalization constants for the associated Gaussians. All functions are defined in the interval $[0, 1]$. The resulting K_1 distributions for $\mathcal{A}_1$ are shown in Figure 1 (bottom row). The values for rejection power are summarized in Table I. For comparison, the values for the χ^2 method ($N = 100$) given by Ref. [3] are approximately 0.81, 0.85, and 0.81, respectively, so our method is comparable in power, at least in the case of these three functions.

In order to apply this method as a goodness-of-fit test for non-uniform null hypotheses the PDF, $f(X)$, must first be transformed to a "flat" variable, Y,

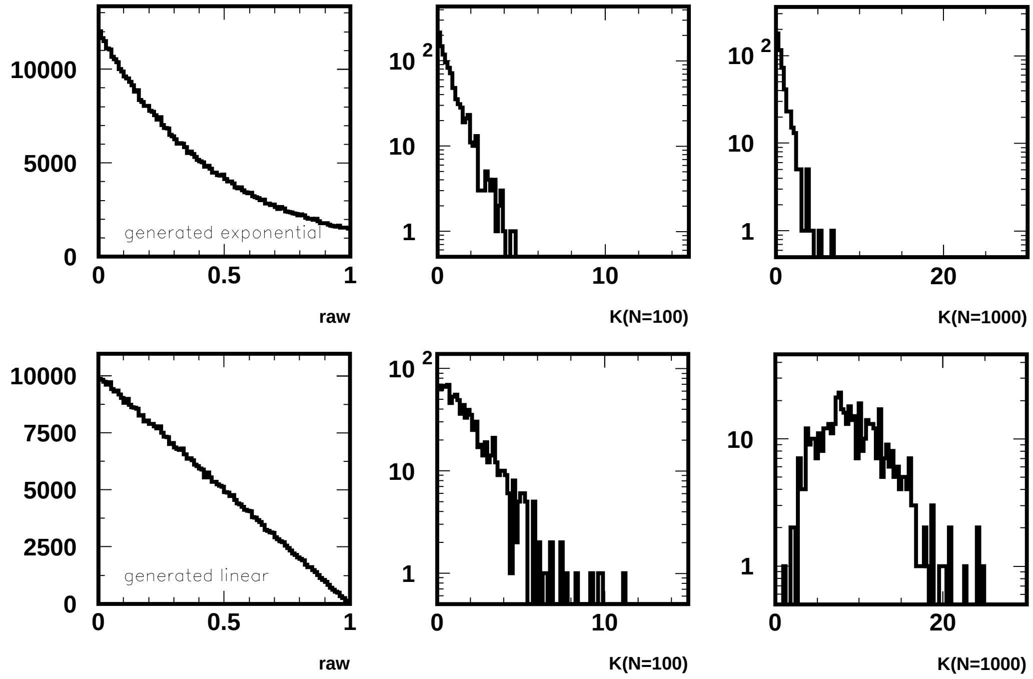


Figure 2: Determination of rejection power for a compound hypothesis: ensembles fitted for decay constant of exponentially decaying form. (top row) PDF matches fit parametrization: (left) Raw distribution, (center, right) distributions in K_1 of fitted, flattened experiments, $N = 100$ and $N = 1000$. (bottom row) PDF inconsistent with parametrization: (left) Raw distribution, (center, right) distributions in K_1 of fitted, flattened experiments, $N = 100$ and $N = 1000$.

where the probability distribution is flat. To form a uniform null hypothesis on a circle one could, for example, construct Y as :

$$Y_i \;=\; 2\pi \int_{X_-}^{X_i} f(X)dX \qquad (5)$$

where the integer subscript i denotes the i^{th} data point and X_- is the lowest possible value of X.

4. COMPOUND HYPOTHESES

The examples considered thus far have been ones where no parameter fitting has occurred. While this has been an instructive exercise, it has limited application, as most measurements in particle physics involve the fitting of measured distributions to determine shapes and to derive some physics quantity or conclusion. We now look at compound hypotheses.

In evaluating rejection of alternative hypotheses via toy MC in the compound case, it is important that the fitting process be integrated into the evaluation procedure. Consider a data set $\{\phi_i\}$ where the PDF is assumed to be parametrizable as $f(\phi;\alpha)$ and the unbinned likelihood is maximum for $\alpha = \alpha_{max}$. The data are then flattened assuming the PDF is $f(\phi;\alpha_{max})$, and the associated K_1 is evaluated. The confidence level of this K_1 value may then be found by referencing the ensemble distribution of K_1 when the true PDF is $f(\phi;\alpha_{max})$, and each experiment of the ensemble is treated as data, fitted and flattened according to the fit.

This procedure was used to evaluate rejection power for pairs of similarly shaped PDF's. Here we show one such result, for the hypothesis $n_4(\alpha)e^{-10X\alpha}$, where n_4 is a normalization constant, the measured quantity is X, and experiments are fitted for α. The alternative PDF was the linear form $f(X) = 2(1 - X)$. Experiments were generated according to the alternative PDF (A), and each was fitted to the hypothesis. The mean maximum likelihood value of α was approximately 4.7. Ensembles (B) were generated according to the hypothesis, with $\alpha = 4.7$, and fitted in the same way. The 5% confidence criterion on K_1 for (B) and acceptance of this criterion for (A) were estimated by counting (Figure 2). The rejection powers were found to be 28% and 99% for $N = 100$ and $N = 1000$, respectively. For comparison we also calculated by the same procedure the rejection of the χ^2 test, using 20 bins in the interval [0,1] and found powers of 13% and 100%, respectively.

We also examined the two-dimensional distribution of fitted α_{max} values *vs.* K_1. Any dependence of the

Table II Inverse decay constants of K_1 distribution for several generated forms, flattened after fitting for parameter(s) $\{\alpha_i\}$. The n_i are normalization constants, which may depend on the parameters α_j. No entry is made for samples where low statistics resulted in best fits which were at the limits of the parametrization.

Form	Generated	Fitted	K_1 (Decay Constant)$^{-1}$ (χ^2/ndf)		
			$N = 10$	$N = 100$	$N = 1000$
$(1-\alpha)+\alpha(2X)$	$\alpha = 0.7$	α	–	–	1.28 ± 0.07 (70/67)
$(1-\alpha)+\alpha[n_2 e^{-64(X-0.5)^2}]$	$\alpha = 0.3$	α	–	1.90 ± 0.06 (230/80)	1.94 ± 0.09 (223/65)
$(1-\alpha)+\alpha[n_3 e^{-256(X-0.5)^2}]$	$\alpha = 0.2$	α	–	1.56 ± 0.05 (203/82)	1.56 ± 0.07 (82/68)
$n_4 e^{-10X/\alpha}$	$\alpha = 1.0$	α	1.23 ± 0.01 (147/133)	1.28 ± 0.04 (68/85)	1.28 ± 0.06 (75/76)
$n_5 e^{-[X-(0.5+\alpha_2)]^2/2(\alpha_1/8)^2}$	$\alpha_1 = 1.0,$	α_1	1.36 ± 0.01 (176/131)	1.38 ± 0.05 (93/85)	1.50 ± 0.07 (56/65)
	$\alpha_2 = 0$	α_2	1.22 ± 0.01 (154/135)	1.25 ± 0.04 (122/96)	1.28 ± 0.06 (73/72)
		α_1, α_2	1.84 ± 0.019 (148/90)	2.00 ± 0.065 (53/59)	2.13 ± 0.095 (47/47)

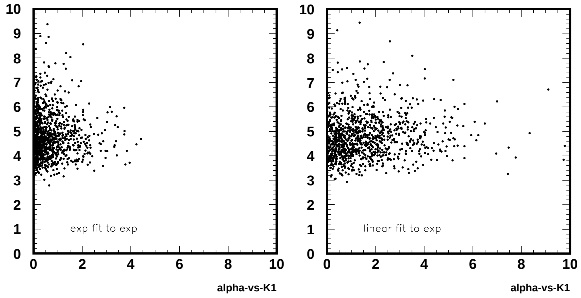


Figure 3: Scatter plots of fitted parameter α_{max} *vs.* K_1 for ensembles shown in Figure 2 ($N = 100$).

test on the fitted rather than underlying parameter value reduces its utility as a goodness-of-fit test; for example, the maximum likelihood value, $\mathcal{L}_{max}$, is not usable as a goodness-of-fit statistic because it depends strongly on the fitted parameter value(s) α_{max} – for a certain class of fitting functions, the correlation is 100%[4]. Figure 3 shows scatter plots of α_{max} and K_1, where the data were generated with $N = 100$ and the generated and fitted forms are those from the example of Figure 2. There appears to be no strong dependence.

In any determination of rejection power with a compound hypothesis, it is necessary to determine the distribution of K_1 for the correct hypothesis. It does not appear that there is a simple ansatz as in the case of binned least squares fitting, where the chisquare converges to a chisquare distribution with the number of degrees of freedom reduced by one unit for each linear fitted parameter. We study this question empirically by generating MC ensembles for a variety of shapes. Each ensemble was generated according to the fitted functional form with parameter value(s) fixed. Each experiment was fitted with parameter(s) floating, and the K_1 value was obtained from the data flattened according to the best fit. The distribution of resultant K_1 values for each ensemble was fitted for the decay constant, assuming an exponentially decaying form. Ensembles with $N = 10$, $N = 100$, and $N = 1000$ were generated. The results are summarized in Table II. There are several notable features. First, while all of the K_1 distributions had a decaying form, as one might expect, and a fit that converged, not all yielded good fits; the exponential form is not preserved under compound hypotheses. Secondly, all inverse decay constants are greater than unity, indicating that the K_1 distribution moves toward zero with fitting. This is not surprising; fitting identifies for each experiment the shape that is "closest" to the data, giving in general a better goodness-of-fit than the generator shape. Finally, there is no obvious pattern in the value of the decay constant with number of floated parameters. However, it is seen that for a given PDF and set of fitted parameters, the shape of the K_1 distribution shows remarkably little change as N is changed by two orders of magnitude.

5. EXTENSION TO MULTIDIMENSIONAL DATA: SPECULATION

Our goal in this investigation has been to arrive at a multidimensional unbinned goodness-of-fit test, one that has rejection power in all dimensions, not just in one-dimensional projections, for multidimensional data. Many unbinned tests depend on the integrated sum of or spacings between neighboring data points, quantities which are not well-defined when extended to more than one dimension. Although the K_1 statistic does not have this property, it is yet to be determined whether there exists an extension that is fully multidimensional; for example, in two dimensions, two components each mapped to a circle corresponds to a data space that is the surface of a toroid, for which there is no obvious nontrivial vector sum that maps to the Random Walk. A fully general extension to multidimensional data will additionally require a flattening algorithm and provisions for data spaces of arbitrary shape. We will continue to explore the possibilities for extending K_1 for use with multidimensional data.

6. SUMMARY

We have explored an unbinned goodness-of-fit test for data in one dimension that is based on the mapping of flattened distributions to a two-dimensional random walk. This method is truly binning-free and scale-independent, and the ensemble distribution for the null hypothesis is well-defined. For a compound hypothesis we specify a procedure to determine the ensemble distribution of the test statistic via Monte Carlo so that rejection power may be readily determined. The distribution is found for several different parametrized forms and shown to be largely independent of statistics. We examine several samples for dependence between the test statistic and fitted parameter values, and find no evidence of any. The rejection power for alternate hypotheses is demonstrated for a few examples and is found to be comparable to that of the chisquare method.

Acknowledgments

The author would like to thank R. Cousins, G. Zech, and B. Yabsley for useful discussions and suggestions, and the organizers of PHYSTAT2003 for a stimulating and interesting conference. This work is supported by Department of Energy grant #DE-FG02-84ER40153.

References

[1] B. Aslan and G. Zech, in *Proc. Conf. on Advanced Statistical Techniques in Particle Physics*, M.R. Whalley and L.Lyons, eds. (2002). `http://www.ippp.dur.ac.uk/Workshops/02/statistics/proceedings.shtml`

[2] *Goodness-Of-Fit Techniques (Statistics: Textbooks and Monographs Series, Vol. 68)*, R.B. D'Agostino and M.A. Stephens, eds., Marcel Dekker, Inc (1986).

[3] B. Aslan and G. Zech, hep-ex/0203010 (2002).

[4] J. Heinrich, "Can the Likelihood Function Value Be Used to Measure Goodness-of-Fit?" /CDF/MEMO/BOTTOM/CDFR/5639 (unpublished); K. Kinoshita, in *Proc. Conf. on Advanced Statistical Techniques in Particle Physics*, M.R. Whalley and L.Lyons, eds. (2002).

A Measure of the Goodness of Fit in Unbinned Likelihood Fits; End of Bayesianism?

Rajendran Raja
Fermilab, Batavia, IL 60510, USA

Maximum likelihood fits to data can be done using binned data (histograms) and unbinned data. With binned data, one gets not only the fitted parameters but also a measure of the goodness of fit. With unbinned data, currently, the fitted parameters are obtained but no measure of goodness of fit is available. This remains, to date, an unsolved problem in statistics. Using Bayes' theorem and likelihood ratios, we provide a method by which both the fitted quantities and a measure of the goodness of fit are obtained for unbinned likelihood fits, as well as errors in the fitted quantities. The quantity, conventionally interpreted as a Bayesian prior, is seen in this scheme to be a number not a distribution, that is determined from data.

1. INTRODUCTION

As of the Durham conference [1], the problem of obtaining a goodness of fit in unbinned likelihood fits was an unsolved one. In what follows, we will denote by the vector s, the theoretical parameters (s for "signal") and the vector c, the experimentally measured quantities or "configurations". For simplicity, we will illustrate the method where both s and c are one dimensional, though either or both can be multi-dimensional in practice. We thus define the theoretical model by the conditional probability density $P(c|s)$. Then an unbinned maximum likelihood fit to data is obtained by maximizing the likelihood [2],

$$\mathcal{L} = \prod_{i=1}^{i=n} P(c_i|s) \tag{1}$$

where the likelihood is evaluated at the n observed data points $c_i, i = 1, n$. Such a fit will determine the maximum likelihood value s^* of the theoretical parameters, but will not tell us how good the fit is. The value of the likelihood $\mathcal{L}$ at the maximum likelihood point does not furnish a goodness of fit, since the likelihood is not invariant under change of variable. This can be seen by observing that one can transform the variable set c to a variable set c' such that $P(c'|s^*)$ is uniformly distributed between 0 and 1. Such a transformation is known as a hypercube transformation, in multi-dimensions. Other datasets will yield different values of likelihood in the variable space c when the likelihood is computed with the original function $P(c|s^*)$. However, in the original hypercube space, the value of the likelihood is unity regardless of the dataset $c'_i, i = 1, n$, thus the likelihood $\mathcal{L}$ cannot furnish a goodness of fit by itself, since neither the likelihood, nor ratios of likelihoods computed using the same distribution $P(c|s^*)$ is invariant under variable transformations. The fundamental reason for this non-invariance is that only a single distribution, namely, $P(c|s^*)$ is being used to compute the goodness of fit.

2. LIKELIHOOD RATIOS

In binned likelihood cases, where one is comparing a theoretical distribution $P(c|s)$ with a binned histogram, there are two distributions involved, the theoretical distribution and the data distribution. The *pdf* of the data is approximated by the bin contents of the histogram normalized to unity. If the data consists of n events, the *pdf* of the data $P^{data}(c)$ is defined in the frequentist sense as the normalized density distribution in c space of n events as $n \to \infty$. In the binned case, we can bin in finer and finer bins as $n \to \infty$ and obtain a smooth function, which we define as the *pdf* of the data $P^{data}(c)$. In practice, one is always limited by statistics and the binned function will be an approximation to the true *pdf*. We can now define a likelihood ratio $\mathcal{L}_\mathcal{R}$ such that

$$\mathcal{L}_\mathcal{R} = \frac{\prod_{i=1}^{i=n} P(c_i|s)}{\prod_{i=1}^{i=n} P^{data}(c_i)} \equiv \frac{P(\mathbf{c_n}|s)}{P^{data}(\mathbf{c_n})} \tag{2}$$

where we have used the notation $\mathbf{c_n}$ to denote the event set $c_i, i = 1, n$. Let us now note that $\mathcal{L}_\mathcal{R}$ is invariant under the variable transformation $c \to c'$, since

$$P(c'|s) = |\frac{dc}{dc'}|P(c|s) \tag{3}$$

$$P^{data}(c') = |\frac{dc}{dc'}|P^{data}(c) \tag{4}$$

$$\mathcal{L}'_\mathcal{R} = \mathcal{L}_\mathcal{R} \tag{5}$$

and the Jacobian of the transformation $|\frac{dc}{dc'}|$ cancels in the numerator and denominator in the ratio. This is an extremely important property of the likelihood ratio $\mathcal{L}_\mathcal{R}$ that qualifies it to be a goodness of fit variable. Since the denominator $P^{data}(\mathbf{c_n})$ is independent of the theoretical parameters s, both the likelihood ratio and the likelihood maximize at the same point s^*. One can also show [3] that the maximum value of the likelihood ratio occurs when the theoretical likelihood $P(c_i|s)$ and the data likelihood $P^{data}(c_i)$ are equal for all c_i.

3. BINNED GOODNESS OF FIT

In the case where the *pdf* $P^{data}(c)$ is estimated by binned histograms and the statistics are Gaussian, it is readily shown [3] that the commonly used goodness of fit variable $\chi^2 = -2log\mathcal{L_R}$. It is worth emphasizing that the likelihood ratio as defined above is needed and not just the negative log of theoretical likelihood $P(\mathbf{c_n}|s)$ to derive this result. The popular conception that χ^2 is -2 log $P(\mathbf{c_n}|s)$ is simply incorrect!. It can also be shown that the likelihood ratio defined above can describe the binned cases where the statistics are Poissonian [4]. In order to solve our problem of goodness of fit in unbinned likelihood cases, one needs to arrive at a method of estimating the data *pdf* $P^{data}(c)$ without the use of bins.

4. UNBINNED GOODNESS OF FIT

One of the better known methods of estimating the probability density of a distribution in an unbinned case is by the use of Probability Density Estimators ($PDE's$), also known as Kernel Density Estimators [5] ($KDE's$). The *pdf* $P^{data}(c)$ is approximated by

$$P^{data}(c) \approx PDE(c) = \frac{1}{n}\sum_{i=1}^{i=n}\mathcal{G}(c-c_i) \qquad (6)$$

where a Kernel function $\mathcal{G}(c-c_i)$ is centered around each data point c_i, is so defined that it normalizes to unity and for large n approaches a Dirac delta function [3]. The choice of the Kernel function can vary depending on the problem. A popular kernel is the Gaussian defined in the multi-dimensional case as

$$\mathcal{G}(c) = \frac{1}{(\sqrt{2\pi}h)^d\sqrt{(det(E))}}exp(\frac{-H^{\alpha\beta}c^\alpha c^\beta}{2h^2}) \qquad (7)$$

where E is the error matrix of the data defined as

$$E^{\alpha,\beta} = < c^\alpha c^\beta > - < c^\alpha >< c^\beta > \qquad (8)$$

and the $<>$ implies average over the n events, and d is the number of dimensions. The Hessian matrix H is defined as the inverse of E and the repeated indices imply summing over. The parameter h is a "smoothing parameter", which has[6] a suggested optimal value $h \propto n^{-1/(d+4)}$, that satisfies the asymptotic condition

$$\mathcal{G}_\infty(c-c_i) \equiv \lim_{n\to\infty}\mathcal{G}(c-c_i) = \delta(c-c_i) \qquad (9)$$

The parameter h will depend on the local number density and will have to be adjusted as a function of the local density to obtain good representation of the data by the PDE. Our proposal for the goodness of fit in unbinned likelihood fits is thus the likelihood ratio

$$\mathcal{L_R} = \frac{P(\mathbf{c_n}|s)}{P^{data}(\mathbf{c_n})} \approx \frac{P(\mathbf{c_n}|s)}{P^{PDE}(\mathbf{c_n})} \qquad (10)$$

evaluated at the maximum likelihood point s^*.

5. AN ILLUSTRATIVE EXAMPLE

We consider a simple one-dimensional case where the data is an exponential distribution, say decay times of a radioactive isotope. The theoretical prediction is given by

$$P(c|s) = \frac{1}{s}\exp(-\frac{c}{s}) \qquad (11)$$

We have chosen an exponential with $s = 1.0$ for this example. The Gaussian Kernel for the PDE would be given by

$$\mathcal{G}(c) = \frac{1}{(\sqrt{2\pi}\sigma h)}\exp(-\frac{c^2}{2\sigma^2h^2}) \qquad (12)$$

where the variance σ of the exponential is numerically equal to s. To begin with, we chose a constant value for the smoothing parameter, which for 1000 events generated is calculated to be 0.125. Figure 1 shows the generated events, the theoretical curve $P(c|s)$ and the PDE curve $P(c)$ normalized to the number of events. The PDE fails to reproduce the data near the origin due to the boundary effect, whereby the Gaussian probabilities for events close to the origin spill over to negative values of c. This lost probability would be compensated by events on the exponential distribution with negative c if they existed. In our case, this presents a drawback for the PDE method, which we will remedy later in the paper using PDE definitions on the hypercube and periodic boundary conditions. For the time being, we will confine our example to values of $c > 1.0$ to avoid the boundary effect.

In order to test the goodness of fit capabilities of the likelihood ratio $\mathcal{L_R}$, we superimpose a Gaussian on the exponential and try and fit the data by a simple exponential. Figure 2 shows the "data" with 1000 events generated as an exponential in the fiducial range $1.0 < c < 5.0$. Superimposed on it is a Gaussian of 500 events. More events in the exponential are generated in the interval $0.0 < c < 1.0$ to avoid the boundary effect at the fiducial boundary at c=1.0. Since the number density varies significantly, we have had to introduce a method of iteratively determining the smoothing factor as a function of c as described in [3]. With this modification in the PDE, one gets a good description of the behavior of the data by the PDE as shown in Figure 2. We now vary the number of events in the Gaussian and obtain the value of the negative log likelihood ratio $\mathcal{NLLR}$ as a function of the strength of the Gaussian. Table I summarizes the results. The number of standard deviations the unbinned likelihood fit is from what is expected is determined empirically by plotting the value of $\mathcal{NLLR}$ for a large number of fits where no Gaussian is superimposed (i.e. the null hypothesis) and determining the mean and RMS of this distribution and using these

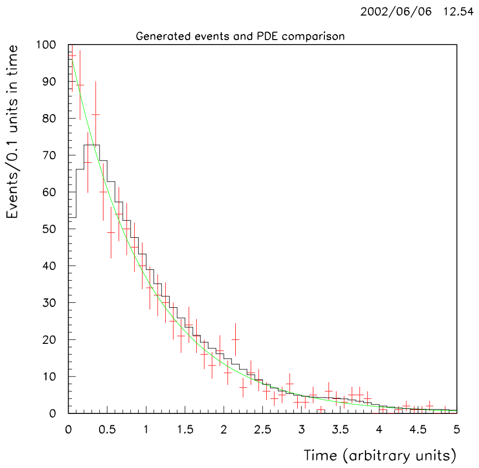


Figure 1: Figure shows the histogram (with errors) of generated events. Superimposed is the theoretical curve $P(c|s)$ and the PDE estimator (solid) histogram with no errors.

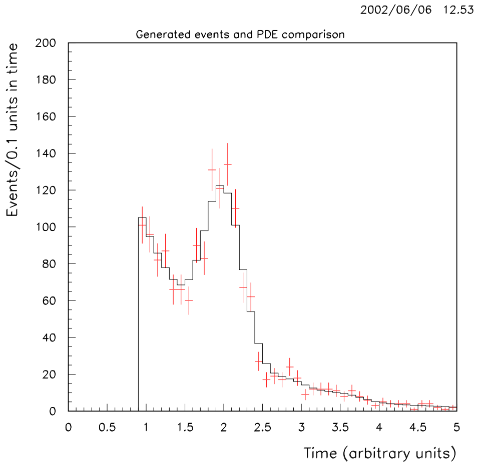


Figure 2: Figure shows the histogram (with errors) of 1000 events in the fiducial interval $1.0 < c < 5.0$ generated as an exponential with decay constant s=1.0 with a superimposed Gaussian of 500 events centered at c=2.0 and width=0.2. The PDE estimator is the (solid) histogram with no errors.

to estimate the number of σ's the observed $\mathcal{NLLR}$ is from the null case. Table I also gives the results of a binned fit on the same "data". It can be seen that the unbinned fit gives a 3σ discrimination when the number of Gaussian events is 85, where as the binned fit gives a χ^2/ndf of 42/39 for the same case. We intend to make these tests more sophisticated in future work.

Figure 3 shows the variation of -log $P(\mathbf{c_n}|s)$ and -log $P^{PDE}(\mathbf{c_n})$ for an ensemble of 500 experiments each with the number of events $n = 1000$ in the exponential and no events in the Gaussian (null hypothesis).

Table I Goodness of fit results from unbinned likelihood and binned likelihood fits for various data samples. The negative values for the number of standard deviations in some of the examples is due to statistical fluctuation.

Number of	Unbinned fit	Unbinned fit	Binned fit χ^2
Gaussian events	$\mathcal{NLLR}$	$N\sigma$	39 d.o.f.
500	189.	103	304
250	58.6	31	125
100	11.6	4.9	48
85	8.2	3.0	42
75	6.3	1.9	38
50	2.55	-0.14	30
0	0.44	-1.33	24

It can be seen that -log $P(\mathbf{c_n}|s)$ and -log $P^{PDE}(\mathbf{c_n})$ are correlated with each other and the difference between the two (-log $\mathcal{NLLR}$) is a much narrower distribution than either and provides the goodness of fit discrimination.

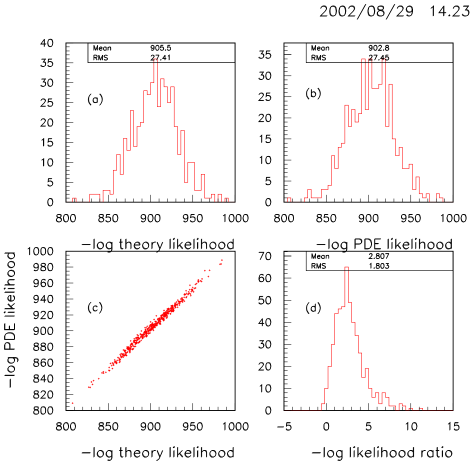


Figure 3: (a) shows the distribution of the negative log-likelihood -$log_e(P(\mathbf{c_n}|s))$ for an ensemble of experiments where data and experiment are expected to fit. (b) Shows the negative log PDE likelihood -$log_e(P(\mathbf{c_n}))$ for the same data (c) Shows the correlation between the two and (d) Shows the negative log-likelihood ratio $\mathcal{NLLR}$ that is obtained by subtracting (b) from (a) on an event by event basis.

5.1. Improving the PDE

The PDE technique we have used so far suffers from two drawbacks; firstly, the smoothing parameter has to be iteratively adjusted significantly over the full range of the variable c, since the distribution $P(c|s)$ changes significantly over that range; and secondly, there are boundary effects at c=0 as shown in figure 1. Both these flaws are remedied if we define the

PDE in hypercube space. After we find the maximum likelihood point s^*, for which the PDE is not needed, we transform the variable $c \to c'$, such that the distribution $P(c'|s^*)$ is flat and $0 < c' < 1$. The hypercube transformation can be made even if c is multi-dimensional by initially going to a set of variables that are uncorrelated and then making the hypercube transformation. The transformation can be such that any interval in c space maps on to the interval $(0,1)$ in hypercube space. We solve the boundary problem by imposing periodicity in the hypercube. In the one dimensional case, we imagine three "hypercubes", each identical to the other on the real axis in the intervals $(-1,0)$, $(0,1)$ and $(1,2)$. The hypercube of interest is the one in the interval $(0,1)$. When the probability from an event kernel leaks outside the boundary $(0,1)$, we continue the kernel to the next hypercube. Since the hypercubes are identical, this implies the kernel re-appearing in the middle hypercube but from the opposite boundary. Put mathematically, the kernel is defined such that

$$\mathcal{G}(c' - c'_i) = \mathcal{G}(c' - c'_i - 1);\ c' > 1 \tag{13}$$

$$\mathcal{G}(c' - c'_i) = \mathcal{G}(c' - c'_i + 1);\ c' < 0 \tag{14}$$

Although a Gaussian Kernel will work on the hypercube, the natural kernel to use considering the shape of the hypercube would be the function $\mathcal{G}(c')$

$$\mathcal{G}(c') = \frac{1}{h};\ |c'| < \frac{h}{2} \tag{15}$$

$$\mathcal{G}(c') = 0;\ |c'| > \frac{h}{2} \tag{16}$$

This kernel would be subject to the periodic boundary conditions given above, which further ensure that every event in hypercube space is treated exactly as every other event irrespective of their co-ordinates. The parameter h is a smoothing parameter which needs to be chosen with some care. However, since the theory distribution is flat in hypercube space, the smoothing parameter may not need to be iteratively determined over hypercube space to the extent that data distribution is similar to the theory distribution. Even if iteration is used, the variation in h in hypercube space is likely to be much smaller.

Figure 4 shows the distribution of the $\mathcal{NLLR}$ for the null hypothesis for an ensemble of 500 experiments each with 1000 events as a function of the smoothing factor h. It can be seen that the distribution narrows considerably as the smoothing factor increases. We choose an operating value of 0.2 for h and study the dependence of the $\mathcal{NLLR}$ as a function of the number of events ranging from 100 to 1000 events, as shown in figure 5. The dependence on the number of events is seen to be weak, indicating good behavior. The PDE thus arrived computed with h=0.2 can be transformed from the hypercube space to c space and will reproduce data smoothly and with no edge effects. We note that it is also easier to arrive at an analytic theory of $\mathcal{NLLR}$ with the choice of this simple kernel.

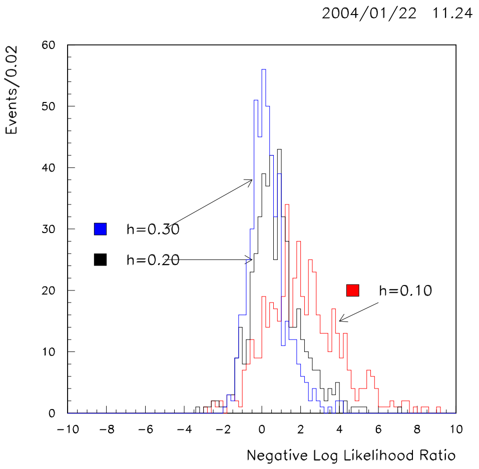


Figure 4: The distribution of the negative log likelihood ratio $\mathcal{NLLR}$ for the null hypothesis for an ensemble of 500 experiments each with 1000 events, as a function of the smoothing factor h=0.1, 0.2 and 0.3

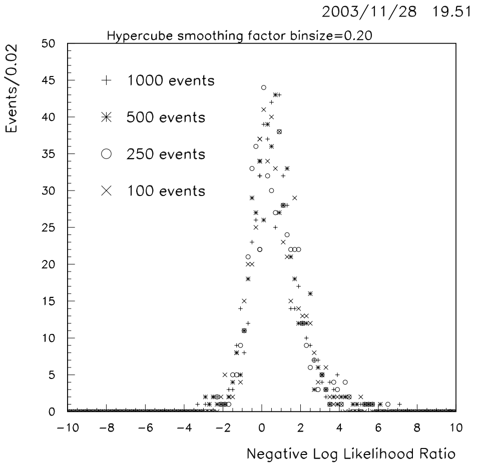


Figure 5: The distribution of the negative log likelihood ratio $\mathcal{NLLR}$ for the null hypothesis for an ensemble of 500 experiments each with the smoothing factor h=0.2, as a function of the number of events

6. END OF BAYESIANISM?

By Bayesianism, we mean the practice of "guessing" a prior distribution and introducing it into the calculations. In what follows we will show that what is conventionally thought of as a Bayesian prior distribution is in reality a number that can be calculated from the data. We are able to do this since we use two *pdf*'s, one for theory and one for data. In what follows, we will interpret the probability distribution of the parameter s in a strictly frequentist sense. The *pdf* of s is the distribution of the best estimator of the true value s_T of s from an ensemble of an infinite number of identical experiments with the same statistical power n.

6.1. Calculation of fitted errors

After the fitting is done and the goodness of fit is evaluated, one needs to work out the errors on the fitted quantities. One needs to calculate the posterior density $P(s|\mathbf{c_n})$, which carries information not only about the maximum likelihood point s^*, from a single experiment, but how such a measurement is likely to fluctuate if we repeat the experiment. The joint probability density $P(s,\mathbf{c_n})$ of observing the parameter s and the data $\mathbf{c_n}$ is given by

$$P^{data}(s,\mathbf{c_n}) = P(s|\mathbf{c_n})P^{data}(\mathbf{c_n}) \qquad (17)$$

where we use the superscript ^{data} to distinguish the joint probability $P^{data}(s,\mathbf{c_n})$ as having come from using the data *pdf*. If we now integrate the above equation over all possible datasets $\mathbf{c_n}$, we get the expression for the *pdf* of s.

$$\mathcal{P}_n(s) = \int P^{data}(s,\mathbf{c_n})d\mathbf{c_n} = \int P(s|\mathbf{c_n})P^{data}(\mathbf{c_n})d\mathbf{c_n} \qquad (18)$$

where we have used the symbol $\mathcal{P}$ to distinguish the fact that it is the true *pdf* of s obtained from an infinite ensemble. We use the subscript n in $\mathcal{P}_n(s)$ to denote that the *pdf* is obtained from an ensemble of experiments with n events each. Later on we will show that $\mathcal{P}_n(s)$ is indeed dependent on n. Equation 18 states that in order to obtain the *pdf* of the parameter s, one needs to add together the conditional probabilities $P(s|\mathbf{c_n})$ over an ensemble of events, each such distribution weighted by the "data likelihood" $P^{data}(\mathbf{c_n})$. At this stage of the discussion, the functions $P^{data}(s|\mathbf{c_n})$ are unknown functions. We have however worked out $\mathcal{L}_{\mathcal{R}}(s)$ as a function of s and have evaluated the maximum likelihood value s^* of s. We can choose an arbitrary value of s and evaluate the goodness of fit at that value using the likelihood ratio. When we choose an arbitrary value of s, we are in fact hypothesizing that the true value s_T is at this value of s. $L_R(s)$ then gives us a way of evaluating the relative goodness of fit of the hypothesis as we change s. Let us now take an arbitrary value of s and hypothesize that that is the true value. Then the joint probability of observing $\mathbf{c_n}$ and s_T being at this value of s is given from the data end by equation 17.

Similarly, from the theoretical end, one can calculate the joint probability of observing the dataset $\mathbf{c_n}$, with the true value being at s. The true value s_T is taken to be the maximum likelihood point of the *pdf* $\mathcal{P}_n(s)$. It may coincide with the mean value of the *pdf* $\mathcal{P}_n(s)$. These statements are assertions of the unbiased nature of the data from the experiment. At this point, there is no information available on where the true value s_T lies. One can make the hypothesis that a particular value of s is the true value and the probability of obtaining a best estimator s^* from experiments of the type being performed in the interval s_T and $s_T + ds_T$ is $\mathcal{P}_n(s_T)ds_T$. The actual value of this number is a function of the experimental resolution and the statistics n of the experiment. The joint probability $P^{theory}(s,\mathbf{c_n})$ from the theoretical end is given by the product of the probability density of the *pdf* of s at the true value of s, namely $\mathcal{P}_n(s_T)$, and the theoretical likelihood $P(c_n|s)$ evaluated at the true value, which by our hypothesis is s.

$$P^{theor}(s,\mathbf{c_n}) = P^{theor}(\mathbf{c_n}|s)\mathcal{P}_n(s_T) \qquad (19)$$

The joint probability $P(s,\mathbf{c_n})$ is a joint distribution of the theoretical parameter s and data $\mathbf{c_n}$. The two ways of evaluating this (from the theoretical end and the data end) must yield the same result, for consistency. This is equivalent to equating $P^{data}(s,\mathbf{c_n})$ and $P^{theor}(s,\mathbf{c_n})$. This gives the equation

$$P(s|\mathbf{c_n})P^{data}(\mathbf{c_n}) = P^{theor}(\mathbf{c_n}|s)\mathcal{P}_n(s_T) \qquad (20)$$

which is a form of Bayes' theorem, but with two $pdf's$ (theory and data). Let us note that the above equation can be immediately re-written as a likelihood ratio

$$\mathcal{L}_{\mathcal{R}} = \frac{P(s|\mathbf{c_n})}{\mathcal{P}_n(s_T)} = \frac{P^{theor}(\mathbf{c_n}|s)}{P^{data}(\mathbf{c_n})} \qquad (21)$$

which is what is used to obtain the goodness of fit. In order to get the fitted errors, we need to evaluate $P(s|\mathbf{c_n})$ which necessitates a better understanding of what $\mathcal{P}_n(s_T)$ is in equation 20. Rearranging equation 20, one gets

$$P(s|\mathbf{c_n}) = \mathcal{L}_{\mathcal{R}}(s)\mathcal{P}_n(s_T) = \frac{P^{theor}(\mathbf{c_n}|s)}{P^{data}(\mathbf{c_n})}\mathcal{P}_n(s_T) \qquad (22)$$

6.1.1. To show that $\mathcal{P}_n(s_T)$ depends on n

In practice, in both the binned and unbinned cases, one only has an approximation to $P^{data}(\mathbf{c_n})$. As $n \to \infty$, in the absence of experimental bias, one expects to determine the parameter set s to infinite accuracy; and $P(s|\mathbf{c_n}) \to \delta(s - s_T)$, where s_T is the true value of s. However, for the null hypothesis, as $n \to \infty$, the statistical error introduced by our use of PDE in the unbinned case or by binning in the binned case becomes negligible with the result that the theory *pdf* describes the data for all c at the true value s_T. i.e.

$$\frac{P^{theor}(c|s_T)}{P^{data}(c)} \to 1 \; as \; n \to \infty \qquad (23)$$

When one evaluates the likelihood ratio $\mathcal{L}_{\mathcal{R}}$ over n events, with $n \to \infty$, the likelihood ratio does not necessarily remain unity. This is due to fluctuations in the data which grow as $\sqrt(n)$. For the binned likelihood case with n_b bins, one can show that as $n \to \infty$,

$$\mathcal{L}_{\mathcal{R}} \to e^{-\sum_{i=1}^{i=n_b} \chi_i^2/2} \to e^{-n_b/2} \tag{24}$$

This is just an example of the likelihood ratio theorem. If one uses a binned χ^2 fit, which can also be thought of as maximizing a likelihood ratio, one gets the same limit as when using binned likelihood fits. The point is that $\mathcal{L}_{\mathcal{R}}$ is finite as $n \to \infty$. In the unbinned case, we have currently no analytic theory available. However, one can argue that the binned case with the number of bins $n_b \to \infty$ and $n_b << n$ should approach the unbinned limit. In this case, the unbinned $\mathcal{L}_{\mathcal{R}}$ also is finite for infinite statistics. This implies that $\mathcal{P}_n(s_T) \to \infty$ as $n \to \infty$. i.e $\mathcal{P}_n(s_T)$ depends on n. This puts an end to the notion of a monolithic Bayesian prior interpretation for $\mathcal{P}_n(s)$.

6.1.2. To show that $\mathcal{P}_n(s_T)$ is constant with respect to s

When one varies the likelihood ratio in equation 22 as a function of s, for each value of s, one is making a hypothesis that $s = s_T$. As one changes s, a new hypothesis is being tested that is mutually exclusive from the previous one, since the true value can only be at one location. So as one changes s, one is free to move the *distribution* $\mathcal{P}_n(s)$ so that s_T is at the value of s being tested. This implies that $\mathcal{P}_n(s_T)$ does not change as one changes s and is a constant *wrt* s, which we can now write as α_n. Figure 6 illustrates these points graphically. Thus $\mathcal{P}_n(s_T)$ in our equations is a number, not a function. The distribution $\mathcal{P}_n(s)$ should not be thought of as a "prior" but as an "unknown concomitant", which depends on the statistics and the measurement capabilities of the apparatus. For a given apparatus, there are a denumerable infinity of such distributions, one for each n. These distributions become narrower as n increases and $\mathcal{P}_n(s_T) \to \infty$ as $n \to \infty$.

6.2. New form of equations

Equation 22 can now be re-written

$$P(s|\mathbf{c_n}) = \frac{P(\mathbf{c_n}|s)\alpha_n}{P^{data}(\mathbf{c_n})} \tag{25}$$

Since $P(s|\mathbf{c_n})$ must normalize to unity, one gets for α_n,

$$\alpha_n = \frac{P^{data}(\mathbf{c_n})}{\int P(\mathbf{c_n}|s)ds} = \frac{1}{\int \mathcal{L}_{\mathcal{R}}(s)\ ds} \tag{26}$$

We have thus determined α_n, the value of the "unknown concomitant" at the true value s_T using our data set c_n. This is our *measurement* of α_n and different datasets will give different values of α_n, in other words α_n will have a sampling distribution with an expected value and standard deviation. As $n \to \infty$, the likelihood ratio $\mathcal{L}_{\mathcal{R}}$ will tend to a finite value at the true value and zero for all other values, and $\alpha_n \to \infty$ as a result.

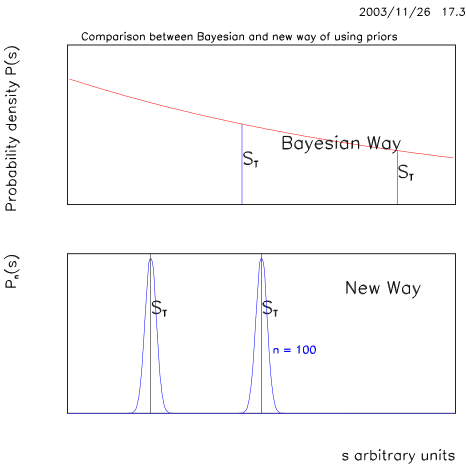


Figure 6: Comparison of the usage of Bayesian priors with the new method. In the upper figure, illustrating the Bayesian method, an unknown distribution is guessed at by the user based on "degrees of belief" and the value of the Bayesian prior changes as the variable s changes. In the lower figure, an "unknown concomitant" distribution is used whose shape depends on the statistics. In the case of no bias, this distribution peaks at the true value of s. As we change s, we change our hypothesis as to where the true value of s lies, and the distribution shifts with s as explained in the text. The value of the distribution at the true value is thus independent of s.

Note that it is only possible to write down an expression for α_n dimensionally when a likelihood ratio $\mathcal{L}_{\mathcal{R}}$ is available. This leads to

$$P(s|\mathbf{c_n}) = \frac{\mathcal{L}_{\mathcal{R}}}{\int \mathcal{L}_{\mathcal{R}}\ ds} = \frac{P(\mathbf{c_n}|s)}{\int P(\mathbf{c_n}|s)ds} \tag{27}$$

The last equality in equation 27 is the same expression that "frequentists" use for calculating their errors after fitting, namely the likelihood curve normalized to unity gives the parameter errors. If the likelihood curve is Gaussian shaped, then this justifies a change of negative log-likelihood of $\frac{1}{2}$ from the optimum point to get the 1σ errors. Even if it is not Gaussian, as we show in section (8), we may use the expression for $P(s|\mathbf{c_n})$ as a *pdf* of the parameter s to evaluate the errors.

The normalization condition

$$P(\mathbf{c_n}) = \int P^{theory}(s,\mathbf{c_n})ds = \int P(c_n|s)\mathcal{P}_n(s_T)ds \tag{28}$$

is obeyed by our solution, since

$$\int P(\mathbf{c_n}|s)\mathcal{P}_n(s_T)\ ds = \int \alpha_n P(\mathbf{c_n}|s)\ ds \equiv P^{data}(\mathbf{c_n}) \tag{29}$$

The expression $\int \alpha_n P(\mathbf{c_n}|s)\ ds$ in the above equation may be thought of as being due to an "unknown concomitant" whose peak value is distributed uniformly in s space. The likelihoods of the theoretical prediction $P(\mathbf{c_n}|s)$ contribute with equal probability each with a weight α_n, to sum up to form the data likelihood $P^{data}(\mathbf{c_n})$. i.e. the data, due to its statistical inaccuracy will entertain a range of theoretical parameters. However, equation 29 does not give us any further information, since it is obeyed identically. Fitting for the maximum likelihood value s^* of s is attained by maximizing the likelihood ratio $\mathcal{L_R} = \frac{P(\mathbf{c_n}|s)}{P^{data}(\mathbf{c_n})}$. The goodness of fit is obtained using the value of $\mathcal{L_R}$ at the maximum likelihood point. The best theoretical prediction is $P(c|s^*)$, and this prediction is used to compare to the data *pdf* $P^{data}(c)$. Note that the maximum likelihood value s is also the same point at which the posterior density $P(s|c)$ peaks. This is true only in our method. When an arbitrary Bayesian prior is used, the maximum likelihood value is not the same point at which the posterior density will peak. Note also that the normalization equation $\int \mathcal{P}_n(s)\ ds$=1 is still valid. The integral

$$\int \alpha_n\ ds \neq 1 \tag{30}$$

since α_n is our measurement of the value of $\mathcal{P}_n(s)$ at the true value. It is a measure of the statistical accuracy of the experiment. The larger the value of α_n, the narrower the distribution $\mathcal{P}_n(s)$ and the more accurate the experiment.

7. COMBINING RESULTS OF EXPERIMENTS

Each experiment should publish a likelihood curve for its fit as well as a number for the data likelihood $P^{data}(\mathbf{c_n})$. Combining the results of two experiments with m and n experiments each, involves multiplying the likelihood ratios.

$$\mathcal{L_R}_{m+n}(s) = \mathcal{L_R}_m(s)\times\mathcal{L_R}_n(s) = \frac{P(\mathbf{c_m}|s)}{P^{data}(\mathbf{c_m})}\times\frac{P(\mathbf{c_n}|s)}{P^{data}(\mathbf{c_n})} \tag{31}$$

Posterior densities and goodness of fit can be deduced from the combined likelihood ratio.

8. INTERPRETING THE RESULTS OF ONE EXPERIMENT

After performing a single experiment with n events, we now can calculate $P(s|\mathbf{c_n})$, using equation 27. Equation 18 gives the prescription for arriving at $\mathcal{P}_n(s)$, given an ensemble of such experiments, the contribution from each experiment being weighted by the "data likelihood" $P^{data}(\mathbf{c_n})$ for that experiment. The "data likelihoods" integrate to unity, i.e $\int P^{data}(\mathbf{c_n})d\mathbf{c_n} = 1$. In the case of only a single experiment, with the observed $\mathbf{c_n}$ being denoted by $\mathbf{c_n^{obs}}$,

$$P^{data}(\mathbf{c_n}) = \delta(\mathbf{c_n} - \mathbf{c_n}^{obs}) \tag{32}$$

Equation 18, for a single experiment, then reduces to

$$\mathcal{P}_n(s) = \int P(s|\mathbf{c_n})P^{data}(\mathbf{c_n})d\mathbf{c_n} = P(s|\mathbf{c_n}^{obs}) \tag{33}$$

i.e. given a single experiment, the best estimator for $\mathcal{P}_n(s)$, the *pdf* of s, is $P(s|\mathbf{c_n}^{obs})$ and thus the best estimator for the true value s_T is s^{*obs} deduced from the experiment. We can thus use $P(s|\mathbf{c_n}^{obs})$ as though it is the *pdf* of s and deduce limits and errors from it. The proviso is of course that these limits and errors as well as s^{*obs} come from a single experiment of finite statistics and as such are subject to statistical fluctuations.

9. COMPARISON WITH THE BAYESIAN APPROACH

In the Bayesian approach, an unknown Bayesian prior $P(s)$ is assumed for the distribution of the parameter s in the absence of any data. The shape of the prior is guessed at, based on subjective criteria or using other objective pieces of information. However, such a shape is not invariant under transformation of variables. For example, if we assume that the prior $P(s)$ is flat in s, then if we analyze the problem in s^2, we cannot assume it is flat in s^2. This feature of the Bayesian approach has caused controversy. Also, the notion of a *pdf* of the data does not exist and $P(c)$ is taken to be a normalization constant. As such, no goodness of fit criteria exist. In the method outlined here, we have used Bayes' theorem to calculate posterior densities of the fitted parameters while being able to compute the goodness of fit. The formalism developed here shows that what is conventionally thought of as a Bayesian prior distribution is in fact a normalization constant and what Bayesians think of as a normalization constant is in fact the *pdf* of the data. Table II outlines the major differences between the Bayesian approach and the new one.

Table II The key points of difference between the Bayesian method and the new method.

Item	Bayesian Method	New Method
Goodness of fit	Absent	Now available in both binned and unbinned fits
Data	Used in evaluating theory *pdf* at data points	Used in evaluating theory *pdf* at data points as well as evaluating data *pdf* at data points
Prior	Is a distribution that is guessed based on "degrees of belief" Independent of data, monolithic	No prior needed. One calculates a constant from data $\alpha_n = \frac{P^{data}(\mathbf{c_n})}{\int P(\mathbf{c_n}\|s)ds}$ $\rightarrow \infty$ as $n \rightarrow \infty$
Posterior density $P(s\|\mathbf{c_n})$	Depends on Prior. $\frac{P(\mathbf{c_n}\|s)P(s)}{\int P(\mathbf{c_n}\|s)P(s)\ ds}$	Independent of prior. same as frequentists use $\frac{P(\mathbf{c_n}\|s)}{\int P(\mathbf{c_n}\|s)\ ds}$

10. FURTHER WORK TO BE DONE

Equation 18 can be used to show that the expectation value of $E(s)$ of the parameter s is given by

$$E(s) = \int s\mathcal{P}_n(s)ds = \int d\mathbf{c_n}P(\mathbf{c_n})\int sP(s|\mathbf{c_n})ds \quad (34)$$

$$= \int \bar{s}(\mathbf{c_n})P(\mathbf{c_n})d\mathbf{c_n} \quad (35)$$

where $\bar{s}(\mathbf{c_n})$ is the average of s for individual experiments. Equation 35 states $E(s)$ is the weighted average of $\bar{s}(\mathbf{c_n})$ obtained from individual measurements, the weight for each experiment being the "data likelihood" $P(\mathbf{c_n})$ for that experiment. In the absence of experimental bias, $E(s)$ would be identical to the true value s_T. It remains to be shown that the weighted average of maximum likelihood values s^* from individual experiments also converge to the maximum likelihood point of $\mathcal{P}_n(s)$.

Also one needs to develop an analytic theory of the goodness of fit for unbinned likelihood fits. Finally, one needs to investigate a bit more closely the transformation properties of $\mathcal{P}_n(s)$ under change of variable.

11. CONCLUSIONS

To conclude, we have proposed a scheme for obtaining the goodness of fit in unbinned likelihood fits. This scheme involves the usage of two *pdf*'s, namely data and theory. In the process of computing the fitted errors, we have demonstrated that the quantity in the joint probability equations that has been interpreted as the "Bayesian prior" is in reality a number and not a distribution. This number is the value of the *pdf* of the parameter, which we call the "unknown concomitant" at the true value of the parameter. This number is calculated from a combination of data and theory and is seen to be an irrelevant parameter. If this viewpoint is accepted, the controversial practice of guessing distributions for the "Bayesian Prior" can now be abandoned, as can be the terms "Bayesian" and "frequentist". We show how to use the posterior density to rigorously calculate fitted errors.

Acknowledgments

This work is supported by Department of Energy. The author wishes to thank Jim Linnemann and Igor Volobouev for useful comments.

Editor's Note: The reviewer of this article found the discussion in Sections 6-11 unconvincing, and considered that some statements about frequentism were inconsistent with accepted Frequentist statistical practice.

Author's Response: Sections 1-6 of the paper motivate the need for two pdf's for obtaining the goodness of fit. Sections 6-11 deal with the mathematical consequences of the two pdf's in calculating posterior densities of fitted parameters. The conclusions of the paper are based on mathematical equations (Equation 18 is a key equation). To rebut the conclusions of the paper, it is necessary to point to errors in the mathematics.

References

[1] K. Kinoshita, "Evaluating Quality of Fit in Unbinned Maximum Likelihood fitting", Proceedings of the Conference on Advanced Statistical Techniques in Particle Physics, Durham, March 2002 IPPP/02/39, DCPT/02/78.
B. Yabsley,"Statistical Practice at the BELLE Experiment, and some questions", *ibid.*
R. D. Cousins,"Conference Summary", *ibid.*

[2] R. A. Fisher,"On the mathematical foundations of theoretical statistics", *Philos. Trans. R. Soc. London Ser. A* **222**, 309-368(1922);
R. A. Fisher,"Theory of statistical estimation", *Proc. Cambridge Philos. Soc.* **22**, 700-725 (1925).

[3] "A measure of the goodness of fit in unbinned likelihood fits", R. Raja, long write-up, `http://www.slac.stanford.edu/econf/C030908/`

[4] "End of Bayesianism?", R. Raja, `http://www.slac.stanford.edu/econf/C030908/`

[5] E. Parzen, "On estimation of a probability density function and mode" *Ann.Math.Statis.* **32**, 1065-1072 (1962).

[6] D. Scott. *Multivariate Density Estimation.* John Wiley & Sons, 1992.
M. Wand and M. Jones, *Kernel Smoothing.* Chapman & Hall, 1995.

Goodness of Fit: What Do We Really Want to Know?

I. Narsky
California Institute of Technology, Pasadena, CA 91125, USA

Definitions of the goodness-of-fit problem are discussed. A new method for estimation of the goodness of fit using distance to nearest neighbor is described. Performance of several goodness-of-fit methods is studied for time-dependent CP asymmetry measurements of $\sin(2\beta)$.

1. INTRODUCTION

The goodness-of-fit problem has recently attracted attention from the particle physics community. In modern particle experiments, one often performs an unbinned likelihood fit to data. The experimenter then needs to estimate how accurately the fit function approximates the observed distribution. A number of methods have been used to solve this problem in the past [1], and a number of methods have been recently proposed [2, 3] in the physics literature.

For binned data, one typically applies a χ^2 statistic to estimate the fit quality. Without discussing advantages and flaws of this approach, I would like to stress that the application of the χ^2 statistic is limited. The χ^2 test is neither capable nor expected to detect fit inefficiencies for all possible problems. This is a powerful and versatile tool but it should not be considered as the ultimate solution to every goodness-of-fit problem.

There is no such popular method, an equivalent of the χ^2 test, for unbinned data. The maximum likelihood value (MLV) test has been frequently used in practice but it often fails to provide a reasonable answer to the question at hand: how well are the data modelled by a certain density [4]? It is only natural that goodness-of-fit tests for small data samples are harder to design and less versatile than those for large samples. For small data samples, asymptotic approximations do not hold and the performance of every goodness-of-fit test needs to be studied carefully on realistic examples. Thus, the hope for a versatile unbinned goodness-of-fit procedure expressed by some people at the conference seems somewhat naive.

A more important practical question is how to design a powerful goodness-of-fit test for each individual problem. It is not possible to answer this question unless we specify in more narrow terms the problem that we are trying to solve.

2. WHAT IS A GOODNESS-OF-FIT TEST?

A hypothesis test requires formulation of null and alternative hypotheses. The confidence level, $1-\alpha_I$, of the test is then defined as the probability of accepting the null hypothesis given it is true, and the power of the test, $1-\alpha_{II}$, is defined as the probability of rejecting the null hypothesis given the alternative is true. Above, α_I and α_{II} denote Type I and Type II errors, respectively. An ideal hypothesis test is uniformly most powerful (UMP) because it gives the highest power among all possible tests at the fixed confidence level. In most realistic problems, it is not possible to find a UMP test and one has to consider various tests with acceptable power functions.

There is a long-standing controversy about the connection between hypothesis testing and the goodness-of-fit problem. It can be argued [5] that there can be no alternative hypothesis for the goodness-of-fit test. In this approach, however, the experimenter does not have any criteria for choosing one goodness-of-fit procedure over another. One can design a goodness-of-fit test using first principles, advanced computational methods, rich intuition or black magic. But the practitioner wants to know how well this method will perform in specific situations. To evaluate this performance, one needs to study the power of the proposed method against a few specific alternatives. A certain, perhaps vague, notion of an alternative hypothesis must be adopted for this exercise; hence, a certain, perhaps vague, notion of the alternative hypothesis is typically used to design a goodness-of-fit test.

Consider, for example, testing uniformity on an interval. The alternative is usually perceived as presence of peaks in the data. Suppose we design a procedure that gives the highest goodness-of-fit value for equidistant experimental points. This test will perform well for the chosen alternative. In reality, however, we may need to test exponentiality of the process. For instance, we use a Geiger counter to measure elapsed time between two consecutive events and plot these time intervals next to each other on a straight line. In this case, equidistant data would imply that the process is not as random as we thought, and the designed goodness-of-fit procedure would fail to detect the inconsistency between the data and the model. Tests against highly structured data (e.g., equidistant one-dimensional data) have been, in fact, a subject of statistical research on goodness-of-fit methods.

The question therefore is how to state the alternative hypothesis in a way appropriate for each individual problem. I emphasize that I am not suggesting to use a directional test for one specific well-defined alternative. The goal is to design an omnibus goodness-of-fit test that discriminates against at least several plausible alternatives.

The null hypothesis is defined as

$$H_0 : \quad X \sim f(x|\theta_0, \eta) \; , \tag{1}$$

where X is a multivariate random variable, and f is the fit density with a vector of arguments x, vector of parameters θ and vector of nuisance parameters η. The alternative hypothesis is stated in the most general way as

$$H_1 : \quad X \sim g(x) \quad with \quad g(x) \neq f(x|\theta_0, \eta) \; . \tag{2}$$

A specific subclass of this alternative hypothesis that is sometimes of interest is expressed as

$$H_1 : \quad X \sim f(x|\theta, \eta) \quad with \quad \theta \neq \theta_0 \; . \tag{3}$$

In other words, most usually we would like to test the fit function against different shapes (2). For the test (3), we assume that the shape of the fit function is correctly modelled and we only need to cross-check the value of the parameter.

If a statistic $S(x)$ is used to judge the fit quality, the goodness-of-fit is given by

$$1 - \alpha_I = \int_{f_S(s) > f_S(s_0)} f_S(s) ds \; , \tag{4}$$

where s_0 is the value of the statistic observed in the experiment, and $f_S(s)$ is the distribution of the statistic under the null hypothesis.

In practice, the vector of parameter estimates θ_0 is usually extracted from an unbinned maximum likelihood (ML) fit to the data: $\theta_0 = \hat{\theta}(x)$. In this case, the goodness-of-fit statistic must be independent of, or at most weakly correlated to, the ML estimator of the parameter: $\rho(S(x), \hat{\theta}(x)) \approx 0$, where ρ is the correlation coefficient computed under the null hypothesis. If $S(x)$ and $\hat{\theta}(x)$ are strongly correlated, the goodness-of-fit test is redundant.

The ML estimator itself is usually a powerful tool for discrimination against the alternative (3). In this case, the statistic $S(x) \neq \hat{\theta}(x)$ can be treated as an independent cross-check of the parameters θ_0.

The nuisance parameters η should affect our judgment about the fit quality as little possible. Discussion of methods for handling nuisance parameters is beyond the scope of this note.

3. DISTANCE TO NEAREST NEIGHBOR TEST

The idea of using Euclidian distance between nearest observed experimental points as a goodness-of-fit measure is not new. Clark and Evans [6] used an average distance between nearest neighbors to test two-dimensional populations of various plants for uniformity. Later they extended this formalism to a higher number of dimensions. Diggle [7] proposed to use an entire distribution of ordered distances to nearest neighbors and apply Kolmogorov-Smirnov or Cramer-von Mises tests to evaluate consistency between experimentally observed and expected densities. Ripley [8] introduced a function, $K(t)$, which represents a number of points within distance t of an arbitrary point of the process; he used the maximal deviation between expected and observed $K(t)$ as a goodness-of-fit measure. Bickel and Breiman [9] introduced a goodness-of-fit test based on the distribution of the variable $\exp(-Nf(x_i)V(x_i))$, where $f(x_i)$ is the expected density at the observed point x_i, $V(x_i)$ is the volume of a nearest neighbor sphere centered at x_i and N is the total number of observed points. An approach closely related to distance-to-nearest-neighbor tests is two-sample comparison based on counts of nearest neighbors that belong to the same sample [10]. These methods received substantial attention from the science community and have been applied to numerous practical problems, mostly in ecology and medical research. Ref. [11] offers a survey of distance-to-nearest-neighbor methods.

The goodness-of-fit test [3] uses a bivariate distribution of the minimal and maximal distances to nearest neighbors. First, one transforms the fit function defined in an n-dimensional space of observables to a uniform density in an n-dimensional unit cube. Then one finds smallest and largest clusters of nearest neighbors whose linear size maximally deviates from the average cluster size predicted from uniformity. The cluster size is defined as an average distance from the central point of the cluster to m nearest neighbors. If the experimenter has no prior knowledge of the optimal number m of nearest neighbors included in the goodness-of-fit estimation, one can try all possible clusters $2 \leq m \leq N$, where N is the total number of observed experimental points. The probability of observing the smallest and largest clusters of this size gives an estimate of the goodness of fit and the locations of the clusters can be used to point out potential problems with data modelling.

This method is a good choice for detection of well-localized irregularities, e.g., unusual peaks in the data. Consider, for example, fitting a normal peak on top of the smooth background, as shown in Fig. 3. The likelihood function is sensitive to the mean and width of the normal component. Hence, if the experimenter is mostly interested in how accurately these parameters

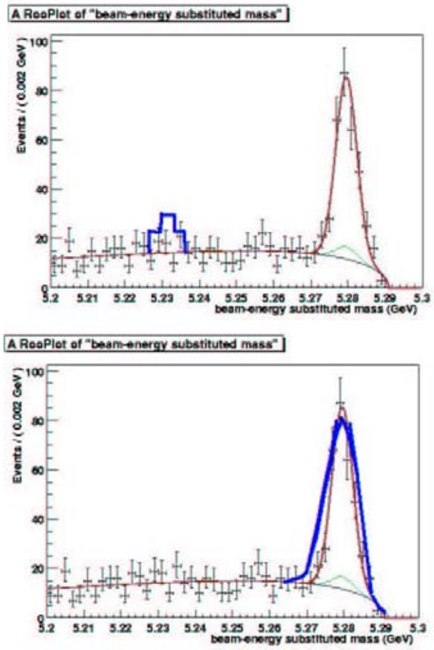


Figure 1: Fits to the sum of a normal signal density and smooth background. Possible deviations of data from the fit function are shown with a thick line. The data exhibit an unusual peak in the background tail (top), and the normal peak in the data is wider than the normal component in the fit function (bottom).

are modelled, the likelihood function can be used to address this question. At the same time, the likelihood shows little sensitivity to the bump in the smooth background tail of the distribution. If the experimenter wants to be aware of such irregularities, the method [3] is a good choice.

The transformation to uniformity is a common technique used in goodness-of-fit methods. One should be keep in mind, however, that a multivariate transformation to uniformity is not necessarily unique. If a different uniformity transformation is chosen, one can obtain a different goodness-of-fit value. One solution to this ambiguity is discussed in Ref. [3].

Table I Statistics that can be applied to an unbinned ML fit in a $\sin(2\beta)$ measurement. For convenience, $\sin(2\beta)$ is replaced in all formulas with θ. Here, the random variable u is obtained by uniformity transformation $u = \int_{-t_{max}}^{t} f(t|\theta_0)dt$, b_j is a Legendre polynomial of order j, F_n is the experimental cumulative density function (CDF), F is the CDF under the null hypothesis, the upper bar denotes averaging, and n is the total number of events observed. The lifetime τ was allowed to vary in the fit.

Method	Formula
likelihood ratio	$L(\theta_0\|x)/L(\hat{\theta}\|x)$
ML estimator	$\hat{\theta}$
score statistic	$(\partial L(\theta\|x)/\partial\theta)\|_{\theta=\theta_0}$
MLV	$L(\theta_0\|x)$
Neyman with $K = 1, 2, 3$	$\sum_{j=1}^{K}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} b_j(u_i)\right)^2$
Kolmogorov-Smirnov	$\max_{i=1,2,\dots,n}\|F_n(u_i) - F(u_i)\|$
Watson U^2	$\int_0^1 (F_n(u) - u)^2 du - \left(\bar{u} - \frac{1}{2}\right)^2$
Anderson-Darling	$\int_0^1 \frac{(F_n(u)-u)^2}{u(1-u)} du$
min distance to nearest neighbor	$d_{\min}(t < 0)/d_{\min}(t > 0)$
max distance to nearest neighbor	$d_{\max}(t < 0)/d_{\max}(t > 0)$

4. COMPARISON OF GOODNESS-OF-FIT METHODS FOR $\sin(2\beta)$ MEASUREMENTS

I apply several methods summarized in Table I to unbinned ML fits in time-dependent CP asymmetry measurements and compare their power functions at fixed confidence levels. The fit function in $\sin(2\beta)$ measurements is given by [12]

$$f(t|\sin(2\beta)) = \frac{1}{2\tau}\exp\left(\frac{-|t|}{\tau}\right)\left[1+\sin(2\beta)\sin(\Delta m t)\right], \tag{5}$$

where positive and negative times t correspond to B tags of opposite flavors in the range -8 ps $\leq t \leq 8$ ps, $\sin(2\beta)$ is a measure of the asymmetry, $\tau = 1.542 \pm 0.016$ ps is the lifetime of the B meson [13], and $\Delta m = 0.489 \pm 0.009$ ps^{-1} [13] is the $B\bar{B}$ mass mixing. For toy Monte Carlo experiments, I generate samples with 100 events, a typical sample size in a BaBar analysis of $B \to J/\psi K_S$ decays in which the J/ψ is reconstructed in hadronic final states [12]. I ignore smearing due to detector resolution and background contributions to the data.

I test the null hypothesis H_0 : $\sin(2\beta) = 0.78$ against H_1 : $\sin(2\beta) \neq 0.78$ and plot correlations $\rho(S(x), \hat{\theta}(x))$ for each listed statistic under the null hypothesis, as well as power functions estimated from Monte Carlo samples generated with different values of $\sin(2\beta)$: 0, 0.5, 0.7, and 1. The power functions for $\sin(2\beta) = 0.5$ and the correlations are shown in Fig. 4. While the top three methods in Table I provide good separation between values of $\sin(2\beta)$, they cannot be used as independent cross-checks of the parameter because of the strong correlation with the ML estimator. The MLV method shows a relatively strong correlation with the ML estimator and a relatively poor power function. The three Neyman smooth tests, as well as Kolmogorov-Smirnov, Watson and Anderson-Darling

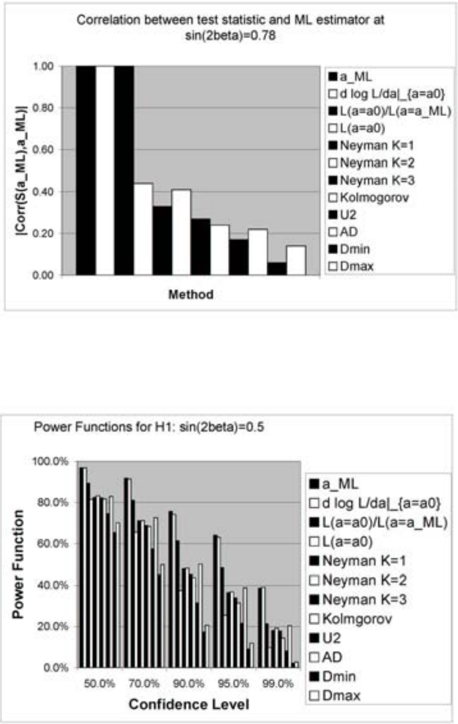


Figure 2: Correlations between the ML estimator of $\sin(2\beta)$ and the chosen statistic (top). Power functions of the hypothesis test $H_0: \ \sin(2\beta) = 0.78$ against $H_1: \ \sin(2\beta) \neq 0.78$ versus confidence level (bottom) at $\sin(2\beta) = 0.5$.

tests, show small correlation to the ML estimator and decent power functions; these tests perform competitively among each other. Yet Kolmogorov-Smirnov and Anderson-Darling tests produce somewhat better combinations of the small correlation and large power function and should be preferred over others. The distance-to-nearest-neighbor test was designed for detection of well-localized regularities and hence was not expected to give a high power function for the hypothesis test discussed here.

This exercise alone is insufficient to conclude that the two recommended tests are in fact the best omnibus tests for fits of $\sin(2\beta)$. One would have to extend this study to include other alternative densities, e.g., specific background shapes, that can distort the experimental data.

5. SUMMARY

An acceptable goodness-of-fit test is defined as an omnibus test that discriminates against at least several plausible alternatives. Numerous distance-to-nearest-neighbor methods for goodness-of-fit estimation have been described in the statistics literature and should be tested in HEP practice. The distance-to-nearest-neighbor test based on minimal and maximal distances should be used for detection of well-localized irregularities in the data. Correlation coefficients and power functions for several statistics are compared for fits of $\sin(2\beta)$ in CP asymmetry measurements for one specific alternative.

Acknowledgments

I wish to thank the organizing committee of PHYSTAT2003 for their effort. Thanks to Bob Cousins for useful comments on this note.

Work partially supported by Department of Energy under Grant DE-FG03-92-ER40701.

References

[1] R. D'Agostino and M. Stephens, "Goodness-of-Fit Techniques", Marcel Decker, Inc., 1986; J. Rayner and D. Best, "Smooth Tests of Goodness of Fit", Oxford Univ. Press, 1989.

[2] B. Aslan and G. Zech, "A new class of binning free, multivariate goodness-of-fit tests: The energy tests", hep-ex/0203010, 2002.

[3] I. Narsky, "Estimation of Goodness-of-Fit in Multidimensional Analysis Using Distance to Nearest Neighbor", physics/0306171, 2003.

[4] J. Heinrich, "Can the likelihood function be used to measure goodness of fit?", CDF/MEMO/BOTTOM/CDFR/5639, Fermilab; also in these Proceedings.

[5] See, for example, F. James' talk at `http://www-conf.slac.stanford.edu/phystat2003/talks/james/james-slac.slides.ps`

[6] P.J. Clark and F.C. Evans, "Distance to Nearest Neighbor as a Measure of Spatial Relationships in Populations", Ecology **35-4**, 445 (1954); "Generalization of a Nearest Neighbor Measure of Dispersion for Use in K Dimensions", Ecology **60-2**, 316 (1979).

[7] P. Diggle, "On Parameter Estimation and Goodness-of-Fit Testing for Spatial Point Patterns", Biometrics **35**, 87 (1979).

[8] B.D. Ripley, "Modelling Spatial Patterns", J. of the Royal Stat. Soc. **B 39-2**, 172 (1977).

[9] P.J. Bickel and L. Breiman, "Sums of functions of nearest neighbor distances, moment bounds, limit theorems and a goodness of fit test", Ann. of Probability **11**, 185 (1983).

[10] J.H. Friedman and L.C. Rafsky, "Multivariate Generalizations of the Wald-Wolfowitz and

Smirnov Two-Sample Tests", Ann. of Statistics **7**, 697 (1979); M.F. Schilling, "Multivariate Two-Sample Tests Based on Nearest Neighbors", J. of the Amer. Stat. Assoc. **81**, 799 (1986); J. Cuzick and R. Edwards, "Spatial Clustering in Inhomogeneous Populations", J. of the Royal Stat. Soc. **B 52-1**, 73 (1990).

[11] P.M. Dixon, "Nearest Neighbor Methods", `http://www.stat.iastate.edu/preprint/articles/2001-19.pdf`

[12] See, for example, BaBar Collaboration, "Measurement of sin2beta using Hadronic J/psi Decays", hep-ex/0309039, 2003.

[13] Phys. Rev. **D 66**, Review of Particle Physics, 2002.

Fundamental Issues in Statistical Detection of Physical Phenomena

P. S. Shawhan
LIGO Laboratory, California Institute of Technology, Pasadena, CA 91125, USA

Many physics experiments are designed to search for some rare, previously unseen phenomenon which would leave a distinctive event signature in the detector. Generally, there are one or more "background" processes which can mimic the signature, so that detecting the new phenomenon is a matter of observing significantly more events than would be expected from background processes. However, the Bayesian approach to credible interval construction does not, in itself, address the question of whether a given excess should be interpreted as a "detection" of the phenomenon. Unified frequentist (*e.g.* Feldman-Cousins) approaches to confidence interval construction dictate when the interval should exclude zero, but are rarely (if ever) calculated using a high confidence level that would be appropriate for making a detection. The standard quantitative way to judge the significance of an apparent signal (in excess of the expected background) is to calculate the *p*-value for the null hypothesis.

1. PREFACE

The goal of this article is to review some important fundamental issues—philosophical, not technical—which arise when interpreting the results of a search for a physical phenomenon which has not yet been observed. We will discuss these issues, starting at the most basic level, in the context of very familiar, straightforward statistical analysis approaches, and will point out that the question of what constitutes a "detection" is not directly addressed by these approaches.

2. "DETECTION" AND UPPER LIMITS IN PHYSICS EXPERIMENTS

A physics experiment typically produces a large amount of data, which needs to be distilled down to something meaningful. In other words, the outcome of an experiment calls for an interpretation. The usual approach is to discard most of the details in the raw data and construct a simple "statistic" to summarize the data, usually a scalar quantity, such as the number of events satisfying a set of selection criteria. For a given choice of statistic, X, the full information content of the experimental result consists of the observed value of the statistic, x, along with the probability function $P(X|\theta)$ which describes the probability of observing any given value of the statistic as a function of one or more imperfectly known physical parameters, denoted by θ. This information content is completely objective, assuming that the decision about what statistic to construct was made without reference to the experimental data.[1]

The final step in an analysis is to extract a physical interpretation from $P(X|\theta)$ and x, normally to infer favored ranges for the values of the physical parameter(s) θ. This may be done in a frequentist sense, by defining an "acceptance region" (of "likely" experimental outcomes) in $P(X|\theta)$ and using the Neyman construction [1] to calculate the resulting confidence interval given the observed value x. Or it may be done in a Bayesian sense, by folding together one's prior belief about θ with the likelihood function $L(\theta) \equiv P(x|\theta)$ to arrive at a posterior probability density function (pdf), and perhaps then go on to derive a credible interval from that pdf. Either approach to interpretation involves a choice about how to define the interval, as we shall discuss further in the later sections of this article.

Many physics experiments are designed to try to detect a distinct signature in the detector from "new physics", some hypothesized physical phenomenon which has not previously been observed. Current examples include the Higgs boson and gravitational waves; past examples have included the top quark, *CP* violation in *B* mesons, etc. In some cases, there are good theoretical reasons, or indirect information from other experiments, to believe that the effect exists, and there may even be an estimate of its magnitude or event rate.[2] In other cases, the magnitude of the effect is unknown, and may even be unmeasurably small or nonexistent. Generally, there are one or more "background" processes which can mimic the signature in the detector, so that detecting the new phenomenon is a matter of observing significantly more events than would be expected from background processes. Of course, even if the *average* magnitude of the background is known accurately, the statistical analysis must allow for fluctuations.

A common aspect of searches for "new physics" is that physicists generally take a conservative approach (in a sociological sense, not the statistical sense) to claiming a "detection". In other words, they require a high standard of evidence. This is sometimes expressed in terms of an equivalent number of standard deviations for

[1] Note that the probability function contains everything which is known about the random aspects of the experiment, regardless of whether a frequentist or a Bayesian approach is to be used to interpret it. Thus, despite occasional claims to the contrary, frequentist and Bayesian analyses are equally dependent on the concept of randomness (sometimes discussed in conceptual terms as an ensemble of identical experiments).

[2] It might seem that the best determination of the event rate or other physical parameters would come from a Bayesian analysis using the theoretical or indirect information in the prior. However, physicists often want to *test* the theory or the consistency of the indirect information, so using that information in the analysis would lead to circular reasoning.

a Gaussian random process, *e.g.* "5 sigma", even when the distribution of the statistic is not Gaussian; the intent is to convey the false detection probability (less than 10^{-6} in this case).

More often than not, these experiments fail to observe clear evidence for the physical effect being looked for. The absence of a significant excess means that the rate or magnitude of the physical effect is unlikely to be very large; this may be expressed quantitatively as an *upper limit* on the event rate or magnitude.[3] Upper limits are typically reported with a 90% or 95% confidence level, depending on conventions established by past experiments in each field of research.

3. DETECTION ISSUES IN A BAYESIAN ANALYSIS

To illustrate some issues which are encountered in a Bayesian analysis, we consider the archetypal "Poisson process with background" case considered, for instance, by Feldman and Cousins [2]. This represents a "counting" experiment, in which the statistic used to summarize the data is the number of events, n, which satisfy a set of selection criteria designed to keep most signal events (if any exist) and reject uninteresting events. If μ is the mean number of signal events expected (an unknown physical parameter, in the range $0 \leq \mu < \infty$) and b is the mean number of background events expected (and is known accurately), then the likelihood function is the Poisson distribution with mean $\mu+b$:

$$L(\mu) = (\mu+b)^n \, e^{-(\mu+b)} \, / \, n!$$

Given some prior belief about the relative probabilities of different values of μ, we apply Bayes' theorem to get a posterior probability density function (pdf). For example, Figure 1 shows the posterior pdf if 7 events are observed, assuming b=3 and a constant prior pdf for μ.

A true Bayesian might consider this posterior pdf to be the final product of the analysis, but most physicists, I think, would want to go one step further and summarize the result with a credible interval. There is no objective rule which dictates what sort of credible interval should be constructed; three possibilities are illustrated in Figure 2. Choosing a credible interval which excludes zero is, in essence, a decision to interpret the result as an apparent detection with some degree of confidence. Is that an appropriate choice in this case? The fact that the pdf is distinctly peaked away from zero is certainly suggestive, but how robust is that as an indicator?

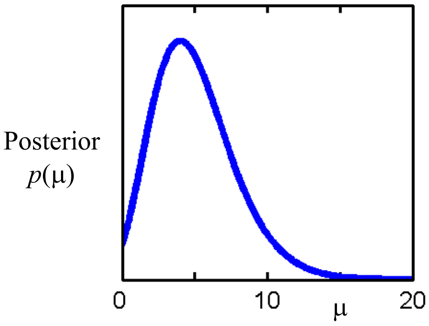


Figure 1: A posterior pdf for the example considered in the text, if n=7.

The peakedness of the posterior pdf depends, in part, on the choice of prior. Figure 3 shows the posterior pdfs for values of n between 6 and 10, for three different priors. Which ones do you think look significant enough that you would be comfortable publishing a paper claiming a detection? How often are you willing to be wrong? In the case of the constant prior, the posterior pdf is noticeably peaked away from zero even for n=6, but it turns out that the background will fluctuate up to 6 or more events 8.4% of the time,[4] so the presence of a peak is not necessarily a reliable indicator. This reflects the fact that a constant prior is too optimistic when searching for a signal which is likely to be small. In fact, if we are completely ignorant about the value of μ (which is a scale parameter in the likelihood), then the Principle of Maximum Entropy [3] suggests that we should use a prior of the form $1/\mu$. In this case, the posterior pdf develops a peak at somewhat higher values of μ, but it is improper for all values of n, so we cannot calculate credible intervals at all! In essence, this prior would lead us to conclude that *any* number of excess events is more likely to be a background fluctuation than to be a real signal.[5] The final prior considered in Figure 3, 1/sqrt(μ), represents a sort of compromise: it emphasizes small values of μ, but yields integrable posterior pdfs. Still, there is no guidance about what is significant enough to represent a detection, other than by considering the false detection probability (a frequentist concept!).

[3] In fact, some experiments / analyses, for which detection is unexpected according to theoretical predictions, are optimized so as to minimize the expectation value of the upper limit (assuming that no signal is seen).

[4] For reference, the background (3 events on average) will fluctuate up to 7 or more events 3.3% of the time, 8 or more 1.2% of the time, 9 or more 0.4% of the time, and 10 or more 0.1% of the time.

[5] One might be tempted to use a prior of the form $1/(\mu+b)$, in which case a change of variables seems to reduce the problem to the simple Poisson case without background. However, this is not quite true, because the domain of the Poisson mean parameter becomes $[b, \infty)$, not $[0, \infty)$. In any case, this prior is conceptually flawed: one's *prior* belief about a physical parameter cannot depend on the properties of the *present* experiment!

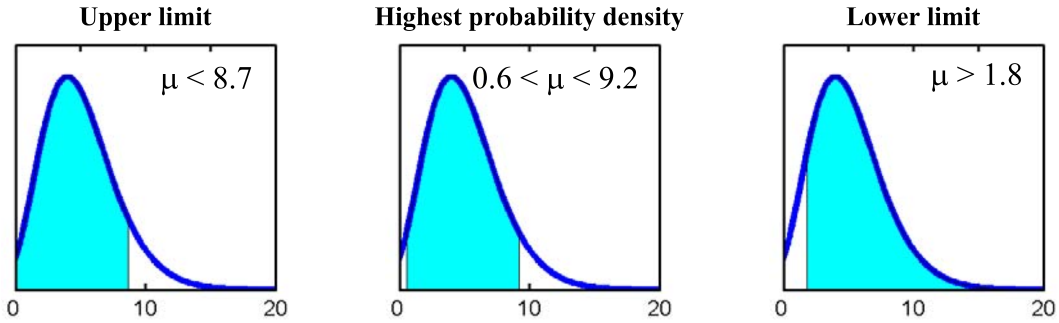


Figure 2: Three possible 90% credible intervals constructed from the posterior pdf shown in Figure 1.

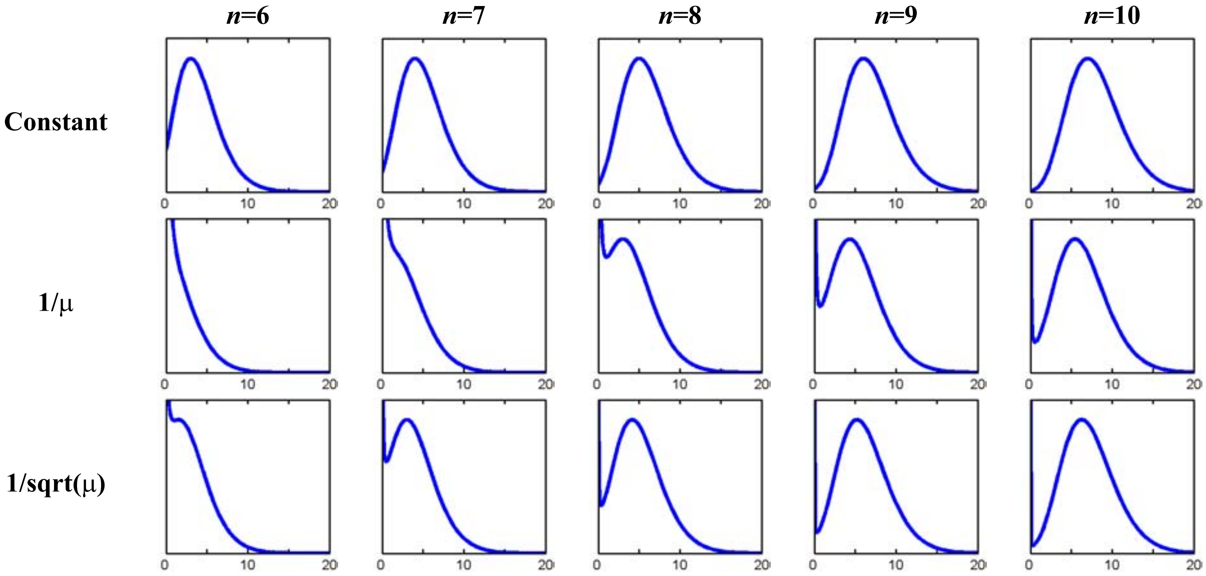


Figure 3: Posterior pdfs for the example considered in the text, for three different functional forms of the prior pdf: constant, 1/μ, and 1/sqrt(μ). Note that in the latter two cases, the posterior pdf diverges as $\mu \to 0$, for all values of *n*.

4. DETECTION ISSUES IN A UNIFIED FREQUENTIST ANALYSIS

Feldman and Cousins have popularized a "unified" frequentist approach, in which the classical Neyman construction is performed with an alternative ordering principle based on likelihood ratios [2]. This approach (which has a few variations) yields confidence intervals which transition smoothly from one-sided to symmetric two-sided as *n* increases, maintaining the desired minimum coverage. This may seem to provide a well-defined detection criterion, at the point of the transition to two-sided intervals, but there is a crucial caveat (originally pointed out by Feldman and Cousins): this type of confidence interval is almost always calculated for a 90% confidence level, so an interval which excludes zero does not necessarily represent a detection at the higher confidence level that we want to require. It is, of course, possible to calculate the interval for a higher confidence level (say, 99.9%), but then the upper end of the interval will no longer be analogous to a traditional 90% upper limit, which is generally considered to be a desirable feature. Faced with this situation, some collaborations follow a policy of giving a 90% unified confidence interval (which may be two-sided) and stating separately whether there is a "detection", based on the *p*-value for the null hypothesis. For example, Figure 4 shows the Feldman-Cousins construction for our Poisson process with background, which yields two-sided intervals for $n \geq 6$, whereas a detection with a *p*-value of 0.01 or less would require $n \geq 9$.

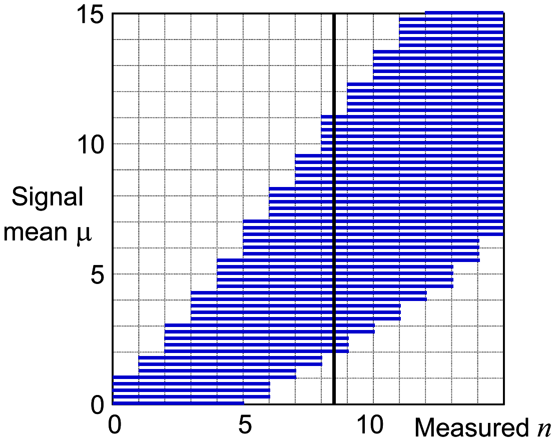


Figure 4: Feldman-Cousins confidence intervals for Poisson process with expected background *b*=3 (adapted from Figure 6 of Ref. [2]). The horizontal bands are "acceptance regions" for various values of μ. The thick vertical line indicates the mean number of observed events, *n* ≥ 9, which would be required to make a detection with a false detection probability less than 1%.

5. SUMMARY

Physicists generally expect an interpretation of the outcome of an experiment, such as a statement about the significance of any excess events observed. In the case of a search for a new phenomenon, a high standard of evidence is required to support a claim of a "detection". Even in the absence of a signal, a Bayesian analysis may occasionally yield a pdf which is peaked away from zero, and does not provide a quantitative measure of significance. A unified frequentist approach could in principle provide a well-defined detection criterion, but is not customarily calculated for an appropriately high confidence level. The best established quantitative approach to evaluate an apparent detection of a new physical phenomenon is to calculate the *p*-value for the null hypothesis. Of course, human judgment is still required to decide how low the *p*-value must be to be interpreted as a detection.

References

[1] J. Neyman, Philos. Trans. R. Soc. London **A236**, 333 (1937). Reprinted in *A selection of Early Statistical Papers on J. Neyman* (University of California Press, Berkeley, 1967), pp. 250-289. Also see the discussion in reference [2], below.

[2] G. J. Feldman and R. D. Cousins, Phys. Rev. D **57**, 3873 (1998).

[3] E. T. Jaynes, IEEE Trans. Syst. Sci. Cybernet. **SSC-4**, 227 (1968).

Sensitivity of Searches for New Signals and Its Optimization

Giovanni Punzi
Scuola Normale Superiore and INFN, Pisa, Italy

A frequentist definition of sensitivity of a search for new phenomena is discussed, that has several useful properties. It is based on completely standard concepts, is generally applicable, and has a very clear interpretation. It is particularly suitable for optimization, being independent of a-priori expectations about the presence of a signal, thus allowing the determination of a single set of cuts that is optimal both for setting limits and for making a discovery. Simple approximate formulas are given for the common problem of Poisson counts with background.

1. INTRODUCTION

The question of the sensitivity of a search for new phenomena is a very common one. The need may arise either by the wish to predict the outcome of an experiment and compare several possible experiments or different configurations of the same experiment. Several different ways have been used to quantify the sensitivity of a search, which makes it sometimes difficult to compare them. In particular, two different sensitivity figures are often quoted, one that is relative to the potential for actually making a discovery, and another to characterize how strong a constraint is imposed on the unknown phenomena if no evidence is found for a deviation from the standard theory. This situation makes it difficult to optimize the design of an experiment, because it is not clear what should be maximized. I describe here a definition of sensitivity which is unique and well-defined for any experiment. This is based on purely frequentist ideas, which avoids the issue of the choice of an a-priori distribution for a new and unknown phenomena.

2. STATEMENT OF THE PROBLEM

The problem of searches for new phenomena can be stated formally in classical statistics as one of "Hypothesis testing". We have a "default hypothesis" H_0, that is our current best theory, and as a result of the experiment we wish to either confirm or disprove the theory H_0, in favor of an alternative theory H_m, where m indicates the free parameters of the new theory (mass or set of masses of new particles, coupling constants, production cross sections, etc.). The experiment consists of measuring the value of a set of observables X (possibly a large number) whose distribution depends on the true state of nature being H_0 or H_m. In a simple counting experiment, the observable X is the number of observed counts, and hypothesis H_0 is defined as the distribution of X being a Poisson with the mean equal to the number of expected background events B. Hypothesis H_m is that the distribution is instead a Poisson with a larger mean $B + S_m$, where S_m is the expected contribution of the "new signal", which is a function of the unknown free parameters of the new theory, m. A test of H_0 is specified by defining the set of values of X that will make us decide that H_0 must be rejected ("critical region"); the *significance level* of the test, indicated by α, is the probability of rejecting H_0 when it is indeed true; that is to say, α is the probability for X to fall within the critical region, calculated under the assumption that H_0 is true. There are many possible choices of the critical region, therefore many possible different tests at the given significance level α, and we will not be concerned here with the way the choice is made; all of the present discussion is independent of the way the test was chosen.

What about the value of α ? This is a "small number", common practice for really new physics discovery being to require α to correspond to the 5σ single tail of a gaussian distribution.

The other element to be considered in a test is the probability that a discovery is made. The classical way to express this is by the *power function* $1 - \beta(m)$, that is, the probability that X will fall in the critical region (=the probability that a discovery will be claimed) assuming H_m is true, as a function of the parameters m. It is clearly desirable to have the greatest possible power. However, it is well known that only in very few special problems it is possible to maximize the power simultaneously for every m. For this reason, trying to optimize the power is subject to a judgement about what values of the parameters are more important; in the next section we will show how to solve the issue by attacking the problem from a different angle.

After a measurement is performed, if no discovery is made the experimenter will usually produce an additional piece of information: a confidence region for the unknown parameters m. This part is in principle completely independent from the "testing" part, and interesting issues arise when one tries to make sure the two kinds of information are coherent. For instance, limits are often desired at a confidence level lower than the level of significance required for claiming a discovery; this can lead easily to situations where no discovery is claimed, and yet limits are quoted that do not include the H_0 hypothesis. For the purpose of the present discussion we don't need to deal with such difficult issues and we will make only minimal assumptions about the relationship between the test and the algorithm adopted for setting limits. We will just assume that the confidence band for m be built in such a way to exclude, whenever possible, all values of X falling within the acceptance region for H_0; (this can be done for every m such that $1 - \beta(m) > CL$, where CL is the desired Confidence Level). This is quite natural, and usually happens spontaneously, because it makes for tighter confidence regions when no discovery is made, at no expense.

If a discovery is indeed made, the most interesting piece of information in the result will be the discovery itself, and maybe an estimate of the parameters m, so we will not be concerned with limit setting in case of discovery, only with the probability that it happens.

3. DEFINITION OF SENSITIVITY OF A SEARCH EXPERIMENT

Many definitions of sensitivity for a search have to do with either the "average limit" produced if H_0 is true (defined in various ways), or with the significance of an observed signal, assuming the observation is exactly equal to the expected value in presence of a signal at m.

We suggest to characterize the sensitivity of an experiment in the following way. Correct statistical practice requires to decide before the experiment the values of α and CL, so we assume their values are given. Then one can proceed by quoting the region of the parameters m for which *the power of the chosen test is greater or equal to the Confidence Level chosen for the limits in case there is no discovery*:

$$1 - \beta_\alpha(m) > CL \tag{1}$$

This region of m can be thought of as a region of parameters to which the experiment is "sufficiently sensitive". While it is always possible to provide additional information by plotting contours of constant power in the m space for values different from the CL, the specific region defined by eq. (1) is particularly informative because it has a very simple and clear-cut interpretation. In fact, it is easy to verify that the following two statements hold simultaneously:

- If the true value of m satisfies (1), then there is a probability at least CL that performing the experiment will lead to discovery (with the chosen significance α).
- If performing the experiment does not lead to discovery, the resulting limits will exclude (at least) the entire region defined by (1), at the chosen CL. (N.B. this relies on the minimal assumption of a "reasonable algorithm" for setting limits made in previous section, and holds independently of the true value of m.)

In short, eq. (1) defines the region in the parameter space for which the experiment will *certainly* give an answer: that region will be excluded, or a discovery will be claimed, with no possible in-between. This double discovery/exclusion interpretation suggests that it deserves to be named *sensitivity region* for the experiment and to be quoted as the single most useful information to characterize its potential and optimize it. Note explicitly that there is no possibility for an experimental fluctuation to jeopardize the result; it is possible for a fluctuation to increase the region of exclusion, but not to diminish it. In particular, if the parameter region covers the whole range of physically interesting values for m, the experiment can very well been said to be conclusive. This *sensitivity region* appears to be a more useful information than others commonly quoted, that have a more vague meaning, like:

- the "average" excluded region, *if* H_0 is true (tells you nothing certain about the actual limits that will be quoted; tells you nothing about what will happen if the signal exists but it is small)
- an "average number of sigmas", for given values of m, or the number of sigmas you would get in case exactly the expected number of signal events is observed (tells you nothing about the limits in case there is no observation; tells you little about how likely it is that a signal will actually be observed, due to the effect of statistical fluctuations)

Comparison between two experiments, or experimental settings, should be made on the basis of whether one sensitivity region includes the other. It is still possible for two experiments to be non-comparable, by having none of the two region completely include the other; in that case, the issue of which is preferable cannot be resolved on a statistical basis, but it is a question of strategy. If the sensitivity regions are very different, the actual conclusion is that the two experiments are somehow 'complementary', probing different regions of the parameters space.

There are a few other arguments in favor of quoting this quantity to characterize the sensitivity of an experiment:

- The definition is independent of the choice of metric (in both observable and parameter space).
- It does not require a choice of priors
- It is straightforward (and meaningful) to apply even in complex situations. For instance:

- 1-D problems with a "non monotonic" structure. Example: search for a CP violation effect, where one measures the sine of an angle, with the range $[-1, 1]$. In this case H_0 is in the middle, and it makes no sense to quote "average upper limit".
- multidimensional parameter problems. Examples of this kind are neutrino oscillation searches, where the space is 2-D. Even more complex examples are found in CP-violation measurements in neutral B mesons oscillations, where both a direct and a mixed component are possible; in this case the allowed region for the parameters is circle of unit radius, H_0 being at the center, and it is impossible to use concepts like "average upper limit", or even "median of the limit".

- It is independent of the expectations for a signal to be present, thus allowing an unbiased optimization.
- It allows you to optimize what you really want for a search, without being distracted by other elements. For instance, if one had to concentrate on getting the maximum possible power (e.g. by looking at its average it over a chosen region), one can easily be fooled into preferring an experiment that has a very high power in a region where the power is pretty high anyway, over one that has a more even distribution of power, that is actually much more likely to provide useful information, since in a discovery measurement the power counts the most where it is "intermediate". Considering the region rather than power in itself takes this into account.

4. OPTIMIZATION OF A COUNTING EXPERIMENT

We will now apply the ideas discussed in the previous section to the very common problem of a counting experiment in presence of background. In this case, we have the discrete observable n, the number of events observed, which is Poisson-distributed with a mean determined by B, the expected number of background events (supposed known), and the possible contribution of signal events S_m:

$$p(n|H_0) = e^{-B}B^n/n! \qquad (2)$$

$$p(n|H_m) = e^{-B-S_m}(B+S_m)^n/n! \qquad (3)$$

For this problem, the only sensible definition of a critical region for the presence of non-zero signal S_m takes the form of a condition like

$$n > n_{min}$$

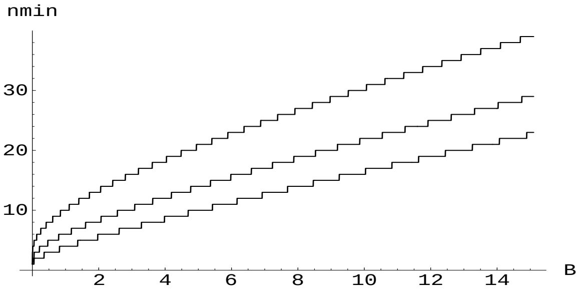


Figure 1: Minimum number of observed events needed to claim discovery with 95%, 3σ, 5σ significance, vs expected background.

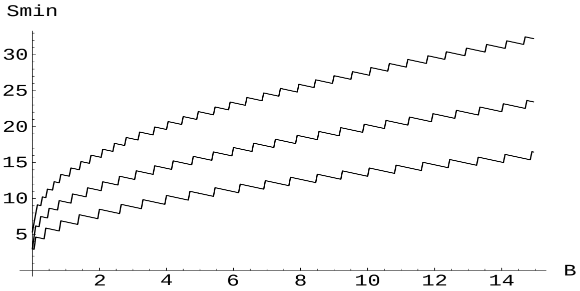


Figure 2: The lower limit of the sensitivity region S_{min}, for a search experiment with (significance, CL) respectively of (95%,95%), (3σ,95%), (5σ,90%).

Therefore, the test is completely defined once the desired significance level α is chosen. Figure 1 shows the value of n_{min} as a function of B, for given values of α, obtained by numerical calculation of sums of Poisson probabilities.

Having completely defined the test, we can now evaluate its power as a function of m, and determine the set of values for m such that eq. (1) holds. Since the power of a test of the form $n > n_{min}$ grows monotonically with S_m, it is easy to see that eq. (1) leads to simple inequalities of the form:

$$S_m > S_{min}$$

Therefore, all is needed to completely characterize the solution of our problem is the value of S_{min}, that is in general a function of α, β, and B. Plots of S_{min} from numerical calculation are shown in Fig. 2.

Tables of this kind of data can in principle be used to compare different experimental settings, by determining for each of them the set of values of m such that $S_m > S_{min}$, and choosing the one with the largest set. However, it is much easier to perform such optimizations tasks with the help of an analytic parametrization. For the purpose of optimization, an approximation of the exact result is usually sufficient; in particular, there is no need to account for the discretization effects.

A simple parametrization of our result can be obtained by means of Gaussian approximation of the Poisson. It is easy to see that in this approximation, condition (1) translates into the following equation for S_{min}:

$$S_{min} = a\sqrt{B} + b\sqrt{B + S_{min}} \tag{4}$$

where a and b are the number of sigmas corresponding to one-sided Gaussian tests at significance α and β respectively.

Solving eq. (4) for S_{min} yields the solution:

$$S_{min} = \frac{b^2}{2} + a\,\sqrt{B} + \frac{b}{2}\,\sqrt{b^2 + 4\,a\,\sqrt{B} + 4\,B} \tag{5}$$

This expression holds for one specific set of data selection criteria. Now consider the common situation where one has to decide on the set of cuts to be used in the analysis. This means that both the background B and the number of expected signal events S_m will depend on the cuts (let's indicate the whole set of cuts with the symbol t). In a completely general case, in order to decide which set of cuts t is best, one needs to determine for every t the set of values $\tilde{m}$ to which the experiment is sensitive, by solving for $\tilde{m}$ the inequality:

$$S_{\tilde{m}}(t) \geq \frac{b^2}{2} + a\,\sqrt{B(t)} + \frac{b}{2}\,\sqrt{b^2 + 4\,a\,\sqrt{B(t)} + 4\,B(t)}$$

and then choose the cuts t yielding the most extended region. The situation is much simpler when the efficiency ϵ of the chosen cuts on the signal is independent of m, that is when one can write:

$$S_m(t) = \epsilon(t) \cdot L \cdot \sigma_m$$

where L is the integrated luminosity and σ_m is the cross section of the process being searched for.

In this case one can simply invert the above equation to write down the minimum "detectable" (according to our criteria) cross section:

$$\sigma_{min} = \frac{\frac{b^2}{2} + a\,\sqrt{B(t)} + \frac{b}{2}\,\sqrt{b^2 + 4\,a\,\sqrt{B(t)} + 4\,B(t)}}{\epsilon(t) \cdot L}$$

Obviously, the maximum sensitivity is attained when σ_{min} is smallest, that is when the quantity:

$$\frac{\epsilon(t)}{b^2 + 2\,a\,\sqrt{B(t)} + b\,\sqrt{b^2 + 4\,a\,\sqrt{B(t)} + 4\,B(t)}} \tag{6}$$

reaches its maximum. Note explicitly that, in the given assumption of the efficiency being independent of m, the optimal choice of cuts *does not depend* on the assumed cross section for the new process σ_m. This is a very useful feature, since this parameter is often unknown, and it is a direct consequence of the chosen approach, that focuses on maximizing the power where it is really necessary, that is at the threshold of visibility. Expression (6) becomes even simpler when the choice $b = a$ is made:

$$\frac{\epsilon(t)}{a/2 + \sqrt{B(t)}} \tag{7}$$

This simple expression is adequate in most problems of search optimization; also, it is readily compared with some ' 'significance-like" expressions that are commonly used for optimization purposes:

a) $\frac{S}{\sqrt{B}}$

b) $\frac{S}{\sqrt{B+S}}$

Note that expression b) cannot be maximized without knowing explicitly the cross section for the searched signal. Also, it does not quite represent what one wants to maximize for a search, being more directly related to the relative uncertainty in the measurement of the yield of a new process, if found, than to significance. Expression a), being linear in S, shares with expression (7) the good property of being independent of the cross section of the new process, but it has the important problem of breaking down at small values of B. Imposing maximization of a) may push the experiment efficiency down to very small values. In order to see the failure of expression a), it is sufficient to consider, for instance, that it will prefer an expectation of 0.1 signal events with a background of 10^{-5} over a situation with 10 signal events expected and a background of 1 event.

It should be apparent that expression (7) (or its slightly more sophisticated form (6)), compared with "significance" a) and b), is not only better motivated, but also unambiguously preferable from a practical viewpoint.

The features of the discussed formulas are more easily seen by plotting the factor $1/S_{min}$ from the exact calculation (that is proportional to the quantity that needs to be maximized, as in eq. (6)) together with the two significance–like expressions discussed above: they all behave as $1/\sqrt{B}$ at large B, and it is therefore possible to normalize them to converge as $B \to \infty$. Expression b) is not simply proportional to S, so we had to make a choice and we put $\frac{S}{\sqrt{B+S}} = a$, in agreement with the spirit of our current approach of focusing on the point where significance is at the threshold, and solved for $1/S$ to obtain a function of B only.

The comparison is shown in Fig. 3, where it appears that our suggested solution lies between a) and b), where a) largely overestimates the "sensitivity" at low backgrounds, as expected, and conversely b) underestimates it, especially for high significance settings.

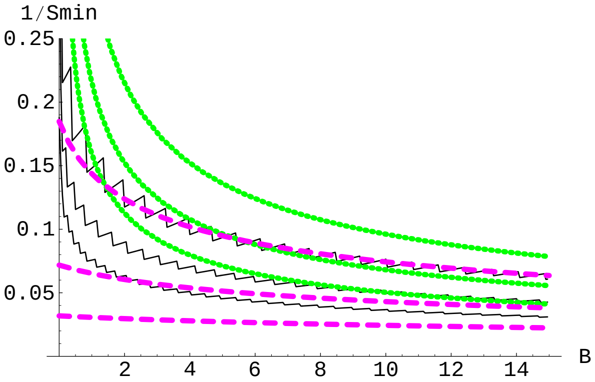


Figure 3: Comparison of $1/S_{min}$ with the corresponding sensitivity factor given by $S/\sqrt{B}$ (dotted) and $S/\sqrt{S+B}$ (dashed), for a search experiment with (significance, CL) respectively of (95%,95%), (3σ,95%), (5σ,90%)

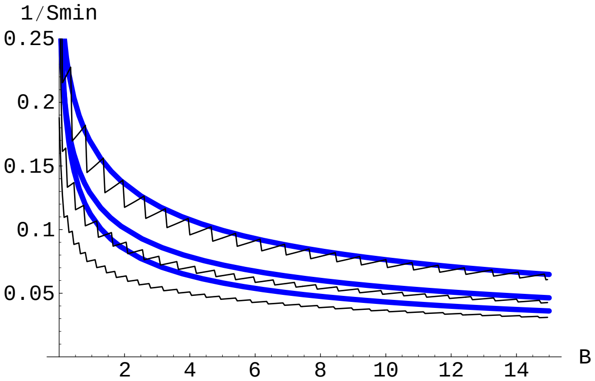


Figure 4: Gaussian approximation of the “Sensitivity factor” $1/S_{min}$ (eq. (6)) for a search experiment with (significance, CL) respectively of (95%,95%), (3σ,95%), (5σ,90%)

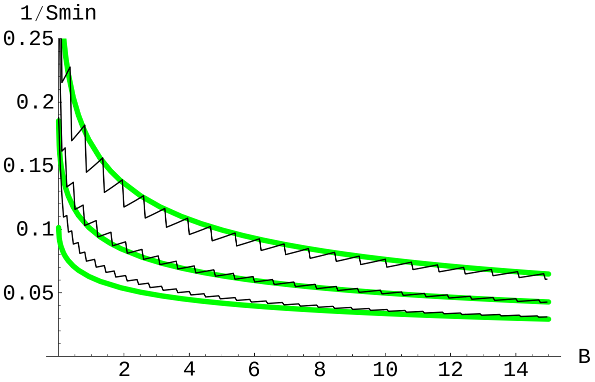


Figure 5: Gaussian approximation of $1/S_{min}$ in the $b \approx a$ approximation (eq. (6)), for a search experiment with (significance, CL) respectively of (95%,95%), (3σ,95%), (5σ,90%). Curves are normalized to the asymptotic limit.

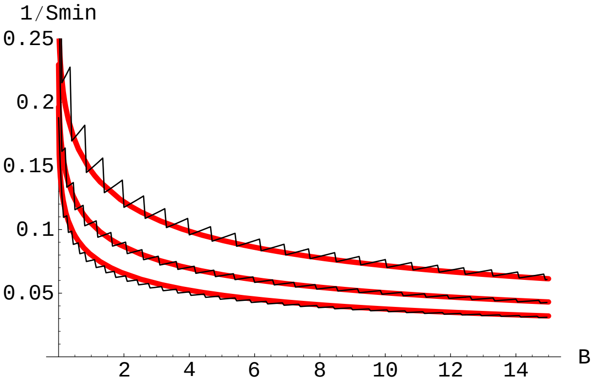


Figure 6: Improved Gaussian approximation of the “Sensitivity factor” $1/S_{min}$ (eq. (8) for a search experiment with (significance, CL) respectively of (95%,95%), (3σ,95%), (5σ,90%)

The Gaussian approximation to the exact solution is shown instead in fig. 4, and its special case for $b \approx a$ in fig. 5.

It can be seen that the approximate formulas work well at moderate values of a and b, but become less accurate when high significance/CL are desired, due to the larger deviations from Gaussian behavior that occur in the Poisson far tails. However, the Gaussian approximation can easily be improved, without losing the good features of the solutions. For instance, it is possible to obtain a more accurate expression by accounting for differences between Gaussian and Poisson tail integrals at the next order in a and b, simply by performing an empirical fit. This results in the following improved expression for S_{min}:

$$S_{min} = \frac{a^2}{8} + \frac{9\,b^2}{13} + a\sqrt{B} + \frac{b}{2}\sqrt{b^2 + 4\,a\sqrt{B} + 4\,B} \quad (8)$$

Fig. 6 shows this slightly modified expression to be considerably accurate even at high significance, which makes it suitable also for searches of “really new” effects, where a significance level of 5σ is a customary requirements.

Acknowledgments

The author wishes to thank Louis Lyons for many helpful comments.

Systematic Analysis of HEP Collider Data

Bruce Knuteson*
Massachusetts Institute of Technology

Compelling arguments suggest the presence of new physics at energy scales that will be probed by frontier energy colliders over the next decade. Arguments for each of the many flavors of new physics that have been proposed seem much less compelling. The wide variety of experimental signatures by which new physics may manifest itself suggests the desirability of analyzing all high energy collider data in one systematic framework. These proceedings describe two potentially useful pieces of such a framework: Sleuth enables a model-independent search for new high-p_T physics, and Quaero automates tests of particular hypotheses against high energy collider data. A sampling of algorithmic detail is provided in the form of a procedure for choosing an optimal binning when computing likelihood ratios.

1. CONTEXT

The audience for this talk (and these proceedings) comprises astrophysicists, cosmologists, and statisticians, in addition to high energy experimentalists. It is therefore worth beginning by discussing the nature of high energy collider data, particularly those features that make these data amenable to the algorithms described here. These data are collected by large, complex detectors that record on roughly a million channels the debris from the collisions of particles (protons, electrons, and their antimatter counterparts) travelling within a few hundred miles per hour of the speed of light.

The information contained in these million channels of electronics is reduced through a series of steps to roughly one dozen numbers, corresponding to the energies and directions (polar and azimuthal angles) of the elementary objects emerging from the collision. This severe reduction in detail facilitates a direct connection to the underlying theory. The underlying theory is most easily understood graphically in terms of Feynman diagrams, an example of which is shown in Fig. 1. Our detectors and algorithms (imperfectly) reconstruct the outgoing particles in collisions like that depicted in Fig. 1. The goal is to figure out, from the debris of trillions of particle collisions, the rules corresponding to graphs such as that shown in Fig. 1: rules for what types of graphs can be drawn, and rules for calculating observable quantities from them. In doing so, we infer from measurements on scales of meters the laws of Nature on scales of 10^{-16} meters and below.

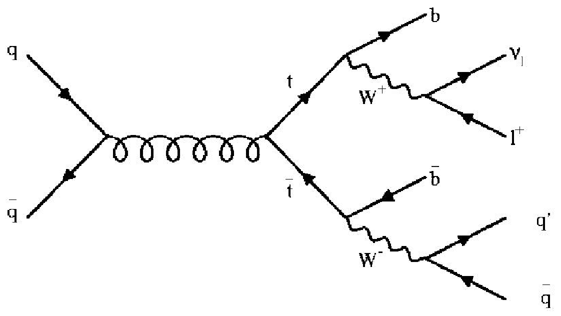


Figure 1: A Feynman diagram, showing the annihilation of a quark (q) and antiquark ($\bar{q}$), and the subsequent production and decay of a top quark (t) and an antitop quark ($\bar{t}$). Time increases to the right.

Figure 2: The theoretical landscape, as depicted in the summary talk of this year's Lepton Photon conference [1].

The theoretical context in which we work is grounded in the standard model of particle physics, which predicts the results of nearly all experiments performed to date with extraordinary accuracy — and in many cases also with extraordinary precision. This standard model represents a canonical reference model, the null hypothesis in our field.

The theoretical landscape beyond the standard model is much less clear. Hundreds of different scenarios have been proposed, each containing many parameters. The lack of clarity in this picture is nicely captured in a slide shown during the summary talk of Lepton Photon 2003, reproduced in Fig. 2.

*URL: `http://mit.fnal.gov/~knuteson/`; Electronic address: `knuteson@mit.edu`

2. Sleuth

The jumbled theoretical landscape in Fig. 2, reflecting the plethora of possible extensions to the standard model, calls into question the paradigm currently being used to explore that landscape. At present roughly one graduate student is consumed for each model tested.

An alternative way to proceed is to systematically search for any evidence of new physics that lies in the data, in a manner that is as model-independent as possible. A prescription for doing this is an algorithm called Sleuth, used by the DØ experiment in Run I to search a large subset of their data [2–5].

One of many problems faced when searching for new physics in such a directionless landscape is how to take into account the large space of possible signatures that could appear when computing a final measure of the significance of any particular result. If many students look at many plots over an extended period of time, fluctuations at the level of three or more standard deviations are bound to appear simply from the fact that thousands of bins in various histograms have been considered. The difficulty in computing this *trials factor*, the number of possible places that an interesting signal could have appeared, has hamstrung several previous search efforts that have attempted to base themselves on signatures rather than models. A rigorous accounting of the trials factor is crucial to any model-independent search; Sleuth is one of the few algorithms currently on the market that is able to compute this trials factor rigorously and explicitly. The H1 Collaboration has developed an algorithm in similar spirit for HERA physics [6].

Key to a rigorous computation of the trials factor is defining — before the data is collected — the interestingness of any particular signature that might be seen in those data. Sleuth is able to do this by making three well-justified assumptions.

1. The data can be categorized into exclusive final states in such a way that any signature of new physics is apt to appear predominantly in one of these final states.

2. New physics will appear with objects at high transverse momentum (p_T) relative to standard model and instrumental background.

3. New physics will appear as an excess of data over background.

The Sleuth algorithm consists of three steps, following these three assumptions.

In the first step, all of the collisions are partitioned into exclusive final states. The objects used to categorize these final states are high-p_T and isolated electrons (e), muons (μ), taus (τ), photons (γ), jets (j), b-tagged jets (b), and missing transverse energy ($\not\!\!E_T$).

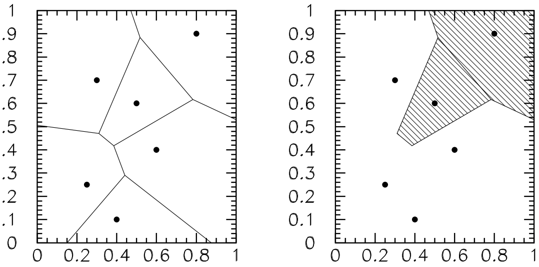


Figure 3: A Voronoi diagram with seven data points (black dots) in a unit square (left). The Run I Sleuth algorithm considers regions that are unions of these cells, such as the shaded region (right).

The second step of the algorithm defines a low-dimensional variable space for each final state. In the Run I implementation of Sleuth, the variables used were

- the summed transverse momentum of any leptons in the event ($\sum {p_T}^{e/\mu/\tau}$);
- the missing transverse energy ($\not\!\!E_T$), if significant in the event;
- the summed transverse momentum of any electroweak gauge bosons in the event ($\sum {p_T}^{W/Z/\gamma}$); and
- the summed transverse momentum of any jets in the event ($\sum {p_T}^{j}$).

The Run II algorithm is simplified enormously by considering only a single variable,

- the summed transverse momentum of all objects in the event ($\sum p_T$).

New high-p_T physics is best searched for by systematically looking for new physics at high p_T.

The algorithm's third step involves searching for regions in which more events are seen in the data than expected from standard model and instrumental background. This search is performed in the variable space defined in the second step of the algorithm, for each of the exclusive final states defined in the first step.

The details of the search are somewhat involved, but both the input and output are exceptionally simple. For each final state, the input is simply the events seen in the data, and the expected background. The steps of the search can be sketched as follows.

- The variable space is transformed into the unit box — the unit interval in one dimension, unit square in two dimensions, unit cube in three dimensions, and unit hypercube in four dimensions.

- The notion of regions about sets of data points is rigorously defined using the concept of Voronoi diagrams, borrowed from the field of computational geometry. Figure 3 shows an example of a unit square containing seven data points, shown as black dots. The perpendicular bisectors of line segments connecting each pair of data points connect to form the Voronoi diagram.

- The interestingness of any particular region (or union of such regions) in Fig. 3 is the Poisson probability that the background in that region would fluctuate up to or above the observed number of events in that region.

- The most interesting region $\mathcal{R}$ is found using a search heuristic to explore the space of potentially interesting regions.

- Pseudo experiments are performed to determine the fraction $\mathcal{P}$ of hypothetical similar experiments in which something more interesting than $\mathcal{R}$ would be seen. Here the fact that many different places have been considered is rigorously and explicitly accounted for. SLEUTH and its H1 analogue appear to be the only algorithms currently on the market for frontier energy collider physics that compute this trials factor completely and systematically.

- The results from all final states considered are then combined to form $\tilde{\mathcal{P}}$, which quantifies the interestingness of the most interesting region observed in the data, accounting for the fact that many final states have been considered.

The Run II algorithm is trivial by comparison. In the single variable $\sum p_T$, semi-infinite regions are defined with a lower bound at each data point. The definition of interestingness, running of pseudo experiments, and calculation of $\mathcal{P}$ and $\tilde{\mathcal{P}}$ proceed as above.

The output of the algorithm is the most interesting region $\mathcal{R}$ observed in the data, and a number $\tilde{\mathcal{P}}$ that quantifies the interestingness of $\mathcal{R}$. $\tilde{\mathcal{P}}$ is a number between zero and unity, pulled from a uniform distribution on the unit interval if the data comes from background alone, and expected to be small if the data contain a hint of new physics. A reasonable threshold for discovery is $\tilde{\mathcal{P}} \lesssim 0.001$, which corresponds loosely to the *de facto* 5σ standard in our field after the trials factor is accounted for. [1]

Two questions must now be asked:

- Will SLEUTH find nothing if there is nothing to be found?

- Will SLEUTH find something if there is something to be found?

The answer to the first is "yes," by construction.[2] The answer to the second depends to what extent the new physics waiting to be uncovered satisfies the three assumptions on which SLEUTH is based.

	Omitted		
	WW, $t\bar{t}$	$t\bar{t}$	none
Final state	$\mathcal{P}$		
$e\mu\not{E}_T$	$\mathbf{2.4\sigma}$	1.1σ	1.1σ
$e\mu\not{E}_Tj$	0.4σ	0.1σ	0.1σ
$e\mu\not{E}_Tjj$	$\mathbf{2.3\sigma}$	$\mathbf{1.9\sigma}$	0.5σ
$e\mu\not{E}_Tjjj$	0.3σ	0.2σ	-0.5σ
$\tilde{\mathcal{P}}$	1.9σ	1.2σ	-0.6σ

Table I Summary of a SLEUTH sensitivity study on the $e\mu\not{E}_T$, $e\mu\not{E}_Tj$, $e\mu\not{E}_Tjj$, and $e\mu\not{E}_Tjjj$ final states. When the standard model processes WW and $t\bar{t}$ are omitted from the background estimate (second column), SLEUTH identifies a region of excess in the $e\mu\not{E}_T$ and $e\mu\not{E}_Tjj$ final states (with $\mathcal{P} = 2.4\sigma$ and 2.3σ, respectively), presumably indicating the true presence of WW and $t\bar{t}$ in the data. When the standard model process WW is included and $t\bar{t}$ omitted (third column), SLEUTH identifies a region of excess in the $e\mu\not{E}_Tjj$ final state (with $\mathcal{P} = 1.9\sigma$), presumably indicating the true presence of $t\bar{t}$ in the data. With all standard model processes included to search for new physics (third column), SLEUTH indicates that 72% ($\tilde{\mathcal{P}} = -0.6\sigma$) of background-only hypothetical similar experiments would have produced a region more interesting than the most interesting region observed in these data.

Although no general answer can be given to this second question, an answer can be given for any specific case. Such a specific case is summarized in Table I [2]. Events containing an energetic electron, muon, and possibly other objects ($e\mu X$) are considered. In a first pass, standard model WW and $t\bar{t}$ production are omitted from the background estimate to see if

[1] The threshold of $\tilde{\mathcal{P}} \lesssim 10^{-3}$ follows directly from our field's standard discovery threshold of five standard deviations in the following manner. We recall that 5σ corresponds to a probability of roughly 10^{-7}. In a collaboration the size of CDF there are ≈ 100 graduate students. Each student works 2 years on his analysis, and makes 1 interesting plot per week. 100 students × 50 weeks/year × 2 years = 10^4 different "things" that are looked at. $10^{-7} \times 10^4 = 10^{-3}$, which corresponds to roughly 3 standard deviations. The desire to see a 5σ effect is therefore a desire to see a 3σ effect when the accounting is done for all possible places a signal could have appeared, but did not. SLEUTH includes this accounting (performed much more rigorously than in this footnote) in the calculation of $\tilde{\mathcal{P}}$.

[2] Spurious signals will of course be seen if SLEUTH is provided improperly modeled backgrounds. SLEUTH directly addresses the issue of whether an observed hint is due to a statistical fluctuation; it is unable to address systematic mismeasurement or incorrect modeling (but quite useful in bringing these to your attention).

SLEUTH is able to find evidence of these processes in DØ Run I data, and the result $\mathcal{P}$ obtained in each final state (translated into units of standard deviations) is shown. SLEUTH finds $\mathcal{P} = 2.4\sigma$ and 2.3σ in the final states $e\mu\not\!\!E_T$ and $e\mu\not\!\!E_T jj$; these excesses correspond (presumably) to the true presence of WW and $t\bar{t}$ in these data. For comparison, a dedicated search for WW in Run I at CDF [7] resulted in 5 events observed on a background of 1.2 ± 0.3, corresponding to a significance of 2.3σ; and a dedicated search in $e\mu X$ for $t\bar{t}$ by DØ in Run I [8] resulted in 5 events observed on a background of 1.4 ± 0.4, corresponding to significance of 2.1σ.

The quantity $\mathcal{P}$ obtained from SLEUTH really should not be directly compared to the result of a dedicated search, since the two techniques are intended for very different problems: dedicated searches are clearly preferred if there are well-defined, compelling things to be found, while SLEUTH provides an alternative strategy in their absence. This example nonetheless provides useful intuition for SLEUTH's performance on a difficult test.

In a second pass, standard model WW production is included in the background estimate, with standard model $t\bar{t}$ production still omitted, to see whether SLEUTH could find evidence of $t\bar{t}$ in these data. The results obtained are shown in the third column of Table I, with the excess in the $e\mu\not\!\!E_T jj$ final state corresponding (presumably) to the actual presence of $t\bar{t}$ in these data. The slight indications of excess in these examples clearly fall well short of that needed to make a discovery claim; as indicated above, these are difficult tests.

With all backgrounds included and SLEUTH used to search for new physics in the fourth column of Table I, a null result is obtained. The use of SLEUTH to analyze roughly thirty additional final states at DØ in Tevatron Run I resulted in no evidence of new physics [2–5].

A general model-independent search in similar spirit [6] has recently been presented by the H1 collaboration at the 2003 European Physical Society meeting in Aachen, Germany. It will be interesting to continue to watch their $\mu j\nu$ final state in HERA Run II.

3. QUAERO

The first hint of new physics at Tevatron Run II may come from a model-independent search. Once such a hint is found, it must be interpreted in terms of an underlying physical theory. This interpretation would clearly be facilitated by some means of quickly and efficiently testing the predictions of many different hypotheses against the data. QUAERO (Latin for "I search for," or "I seek") is an algorithm designed for this purpose.

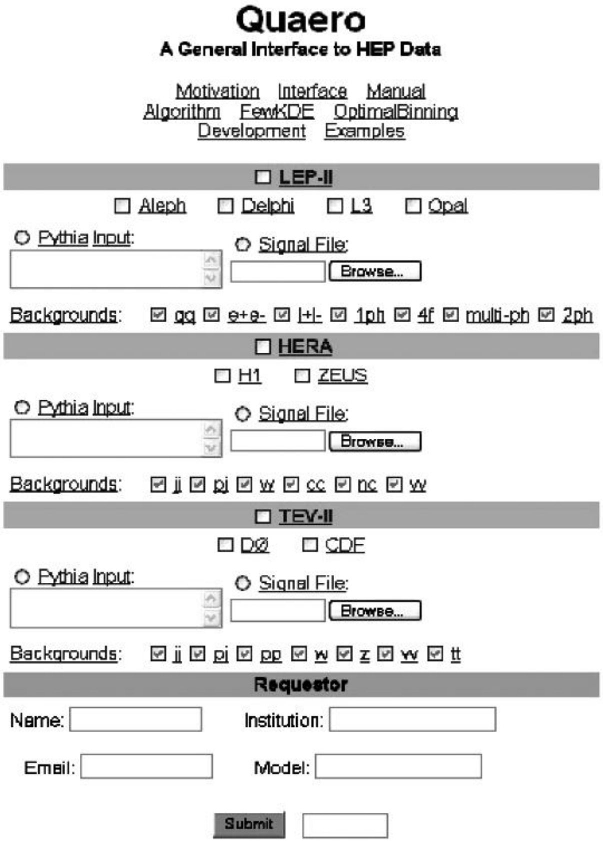
Present practice for testing hypotheses against collider data can be improved upon in several respects. A personal wish list for conducting analyses on high energy collider data includes:

- Reducing the time spent to perform an analysis from two years of one graduate student's life to roughly an hour of CPU time. Achieving this would represent a reduction in the time it takes to perform an analysis by a factor of 10^4.
- Reducing human bias that invariably creeps into analysis on complex data sets.
- Allowing the publication of data in their full dimensionality, unrestricted by the two dimensions of a sheet of paper.
- Providing an alternative to exclusion contours. The exclusion plots often shown make it difficult to understand exactly what model is being tested, together with all assumptions that are made, and difficult to tell what the data have to say about a model that does not lie in that two-dimensional space.
- Automating the optimization of analysis, to ensure the data are used to their fullest.
- Rigorously propagating systematic errors in an intuitive, straightforward, and rigorous way.
- Combining results among correlated experiments in a manner that requires as few *ad hoc* prescriptions as possible.
- Increasing the robustness of our scientific results by using a high-level analysis algorithm that has been validated on hundreds of previous analyses.
- All of this on the web.

A first pass of such an algorithm has been achieved. With the posting of an article entitled "Search for New Physics Using QUAERO: A General Interface to DØ Event Data" [9], DØ has made a subset of data collected in Tevatron Run I available on the web at `http://quaero.fnal.gov/` since June 2001.

Astrophysicists have become accustomed to polished interfaces to their data; the web page served up by the Sloan Digital Sky Survey at `http://www.sdss.org/` is one of many examples. Those in the audience with this image in mind are bound to be disappointed by the look and feel of Fig. 4 — high energy physics is at least a decade behind the astrophysics and astronomy communities on this front.

The essential QUAERO interface, devoid of adornment, is displayed in Fig. 4. A physicist with a particular hypothesis $\mathcal{H}$ to test against high energy collider data should be able to provide his hypothesis in the form of the events his model predicts — either as input to an event generator, or as a file with the events



Quaero

A General Interface to HEP Data

Motivation Interface Manual
Algorithm FewKDE OptimalBinning
Development Examples

□ LEP-II

□ Aleph □ Delphi □ L3 □ Opal

○ Pythia Input: ○ Signal File: Browse...

Backgrounds: ☑ qq ☑ e+e- ☑ l+l- ☑ 1ph ☑ 4f ☑ multi-ph ☑ 2ph

□ HERA

□ H1 □ ZEUS

○ Pythia Input: ○ Signal File: Browse...

Backgrounds: ☑ jj ☑ pj ☑ w ☑ cc ☑ nc ☑ w

□ TEV-II

□ DØ □ CDF

○ Pythia Input: ○ Signal File: Browse...

Backgrounds: ☑ jj ☑ pj ☑ pp ☑ w ☑ z ☑ vv ☑ tt

Requestor

Name: Institution:

Email: Model:

Submit

Figure 4: The QUAERO web page under development for frontier energy collider data.

themselves. These events (the “signal”), together with whatever standard model processes (“backgrounds”) he wishes to include, define his hypothesis for how Nature works at very small distance scales.

After providing his name and the email address to which the result should be sent, the physicist clicks “Submit.” QUAERO then performs the complete analysis, taking into full account the expert knowledge gleaned within the collaboration and packaged into code, and returns a single number, quantifying the extent to which the data (dis)favor the hypothesis relative to the standard model. QUAERO also provides a number of plots showing in detail how the analysis was performed. Far from being a black box, QUAERO arguably provides a much more transparent view into how analyses are performed than our standard publications.

The QUAERO algorithm itself is relatively simple, involving a few straightforward steps.

- The events predicted by the hypothesis $\mathcal{H}$ are run through the detector simulation appropriate for each experiment.
- Events from $\mathcal{H}$, the standard model (SM), and the data ($\mathcal{D}$) are partitioned into exclusive final states. Speaking loosely, these final states are orthogonal (no event belongs to more than one final state) and complete (every event belongs to a final state).
- Variables are chosen automatically within each final state.
- A binning is chosen automatically within the variable space in each final state.
- A binned likelihood is calculated within each final state.
- Results from different final states are combined.
- Results from different experiments are combined.
- Systematic errors are integrated numerically.
- The result returned is a likelihood ratio,

$$\mathcal{L}(\mathcal{H}) = \frac{p(\mathcal{D}|\mathcal{H})}{p(\mathcal{D}|\mathrm{SM})}. \qquad (1)$$

In order to provide a feeling for the details of the algorithm within the space constraints of these proceedings, one piece of the algorithm is highlighted: automatic choice of binning.

4. OPTIMAL BINNING FOR LIKELIHOOD RATIOS

A binned likelihood provides a robust yet sensitive method for discriminating between two hypotheses. But how should the bins be chosen? Somewhat surprisingly, the literature does not yet appear to contain a satisfactory general prescription for choosing an optimal binning. This section suggests such a prescription, investigates its implications in several limiting cases, and provides examples of its use.

Figures 5(a) and (b) show a typical problem. Predicted (analytic) distributions from two hypotheses h and b are shown in Fig. 5(a). Units on the vertical axis are the number of predicted events per unit of x, the observable shown on the horizontal axis. Often the analytic form of the predictions are not known, however; knowledge of the predictions from h and b come in the form of an ensemble of Monte Carlo events, whose statistics are limited by the complexity of the simulation required for each event.

It is desired to perform an experiment to collect data d to determine whether hypothesis h or b is the more accurate description of Nature. The number we would like to determine is the likelihood ratio $p(d|h)/p(d|b)$ — in words, the probability of obtaining the data d assuming the correctness of the hypothesis

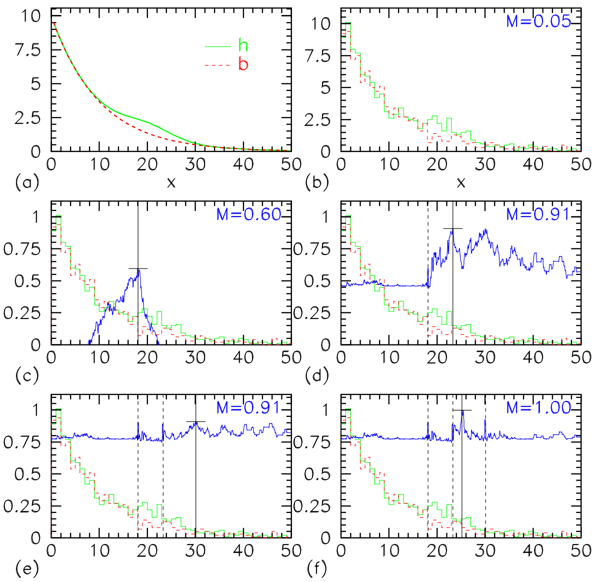


Figure 5: Placement of bins in a toy example: a bump on a falling exponential. The true (unknown, analytic) distributions $h(x)$ (solid, green) and $b(x)$ (dashed, red) are shown in (a); our knowledge of these distributions, in the form of 1000 Monte Carlo points drawn from each, is shown in (b). In this case $b(x)$ is a simple exponential, with $h(x)$ adding a Gaussian bump centered at $x = 20$. The vertical axes in (a) and (b) represent the number of events expected per unit of x. Sequential placement of bin edges is shown in (c)–(f), with the figure of merit $\mathcal{M}$ on the vertical axis.

h, divided by the probability of obtaining the data d assuming the correctness of the hypothesis b.

Given the predictions from h and b shown in Fig. 5(b), how should this likelihood ratio be computed? If the predictions $h(x)$ and $b(x)$ were known as analytic functions of x, as in Fig. 5(a), an unbinned likelihood could be calculated. But the analytic forms $h(x)$ and $b(x)$ are not known. Constructing smooth distributions $h(x)$ and $b(x)$ from Monte Carlo points using smoothing techniques is possible, but the final answer is often unfortunately sensitive to the details of how this smoothing is performed. The only reasonable option appears to involve the introduction of bins, and the computation of a binned likelihood.

But then how should the bins be set? The bins must clearly be fine enough to probe the difference in shape between the two distributions; the bins must just as clearly be large enough that an accurate prediction is obtained for the number of events h_k and b_k in each bin k. The issue at hand is not only how many bins to use, but also where to place their edges.

There is no unique solution to this problem. The best one can do is to define a reasonable prescription for choosing an optimal binning — in effect, by suggesting some reasonable definition of "optimal" — and demonstrate its reasonable behavior on a variety of examples.

The prescription suggested here involves defining a figure of merit $\mathcal{M}$ by

$$\mathcal{M} = \sum_{d_1=0}^{\infty}\sum_{d_2=0}^{\infty}\cdots\left(\prod_k p(d_k|p(h_k))\right)\times \log\left(\prod_k \frac{p(d_k|p(h_k))}{p(d_k|p(b_k))}\right) + (h\leftrightarrow b) - \mathcal{P}. \qquad (2)$$

In words, $\mathcal{M}$ is the evidence the experiment is expected to provide in favor of h if h is correct, plus the evidence the experiment is expected to provide in favor of b is b is correct. The definition of "evidence" here, adopted from Ref. [10], is the logarithm of the likelihood ratio; "expected" is defined in terms of an average over an ensemble of hypothetical experiments, where the correctness of either h or b is assumed in weighting the possible outcomes.

The initial sum in Eq. 2 is over all possible outcomes of the experiment: the number of data events d_k in each bin k is allowed to vary between zero and infinity. The factor $\prod_k p(d_k|p(h_k))$ weights each outcome by the probability of its occurrence, assuming the correctness of h. Here $p(h_k)$ is our knowledge of the number of events predicted by h in bin k; we might have $p(h_k)$ in the form of a Gaussian with mean 7 and width 1.2 if the number of events predicted by h in bin k were 7 ± 1.2. The factor $\log\left(\prod_k \frac{p(d_k|p(h_k))}{p(d_k|p(b_k))}\right)$ is the evidence obtained in favor of h in this outcome. To this is added a similar term with h and b swapped; the second term $(h \leftrightarrow b)$ is the expected evidence in favor of b if b is correct. The third term $\mathcal{P}$ is a penalty term, which provides the stopping condition for the algorithm's placement of bins.

Figures 5 and 6 show how this figure of merit $\mathcal{M}$ can be used to determine the placement of bins. Figure 5(a) shows the true (analytic and unknown) predictions $h(x)$ and $b(x)$ from the hypotheses h and b in the observable x. Figure 5(b) shows our knowledge of the predictions of these two hypotheses, in the form of one thousand Monte Carlo points drawn from the true (unknown) distributions $h(x)$ and $b(x)$. Using a single bin from 0 to 50 in x, the figure of merit computed using Eq. 2 for the points shown in Fig. 5(b) is $\mathcal{M} = 0.05$.

Figures 5(c)–(f) show the successive placement of bin edges. In these plots the vertical axis has units of expected evidence; the predictions of h and b are superimposed with arbitrary scale. At each value of x, the dark (jagged, blue) curve shows the figure of merit $\mathcal{M}$ if a bin edge is placed at that point. In Fig. 5(c), the maximum of this curve is obtained with a bin edge placed at $x = 18$; this raises the figure of merit to $\mathcal{M} =$

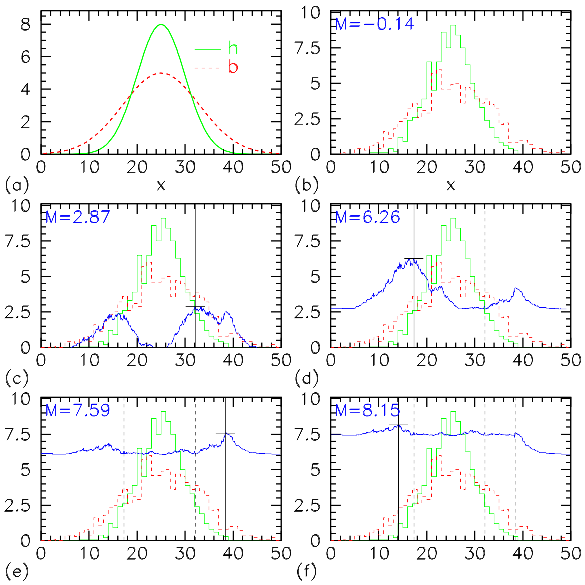


Figure 6: Placement of bins in a toy example: two Gaussians of different widths. The true (unknown, analytic) distributions $h(x)$ (solid, green) and $b(x)$ (dashed, red) are shown in (a); our knowledge of these distributions, in the form of 1000 Monte Carlo points drawn from each, is shown in (b). In this case $b(x)$ is a Gaussian centered at 25 with width 8; $h(x)$ is a Gaussian with the same mean and width 5. The vertical axes in (a) and (b) represent the number of events expected per unit of x. Sequential placement of bin edges is shown in (c)–(f), with the figure of merit $\mathcal{M}$ on the vertical axis.

0.60. Placing a bin edge at this point results in one bin stretching from 0 to 18, and one bin reaching from 18 to 50. Figure 5(d) shows this process repeated, the figure of merit calculated for each possible location of a second bin edge. Maximizing the expected evidence in the dark (jagged, blue) curve requires placement of a bin edge at $x = 23$. This placement leaves three bins: [0–18], [18–23], and [23–50]. Figures 5(e) and (f) show the placement of two more bin edges, at $x = 30$ and at $x = 25$. Further placement of bin edges decreases the figure of merit $\mathcal{M}$, so the algorithm halts.

A second example is shown in Fig. 6. Figure 6(a) shows the true (analytic and unknown) predictions $h(x)$ and $b(x)$, both Gaussians with identical mean and area but different widths. One thousand Monte Carlo points pulled from each of $h(x)$ and $b(x)$ are shown in Fig. 6(b). The use of a single bin from 0 to 50 results in a figure of merit of $\mathcal{M} = -0.14$; a negative value is obtained because the total number of events predicted by h and b in this single encompassing bin is the same (the Gaussians have equal area), and the penalty term $\mathcal{P}$ in Eq. 2 drives the figure of merit $\mathcal{M}$ negative. In Fig. 6(c)–(f), the units of the vertical axes are expected evidence, with the predictions of h and b again superimposed. Notice the difference in vertical scale between Figs. 5(c)–(f) and Figs. 6(c)–(f); the evidence we expect the experiment to provide in favor of h relative to b (or vice versa) is clearly much larger in the example of Fig. 6.

A first bin edge is placed in Fig. 6(c), the figure of merit $\mathcal{M}$ computed as the bin edge's position is scanned in x, resulting in the dark (jagged, blue) curve. As expected by looking at the true distributions for h and b, the algorithm prefers bin placement at $x \approx 15$ or $x \approx 35$, where the analytic predictions for h and b cross in Fig. 6(a). In Fig. 6(c), placement of a bin edge at $x = 32$ is slightly favored. The first bin edge is placed at this point, resulting in one bin ranging from 0 to 32, and a second bin covering 32 to 50. The process is repeated, with the expected evidence curve shown in Fig. 6(d), and a second bin is placed at $x = 17$, raising the total figure of merit to $\mathcal{M} = 6.26$. Figures 6(e) and (f) show the process repeated twice more, raising the total figure of merit to $\mathcal{M} = 8.15$. The algorithm places eight additional bin edges in the regions $x \approx 20$ and $x \approx 30$ before halting.

The algorithm's performance in these two cases is remarkably intuitive. In the first example, the procedure nicely carves out the region around the bump that $h(x)$ shows relative to $b(x)$ in Fig. 5(a), correctly ignoring the bulk of the distribution at $x < 10$ and the tail at $x > 30$. In the second example, the algorithm systematically works from side to side in Fig. 6, from the right of the mean to the left of the mean and back, doggedly separating regions in which h predicts more events than b from regions in which b predicts more than h.

The algorithm presented here has at least two multivariate generalizations. One option iteratively places bin edges in the form of hyperplanes parallel to the variable axes, creating a grid in the multidimensional space. In some cases this may be an acceptable approach, but the resulting rectangular bins are too constrained in shape to adequately handle an arbitrary multidimensional problem. An (improved) alternative is to use kernel density estimation to first reduce the problem to a single dimension, enabling the application of the one-dimensional binning algorithm just described.

5. FewKDE

Standard kernel estimation involves placing bumps of probability, typically in the form of Gaussian kernels, around each Monte Carlo point. Summation of kernels placed around each of an ensemble of Monte Carlo points forms the density estimate.

In this standard approach, the evaluation of the density at any particular point requires the evaluation of a Gaussian centered at each of of the N_{MC}

Monte Carlo points. The time cost of evaluating this density estimate at each of the points used to generate the estimate thus grows as $\mathcal{O}(N_{MC}^2)$, which becomes prohibitive when dealing with samples of $\gtrsim 10^4$ Monte Carlo points. Application to high statistics Tevatron and future LHC analyses is facilitated by noting that distributions derived from four-vector quantities of final state objects in high-p_T collider physics can be approximated satisfactorily by the sum of just a few Gaussians.

An algorithm called FewKDE has been introduced with the generally featureless nature of our distributions in mind, where "FewKDE" is shorthand for "kernel density estimation with few kernels." The parameters of the few Gaussians are chosen to provide the best fit to the data. A novel technique is employed to appropriately handle the types of hard physical boundaries (such as $p_T > 0$) that exist in commonly considered distributions.

6. SUMMARY

These proceedings have briefly sketched a method allowing the systematic analysis of high energy collider data. After briefly providing the experimental and theoretical contexts of frontier energy collider data to the statisticians, astrophysicists, and cosmologists in the audience, a direct solution to a few of the problems we face in the analysis of those data has been described.

Given the variety of possible forms physics beyond the standard model may take, the question of how to search for something when we know only vaguely what it is we are searching for becomes acute. Sleuth is an algorithm that accomplishes this in a rigorous and systematic way, enabling a model-independent search for new high-p_T physics.

Once a hint of new physics is observed, data understood in the context of a systematic search must be interpreted in terms of the underlying physical theory. Accomplishing this requires a procedure for quickly and efficiently testing particular hypotheses against the data. Quaero provides a qualitatively new medium for facilitating this interpretation.

In order to provide a feeling for one of several algorithmic pieces introduced in the development of Quaero, a procedure for optimally choosing a binning for the computation of a binned likelihood ratio has been described. Generalization to the multivariate case makes use of FewKDE, a time-saving variant of the standard procedure for kernel density estimation.

It is our hope that the ideas presented here, developed for a particular problem within high energy physics, may lend themselves to many other problems in the physical sciences.

Acknowledgments

The PHYSTAT2003 conference organizers and SLAC provided a wonderful opportunity to share techniques and ideas in a creative atmosphere. Mark Strovink, Hugh Montgomery, Greg Landsberg, Dave Toback, and others on the DØ collaboration contributed to the development of Sleuth and Quaero, and their application to high energy collider data. The application of Sleuth and Quaero to data has been further advanced by Daniel Whiteson at DØ, Marcello Maggi at ALEPH, André Holzner at L3, Sascha Caron at H1, and others. Additional invaluable logistical assistance has been provided by many, including Bolek Pietrzyk, Jacques Boucrot, and Mark Oreglia. Khaldoun Makhoul and Jang Woo Lee (MIT) have aided the implementation of the new Quaero and Sleuth algorithms. Hannu Miettinen, Paul Padley, David Scott, and Sang-Joon Lee (Rice University) have assisted in the testing of FewKDE and other issues related to kernel density estimation. Henry Frisch and Mel Shochet (University of Chicago) have provided helpful discussions.

This work has been funded by a Department of Defense National Defense Science and Engineering Graduate Fellowship at the University of California, Berkeley; an International Research Fellowship from the National Science Foundation (INT-0107322); a Fermi/McCormick Fellowship at the University of Chicago; and Department of Energy grant DE-FC02-94ER40818.

References

[1] Hitoshi Muryama. Outlook: The Next Twenty Years. In *XXI International Symposium on Lepton and Photon Interactions at High Energies*, Fermilab, IL, USA, 2003. `http://conferences.fnal.gov/lp2003/`.

[2] B. Abbott et al. Search for new physics in $e\mu X$ data at DØ using Sleuth: a quasi model independent search strategy for new physics. *Phys. Rev.*, D62:092004, 2000.

[3] B. Abbott et al. A quasi-model-independent search for new physics at large transverse momentum. *Phys. Rev.*, D64:012004, 2001.

[4] B. Abbott et al. A quasi-model-independent search for new high p_T physics at DØ. *Phys. Rev. Lett.*, 86:3712–3717, 2001.

[5] B. Knuteson. PhD thesis, University of California, Berkeley, 2000.

[6] Martin Wessels. Generic Searches at HERA. In *International Europhysics Conference on High Energy Physics*, Aachen, Germany, 2003. `http://eps2003.physik.rwth-aachen.de/`.

[7] F. Abe et al. Observation of W^+W^- production in $p\bar{p}$ collisions at $\sqrt{s} = 1.8$ TeV. *Phys. Rev. Lett.*, 78:4536–4540, 1997.

[8] S. Abachi et al. Measurement of the top quark pair production cross section in $p\bar{p}$ collisions. *Phys. Rev. Lett.*, 79:1203–1208, 1997.

[9] V. M. Abazov et al. Search for new physics using Quaero: a general interface to DØ event data. *Phys. Rev. Lett.*, 87:231801, 2001.

[10] E. Jaynes. *Probability Theory with Applications in Science and Engineering*. Cambridge University Press, 2003.

Challenges in Moving the LEP Higgs Statistics to the LHC

K.S. Cranmer, B. Mellado, W. Quayle, Sau Lan Wu
University of Wisconsin-Madison, Madison, WI 53706, USA

We examine computational, conceptual, and philosophical issues in moving the statistical techniques used in the LEP Higgs working group to the LHC.

1. INTRODUCTION

Higgs searches at LEP were based on marginal signal expectations and small background uncertainties. In contrast, Higgs searches at the LHC are based on strong signal expectations and relatively large background uncertainties. Based on our experience with the LEP Higgs search, our group tried to move the tools we had developed at LEP to the LHC environment. In particular, our calculation of confidence levels was based on an analytic computation with the Fast Fourier Transform and the log-likelihood ratio as a test statistic (and systematic errors based on the Cousins-Highland approach). We encountered three types of problems when calculating ATLAS' combined sensitivity to the Standard Model Higgs Boson: problems associated with large numbers of expected events, problems arising from very high significance levels, and problems related to the incorporation of systematic errors.

Previously, it was shown that the migration of the statistical techniques that were used in the LEP Higgs Working Group to the LHC environment is not as straightforward as one might naïvely expect [1]. After a brief overview in Section 2, those difficulties and their ultimate solution are discussed in Section 3. Our group has developed two independent software solutions (both in C++; both with FORTRAN bindings; one ROOT based and the other standalone) which can be found at:

http://wisconsin.cern.ch/software

In Section 4 we discuss the incorporation of systematic errors and compare a few different strategies. In Section 5 we present and discuss the discovery luminosity (the luminosity expected to be required for discovery). Lastly, in Section 6 we discuss the statistical notion of *power* (which is related to the probability of Type II error (the probability we do reject the "signal-plus-background hypothesis" when it is true).

2. THE FORMALISM

Our starting point for this note is a brief review of the techniques that were used at LEP. We refer the interested reader to [2] for an introduction to the fundamentals, to [3] for why the likelihood ratio has been chosen as a test statistic, to [4] for a Monte Carlo approach to the calculation and to [5] for the analytic calculation using Fast Fourier Transform (FFT) techniques. For completeness, we introduce the basic approach below using the notation found in [1]. For a counting experiment where we expect, on average, b background events and s signal events, we consider two hypotheses: the null (or background-only) hypothesis in which the number of expected events, n, is described by a Poisson distribution $P(n;b)$ and the alternate (or signal-plus-background) hypothesis in which the number of expected events is described by a Poisson distribution $P(n;s+b)$. Here the number of events serves the purpose of a test statistic: a real number which quantities an experiment.

It is possible to include a discriminating variable x which has some probability density function (pdf) for the background, $f_b(x)$, and some pdf for the signal, $f_s(x)$, both normalized to unity. Given an observation at x we can construct the Likelihood Ratio $Q = (sf_s(x) + bf_b(x))/bf_b(x)$. With several independent observations $\{\hat{x}_i\}$ we can consider the combined likelihood ratio $Q = \prod Q_i$. It is possible, and in some sense optimal, to use Q (or in practice $q = \ln Q$) as a test statistic.

The computational challenge of using the log-likelihood ratio in conjunction with a discriminating variable x is the construction of the log-likelihood ratio distribution for the background-only hypothesis, $\rho_b(q)$, and for the signal-plus-background hypothesis $\rho_{s+b}(q)$. In this case, there are not only the Poisson fluctuations of the number of events, but also the continuously varying discriminating variable x. In particular, for a single background event the log-likelihood ratio distribution, $\rho_{1,b}(q)$, must incorporate all possible values of x. From these single event distributions we can build up the expected log-likelihood ratio distribution by repeated convolution. This is most effectively done by using a Fast Fourier Transform (FFT) where convolution can be expressed as multiplication in the frequency domain (denoted with a bar). In particular we arrive at:

$$\begin{aligned} \overline{\rho_b(q)} &= e^{b[\overline{\rho_{1,b}(q)}-1]} \quad \text{and} \\ \overline{\rho_{s+b}(q)} &= e^{(s+b)[\overline{\rho_{1,s+b}(q)}-1]}. \end{aligned} \tag{1}$$

From the log-likelihood distribution of the two hypotheses we can calculate a number of useful quantities. Given some experiment with an observed log-likelihood ratio, q^*, we can calculate the background-only confidence level, CL_b :

$$CL_b(q^*) = \int_{q^*}^{\infty} \rho_b(q')dq' \tag{2}$$

In the absence of an observation we can calculate the expected CL_b given the signal-plus-background hypothesis is true. To do this we first must find the median of the signal-plus-background distribution $\overline{q}_{s+b}$. From these we can calculate the expected CL_b by using Eq. 2 evaluated at $q^* = \overline{q}_{s+b}$.

Finally, we can convert the expected background confidence level into an expected Gaussian significance, $N\sigma$, by finding the value of N which satisfies

$$CL_b(\overline{q}_{s+b}) = \frac{1 - \mathrm{erf}(N/\sqrt{2})}{2}. \tag{3}$$

where $\mathrm{erf}(N) = (2/\pi)\int_0^N \exp(-y^2)dy$ is a function readily available in most numerical libraries.

3. NUMERICAL DIFFICULTIES

The methods described in the previous section have been applied to the combined ATLAS Higgs effort with some caveats related to numerical difficulties [1]. In particular, in the extreme tails of $\rho_b(q)$, the probability density is dominated by numerical noise. This numerical noise is an artifact of round-off error in the double precision numbers used in the Fast Fourier Transform[1]. The noise is on the order of 10^{-17} (for double precision floating point numbers), which translates into a limit on the significance of about 8σ. For particular values of the Higgs mass, ATLAS has an expected significance well above 8σ with only 10 fb^{-1} of data. In order to produce significance values above the 8σ limit, various extrapolation methods were used in [1]. We now introduce a definitive solution to this problem based on arbitrary precision floating point numbers.

It should be made clear that the numerical precision problem is not due to the fact that the CL_b is so small that the evaluation of the integral in Eq. 2 cannot be treated with double precision floating point numbers. Instead, the numerical precision problem is due to the many (approximately 2^{20}) Fourier modes which must in total produce a number very close to 0. In order to rectify this problem we have implemented the Fast Fourier Transform with the arbitrary-precision floating point numbers provided in the CLN library[2] [6].

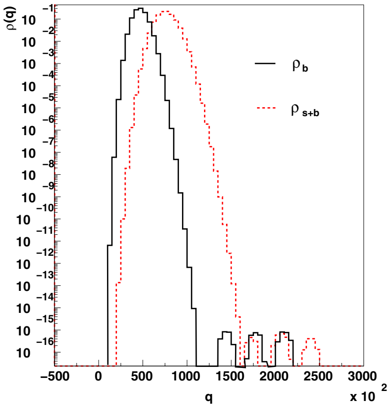


Figure 1: The distribution of the log-likelihood ratio $\rho(q)$ for the null and alternate hypothesis (the axis labels refer to bins of q, not q itself). For $q > 10^5$ the distribution is contaminated by numerical noise (see text for details).

One might protest that above 5σ we are not interested in the precise value of the significance and that this exercise is purely academic. We refer the interested reader to Sections 5 & 6 for different summaries of an experiments discovery potential.

3.1. Extrapolation

While the arbitrary precision FFT approach is the definitive solution to the problem of calculating very high expected significance, it is also incredibly time consuming. A much faster, approximate solution is to approximate the CL_b by fitting the ρ_b distribution to a functional form. The first method of extrapolation studied was a simple Gaussian fit to the ρ_b distribution. This method works fairly well, but tends to overestimate the significance. The second method we studied was based on a Poisson fit to the ρ_b distribution. The Poisson distribution has the desirable properties that it will have no probability below the hard limit $q \geq -s$ and that its shape is more appropriate [1]. Figure 2 compares these different extrapolation methods.

4. INCORPORATING SYSTEMATIC UNCERTAINTY

One encounters both philosophical and technical difficulties when one tries to incorporate uncertainty on the predicted values s and b found in Eq. 1. In a Frequentist formalism the unknown s and b become nuisance parameters. In a Bayesian formalism, s and b can be marginalized by integration over their respective priors. At LEP the practice was to smear ρ_b and ρ_{s+b} by integrating s and b with a multivariate normal distribution as a prior. This smearing technique is

[1] We use the FFTW library: http://www.fftw.org
[2] CLN is available at http://www.ginac.de

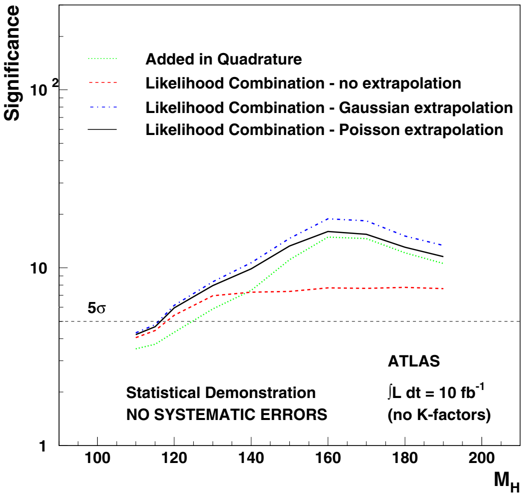


Figure 2: Comparison of the ATLAS Higgs combined significance obtained from several approximate techniques. The (red) dashed line corresponds to the unmodified likelihood ratio which can not produce significance values above about 8σ (see text). This figure is meant to demonstrate the different methods of combination and does not include up-to-date results from the various Higgs analyses.

commonly referred to as the Cousins-Highland Technique, and it is has some Bayesian aspects.

4.1. A Purely Frequentist Technique

At the PhysStat2003 conference a purely frequentist approach to hypothesis testing with background uncertainty was presented [7]. This method relies on the full Neyman construction and uses a likelihood ratio similar to the profile method as an ordering rule. In this formalism, a systematic uncertainty at the level of 10% has a much larger effect than when treated with the Cousins-Highland technique.

4.2. The CousinsHighland Technique

The Cousins-Highland formalism for including systematic errors on the normalization of the signal and background is provided in [8] and generalized in [4, 5]. In particular, for a multivariate normal distribution[3] as a prior for the n_i the distribution of the log-likelihood ratio is given by:

$$\overline{\rho_{sys}(q)} = \int ... \int e^{\sum_i^K n_i[\overline{\rho_{1,i}(q)}-1]} \left(\frac{1}{\sqrt{2\pi}}\right)^K \frac{1}{\sqrt{|S|}} \qquad (4)$$
$$e^{\sum_i^K \sum_j^K -\frac{1}{2}(n_i-\langle n_i\rangle)S_{ij}^{-1}(n_j-\langle n_j\rangle)} \prod_i dn_i$$

where $S_{ij} = \langle (n_i - \langle n_i\rangle)(n_j - \langle n_j\rangle)\rangle$. Reference [5] provides an analytic expression for the resulting log-likelihood ratio distribution including a correlated error matrix; however, this equation was obtained with an integration over negative numbers of expected events and does not hold. Attempts to provide a closed form solution for the positive semi-definite region require analytical continuation of the error function over a wide range of the complex plane. Instead, a numerical integration over the positive semi-definite region has been adopted for our software packages.

5. DISCOVERY LUMINOSITY

Because the calculation of expected significance is technically very difficult at the LHC, other summaries of the discovery potential have been explored. While these techniques are not new, it is important to consider their pros and cons. One such alternate summary of the discovery potential is based on the discovery luminosity". Define the discovery luminosity, $L^*(m_H)$, to be the integrated luminosity necessary for the expected significance to reach 5σ. The discovery luminosity is an informative quantity; however, it must be interpreted with some care:

- Collecting an integrated luminosity equal to the nominal discovery luminosity does not guarantee that a discovery will be made. Instead, with $L^*(m_H)$ of data the median of ρ_{s+b} will be at the 5σ level – which corresponds to a 50% chance of discovery. See Section 6 for more details.

- In practice an analysis' cuts, systematic error, and signal and background efficiencies are luminosity-dependent quantities. When we calculate the discovery luminosity, we treat the analysis as constant.

6. THE POWER OF A 5σ TEST

The traditional quantity which is used to summarize an experiment's discovery potential is the combined significance; however, as was noted in Section 3 this plot becomes very difficult to make when the significance goes beyond about 8σ. Furthermore, the plot itself starts to loose relevance when the significance is far above 5σ. The discovery luminosity is another

[3] In principle, any distribution could be used within this framework.

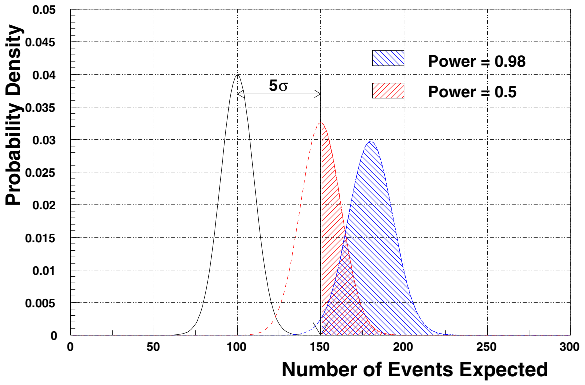


Figure 3: Examples of power for two different signal-plus-background hypotheses with respect to a single background-only hypothesis with 100 expected events (black).

possible way of illustrating an experiment's discovery potential, but it must be interpreted with some care. A third summary of an experiment's discovery potential which is related to the probability of Type II error: the *power*. First, it should be noted that the expected significance is a measure of separation between the *medians* of the background-only and signal-plus-background hypotheses. Thus, when we see the significance curve cross the 5σ line in Fig. 2 there is only a 50% chance that we would observe a 5σ effect if the Higgs does indeed exist at that mass. In practice, we claim a discovery if the observed data exceeds the 5σ critical region, and do not claim a discovery if it doesn't. The meaning of the 5σ discovery threshold is a convention which sets the probability of Type I error to be $2.85 \cdot 10^{-7}$. With that in mind, the idea that the significance is 20σ at $m_H = 160$ GeV is irrelevant. What is relevant is the probability that we will claim discovery of the Higgs if it is indeed there: that quantity is called the power. The power is defined as $1 - \beta$ where β is the probability of Type II error: the probability that we reject the signal-plus-background hypothesis when it is true [2].

Consider Figure 3 with a background expectation of 100 events. The black vertical arrow denotes the 5σ discovery threshold. The (red) dashed curve shows the distribution of the number of expected events for a signal-plus-background hypothesis with 150 events. Normally, we would say the expected significance is 5σ for this hypothesis; however, we can see that only 50% of the time we would actually claim discovery. The rightmost (blue) curve shows the distribution of the number of expected events for a signal-plus-background hypothesis with 180 events. Normally, we would say the expected significance is 8σ for this hypothesis; however, a more meaningful quantity – the power – is associated with the probability we would claim discovery which is about 98%. In addition to the power being a germane quantity, it is much easier to calculate.

7. CONCLUSION

In conclusion, the migration of the statistical toolset developed at LEP to the LHC environment is not as straightforward as one might expect. The first difficulties are computational and arise from the combination of channels with many events and channels with few events (these are easily solved). The next difficulties are numerical and arise from the extremely high expected significance of the high-energy frontier. These problems can be solved by brute force; or they can be reinterpreted as conceptual problems, and solved by asking different questions (i.e. power). Lastly, there is a philosophical split related to the Bayesian and Frequentist approach to uncertainty. At the LHC, the choice of the formalism is no longer a second-order effect, and this problem is not so easy to solve.

Acknowledgments

This work was supported by a graduate research fellowship from the National Science Foundation and US Department of Energy Grant DE-FG0295-ER40896.

References

[1] K.S. Cranmer *et. al.* Confidence level calculations for $H \to W^+W^- \to l^+l^-\not{p}_T$ for $115 < M_H < 130\,\text{GeV}$ using vector boson fusion. ATLAS communication ATL-COM-PHYS-2002-049 (2002).

[2] J.K Stuart, A. Ord and S. Arnold. *Kendall's Advanced Theory of Statistics, Vol 2A (6th Ed.).* Oxford University Press, New York, 1994.

[3] A.L. Read. Optimal statistical analysis of search results based on teh likelihood ratio and its application to the search for the MSM higgs boson at $\sqrt{s} = 161$ and 172 GeV. DELPHI note 97-158 PHYS 737 (1997).

[4] T. Junk. Confidence level computation for combining searches with small statistics. *Nucl. Instrum. Meth.*, A434:435–443, 1999.

[5] J. Nielsen H. Hu. Analytic confidence level calculations using the likelihood ratio and fourier transform. "Workshop on Confidence Limits", Eds. F. James, L. Lyons and Y. Perrin, CERN 2000-005 (2000), p. 109.

[6] C. Bauer *et. al.* Introduction to the GiNaC framework for symbolic computation within the c++ programing language. *J. Symbolic Computation*, 33:1–12, 2002.

[7] K.S. Cranmer. Frequentist hypothesis testing with background uncertainty., 2003. "PhyStat2003", SLAC. physics/0310108.

[8] R.D. Cousins and V.L. Highland. Incorporating systematic uncertainties into an upper limit. *Nucl. Instrum. Meth.*, A320:331–335, 1992.

A Multivariate Two-Sample Test Based on the Concept of Minimum Energy

G. Zech and B. Aslan
University of Siegen, 57072 Siegen, Germany

We introduce a new statistical quantity the *energy* to test whether two samples originate from the same distributions. The energy is a simple logarithmic function of the distances of the observations in the variate space. The distribution of the test statistic is determined by a resampling method. The power of the energy test in one dimension was studied for a variety of different test samples and compared to several nonparametric tests. In two and four dimensions a comparison was performed with the Friedman-Rafsky and nearest neighbor tests. The two-sample energy test is especially powerful in multidimensional applications.

1. INTRODUCTION

Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ and $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m$ be two samples of independent random vectors with distributions F and G, respectively. The classical two-sample problem then consists of testing the hypothesis

$$H_0 : F(\mathbf{x}) = G(\mathbf{x}), \text{ for every } \mathbf{x} \in \mathbb{R}^d,$$

against the general alternative

$$H_1 : F(\mathbf{x}) \neq G(\mathbf{x}), \text{ for at least one } \mathbf{x} \in \mathbb{R}^d,$$

where the distribution functions F and G are unknown.

Testing whether two samples, for example, two data sets taken at different times, are consistent with a single unknown distribution is a task that occurs in many areas of research. Clearly tests based on moments [1–3] are not sensitive to all alternatives H_1. Other tests require binning of data like the power-divergence statistic test [4] and tests of the χ^2 type. However, a high dimensional space is essentially empty, as is expressed in the literature by the term *curse of dimensionality* [5], hence tests based on binning are rather inefficient unless the sample sizes are large. Binning-free tests based on rank statistics are restricted to univariate distributions, and, when applied to the marginal distributions, they neglect correlations. The Friedman-Rafsky test [6] and the nearest neighbor test [7] avoid these caveats.

The *Friedman-Rafsky test* can be seen as a generalization of the univariate Wald-Wolfowitz run test [8]. The problem in generalizing the run test to more than one dimension is that there is no unique sorting scheme for the observations. The minimum spanning tree can be used for this purpose. It is a graph which connects all observations in such a way that the total Euclidean length of the connections is minimum. Closed cycles are inhibited. The minimum spanning tree of the pooled sample $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n, \mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m$ is formed. The test statistic R_{nm} equals the number of connections between observations from different samples. Small values of R_{nm} lead to a rejection of H_0. The statistic R_{nm} is asymptotically distribution-free under the null hypothesis [9].

The *nearest neighbor test* statistic N_{nm} is the sum of the number of observations of the pooled sample where the nearest neighbor is of the same type. In [7] it is shown that the limiting distribution of N_{nm} is normal in the limit $\min(n, m) \to \infty$. Large values of N_{nm} lead to rejection of H_0.

In this paper we propose a new test for the two-sample problem - the *energy test* - which shows high performance independent of the dimension of the variate space and which is easy to implement. Our test is related to Bowman-Foster test [10] but whereas this test is based on probability density estimation and local comparison, the energy test explores long range correlations.

2. THE TWO-SAMPLE *ENERGY* TEST

The basic idea behind using the quantity *energy* to test the compatibility of two samples is simple. We consider the sample $A : \mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ as a system of positive charges of charge $1/n$ each, and the second sample $B : \mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m$ as a system of negative charges of charge $-1/m$. The charges are normalized such that each sample contains a total charge of one unit. From electrostatics we know that in the limit of where n, m tend to infinity, the total potential energy of the combined samples computed for a potential following a one-over-distance law will be minimum if both charge samples have the same distribution. The energy test generalizes these conditions. For the two-sample test we use a logarithmic potential in $\mathbb{R}^d$. In Ref. [11] we show that also in this case, large values of energy indicate significant deviations between the parent populations of the two samples. The proof relies on the fact that the Fourier transform of the kernel function $1/r^\kappa$ is positive definite. The logarithmic function is equivalent to the inverse power function in the limit where the exponent tends to zero.

The test statistic Φ_{nm} consists of three terms, which correspond to the energies of samples A, B and the interaction energy of the two samples

$$\Phi_{nm} = \frac{1}{n^2}\sum_{i<j}^{n} R\left(|\mathbf{x}_i - \mathbf{x}_j|\right) + \\ + \frac{1}{m^2}\sum_{i<j}^{m} R\left(|\mathbf{y}_i - \mathbf{y}_j|\right) + \\ - \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} R\left(|\mathbf{x}_i - \mathbf{y}_j|\right)$$

where $R(r)$ is a continuous, monotonic decreasing function of the Euclidean distance r between the charges. The choice of R may be adjusted to a specific statistical problem. With the choice $R(r) = -\ln r$ the test is scale invariant and offers a good rejection power against many alternatives to the null hypothesis.

To compute the power of the new two-sample *energy* test we use the permutation method [12] to evaluate the distribution of Φ_{nm} under H_0. We merge the $N = m + n$ observations of both samples and draw from the combined sample a subsample of size n without replacement. The remaining m observations represent a second sample. The probability distribution under H_0 of Φ_{nm} is evaluated by determining the values of Φ_{nm} of all $\binom{N}{m} = \frac{N!}{n!m!}$ possible permutations. For large N this procedure can become computationally too laborious. Then the probability distribution is estimated from a random sample of all possible permutations.

We propose to normalize the vectors $\mathbf{z}_i$, $i = 1, 2, \ldots, N$ of the pooled sample to unit variance in all projections, $\overset{*}{z}_{ik} = (z_{ik} - \mu_k)/\sigma_k$, where μ_k, σ_k are mean value and standard deviation of the projection $z_{1k}, \ldots, z_{Nk}$ of the coordinates of the observations of the pooled sample. In this way we avoid situations in which a single projection dominates the value of the distance and consequently of the energy and that other projections contribute only marginally to it. We have not studied the effect of this scaling procedure which probably is sensible for all multidimensional goodness-of fit tests. In the following power comparison of our method with the competing methods, the different projections were not normalized.

3. POWER COMPARISONS

The performance of various tests were assessed for finite sample sizes by Monte Carlo simulations in d = 1, 2 and 4 dimensions. Also the critical values of all considered tests were calculated by Monte Carlo simulation. We chose a 5% significance level.

Table I Four dimensional distributions used to generate the samples.

case	P^X	P^Y
1	$N(\mathbf{0},\mathbf{I})$	$C(\mathbf{0},\mathbf{I})$
2	$N(\mathbf{0},\mathbf{I})$	$N_{\log}(\mathbf{0},\mathbf{I})$
3	$N(\mathbf{0},\mathbf{I})$	$80\%N(\mathbf{0},\mathbf{I})+20\%N\left(\mathbf{0},0.2^2\mathbf{I}\right)$
4	$N(\mathbf{0},\mathbf{I})$	$50\%N(\mathbf{0},\mathbf{I})+50\%N\left(\mathbf{0},\begin{pmatrix}1 & 0.4 & 0.5 & 0.7\\ 0.4 & 1 & 0.6 & 0.8\\ 0.5 & 0.6 & 1 & 0.9\\ 0.7 & 0.8 & 0.9 & 1\end{pmatrix}\right)$
5	$N(\mathbf{0},\mathbf{I})$	Student's t_2
6	$N(\mathbf{0},\mathbf{I})$	t_4
7	$U(\mathbf{0},\mathbf{1})$	$CJ(10)$
8	$U(\mathbf{0},\mathbf{1})$	$CJ(5)$
9	$U(\mathbf{0},\mathbf{1})$	$CJ(2)$
10	$U(\mathbf{0},\mathbf{1})$	$CJ(1)$
11	$U(\mathbf{0},\mathbf{1})$	$CJ(0.8)$
12	$U(\mathbf{0},\mathbf{1})$	$CJ(0.6)$
13	$U(\mathbf{0},\mathbf{1})$	$80\%U(\mathbf{0},\mathbf{1}) + 20\%N(\mathbf{0.5},0.05^2\mathbf{I})$
14	$U(\mathbf{0},\mathbf{1})$	$50\%U(\mathbf{0},\mathbf{1}) + 50\%N(\mathbf{0.5},0.2^2\mathbf{I})$

For the null hypothesis we determine the distribution of Φ_{nm} with the permutation technique, as mentioned above. We followed [12] and generated 1000 randomly selected two-sample subsets in each case and determined the critical values ϕ_c of ϕ_{nm}. For the specific case $n = m = 50$ and samples drawn from a uniform distribution we studied the statistical fluctuations. Transforming the confidence interval of ϕ_c into limits for nominal level $\alpha = 0.05$, we obtain the interval $[0.036, 0.063]$.

Even though the energy test has been designed for multivariate applications, we also investigated its power in one dimension because there a comparison with several well established tests is possible. To avoid a personal bias we drew the two samples from the same probability distributions which have been investigated by [13]. We compared the energy test to the Kolmogorov-Smirnov, Cramèr-von Mises, Wilcox, Lepage and χ^2 test. Details of the comparison are given in Ref. [11]. The results indicate that the power of the energy test in most of the cases is larger than that of the well known χ^2 and Kolmogorov-Smirnov test and comparable to that of the Cramèr-von Mises test.

In the multivariate case we compared the energy test with the Friedman-Rafsky and the nearest neighbor tests.

In order to investigate how the performance of the tests using Φ_{nm}, R_{nm} and N_{nm} changes with the dimension, we have investigated the power in dimensions $d = 2$ and 4. Since the results in both cases are similar, we present in this short writeup only those

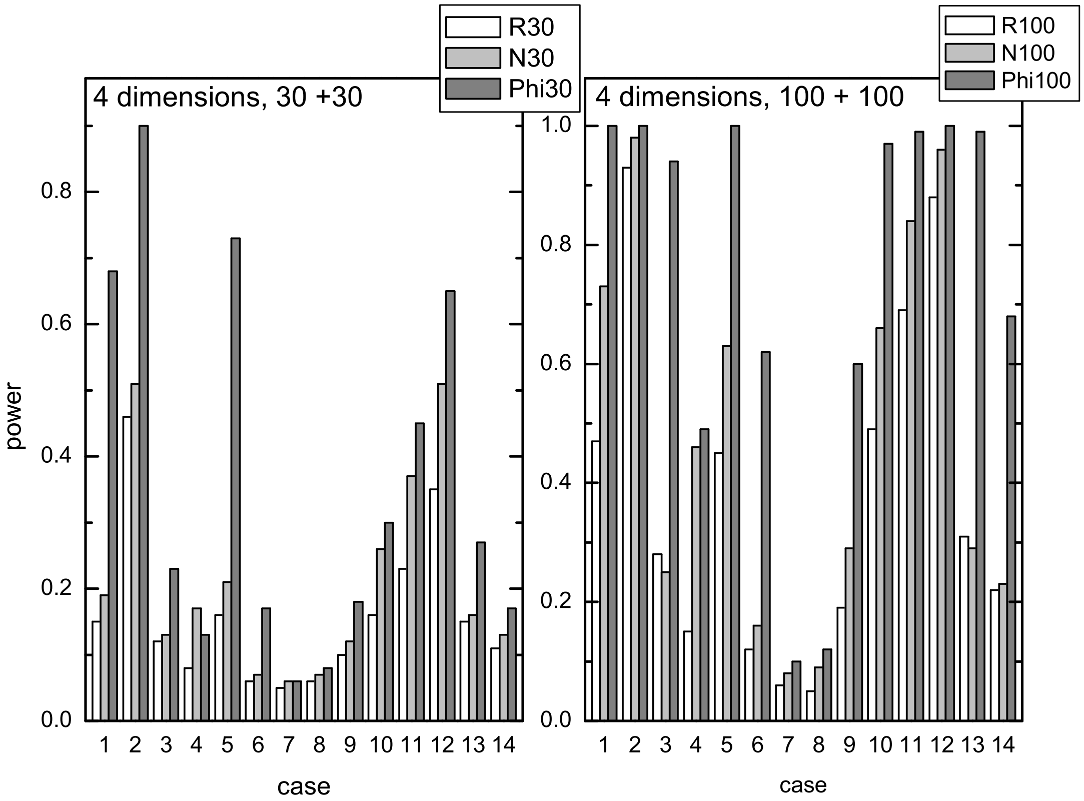


Figure 1: Rejection power of three two-sample tests for different alternatives. The sample sizes are 30 + 30 (left hand) and 100 +100 (right hand). R, N, Phi denote the Friedman-Rafsky test, the nearest neighbor test, and the energy test.

of the four dimensional case. In Table I we summarize the alternative probability distributions P^X and P^Y from which we drew the two samples. The first sample was drawn either from $N(\mathbf{0},\mathbf{I})$ or from $U(\mathbf{0},\mathbf{1})$ where $N(\mu,\mathbf{V})$ is a multivariate normal probability distribution with the indicated mean vector μ and covariance matrix $\mathbf{V}$ and $U(\mathbf{0},\mathbf{1})$ is the multivariate uniform probability distribution in the unit cube. The parent distributions of the second sample were the Cauchy distribution C, the $N_{\log}$ distribution (explained below), correlated normal distributions, the Student's distributions, t_2 and t_4, and Cook-Johnson $CJ(a)$ distributions [14] with correlation parameter $a > 0$. $CJ(a)$ converges for $a \to \infty$ to the independent multivariate uniform distribution and $a \to 0$ corresponds to the totally correlated case $X_{i1} = X_{i2} = ... = X_{id}, i = 1, ..., n$. We generated the random vectors from $CJ(a)$ via the standard gamma distribution with shape parameter a, following the prescription proposed by [15]. The distribution denoted by $N_{\log}$ is obtained by the variable transformation $x \to x' = \ln |x|$ applied to each coordinate of a multidimensional normal distribution and is not to be confused with the log-normal distribution. It is extremely asymmetric. Some of the considered probability densities have also been used in a power study in [16].

The various combinations emphasize different types of deviations between the populations. These include location and scale shifts, differences in skewness and kurtosis as well as differences in the correlation of the variates.

The test statistics Φ_{nm}, R_{nm} and N_{nm} were evaluated.

The power was again computed for 5% significance level and samples of equal size $n = m = 30$, 50, and 100 (small, moderate and large) in two and four dimensions. In Figure 1 we show some of the results. More details can be found in [11].

The Friedman-Rafsky and the nearest neighbor tests show very similar rejection power. For all three sample sizes and dimensions the energy test performed better than the other two tests in almost all considered alternatives. This is astonishing because the logarithmic distance function is long range and the probability distributions in the cases 11 and 12 have a sharp peak in one corner of a d dimensional unit cube and in case 13 a sharp peak in the middle of the unit cube. The multivariate student distribution represents very mild departures from normality, but nevertheless the rejection rate of the energy test is high.

Acknowledgments

We would like to thank the organizers of the conference for this beautiful conference. We acknowledge the financial support by the German Bundesministerium für Bildung und Forschung, Frderkennzeichen 05 HB1PSA/4.

References

[1] Duran, B. S., "A survey of nonparametric tests for scale", Communications in Statistics - Theory and Methods 5 (1976) 1287.

[2] Conover, W. J., Johnson, M. E., and Johnson, M. M., "A comparative study of tests of homogeneity of variances, with applications to the outer continental shelf bidding data", Technometrics 23 (1981) 351.

[3] Buening, H., "Robuste und adaptive Tests", Berlin: De Gruyter (1991).

[4] Read, T. R. C., and Cressie, N. A. C., "Goodness-of-fit statistics for discrete multivariate data", New York: Springer-Verlag (1988).

[5] Scott, D. W., "Multivariate Density Estimation: Theory, Practice and Visualisation", Wiley, New York (1992).

[6] Friedman, J. H., and Rafsky, L. C., "Multivariate generalizations of the Wolfowitz and Smirnov two-sample tests", Annals of Statistics 7 (1979) 697.

[7] Henze, N., "A multivariate two-sample test based on the number of nearest neighbor type coincidences", Annals of Statistics 16 (1988) 772.

[8] Wald, A., and Wolfowitz, J., "On a test whether two samples are from the same population", Ann. Math. Statist. 11 (1940) 147.

[9] Henze, N., and Penrose, M. D., "On the multivariate runs test", Annals of Statistics 27 (1998) 290.

[10] Bowman, A., and Foster, P., "Adaptive smoothing and density-based tests of multivariate normality", J. Amer. Statist. Assoc. 88 (1993) 529.

[11] Aslan, B. and Zech, G., "A new test for the multivariate two-sample problem based on the concept of minimum energy", available on http://arxiv.org/abs/math.PR/0309164 (2003).

[12] Efron, B., and Tibshirani, R., "An Introduction to the Bootstrap", New York: Chapman and Hall (1993).

[13] Buening, H., "Kolmogorov-Smirnov- and Cramèr-von Mises type two-sample tests with various weight functions", Communications in Statistics-Simulation and Computation 30 (2001) 847.

[14] Devroye, L., "Non-Uniform Random Variate Generation", New York: Springer-Verlag (1986).

[15] Ahrens, J. H., and Dieter, U., "Pseudo-Random Numbers", New York: Wiley (1977).

[16] Bahr, R., "A new test for the multi-dimensional two-sample problem under general alternative", Ph.D. Thesis, University of Hannover (1996) (in German).

Some Comments on χ^2 Minimisation Applications

V. Blobel
Institut für Experimentalphysik, Universität Hamburg, Germany

The determination of parameters in fits to measured data is a standard task of data analysis. The popular method called χ^2 minimization, as used in recent publications in a wide range of applications, is analysed and compared to standard statistical methods for parameter estimation, the method of least squares and the maximum likelihood method, which have certain optimal statistical properties.

1. INTRODUCTION

χ^2 minimisation. The determination of parameters in fits to measured data is a standard task of data analysis. The standard method of least squares is often referred to as χ^2 minimisation, which is a confusion in terminology; the minimum of the least squares sum follows often, but not always, the χ^2 distribution. In "χ^2 minimization" used in a wide range of applications from calorimeter calibration to complex analyses like fitting parton densities using data from different experiments ("*. . . to determine these parameters one must minimise a χ^2 which compares the measured values . . . to the calculated ones . . . .*") a variety of different non-standard concepts is used, often motivated by serious problems to handle the experimental data in a consistent way; these methods as used in recent applications may result in a bias of the fitted parameter values. Two examples showing common mistakes in the construction of χ^2-expressions are discussed below.

Calorimeter calibration. Calorimeters for energy measurements in a particle detector require a calibration, usually based on data taken with a fixed beam energy E. The measured data y_{jk} for calorimeter cell j in event k (total N events) have to be related to this known energy E. A method used in many experiments is based on the minimisation of the expression

$$\chi^2 = \frac{1}{N} \sum_{k=1}^{N} \left(a_1 y_{1,k} + a_2 y_{2,k} + \ldots + a_n y_{n,k} - E\right)^2$$

for the determination of the a_j, and this can produce biased results, as pointed out by D. Lincoln at al.[1]. To simplify the discussion single cell measurements y_k are assumed with standard deviation σ, with a mean value from N measurements of $\bar{y} = \sum_k y_k/N$; the intended result for the calibration factor is simply $a = E/\bar{y}$. The one-cell version of the above χ^2 definition

$$\chi^2 = \frac{1}{N} \sum_{k=1}^{N} \left(a \cdot y_k - E\right)^2$$

would produce the biased result $a = E \cdot \bar{y}/(\bar{y}^2 + \sigma^2) \neq E/\bar{y}$; the bias mimics a non-linear response of the calorimeter (a *known* bias can of course be corrected for). Using a fixed factor $1/\sigma^2$ instead of the unconventional $1/N$ would give the identical result. There would be no bias, if either a factor $1/(a \cdot \sigma)^2$ would be introduced, or if the inverse constant a_{inv} would have been determined from a modified χ^2 expression with $(y_k - a_{\text{inv}} \cdot E)$ instead of $(a \cdot y_k - E)$.

Normalisation errors. In several publications with χ^2 minimisation it is mentioned that normalisation errors can produce biased fit results ("*. . . that including normalisation errors in the correlation matrix will produce a fit which is biased towards smaller values . . .*"). This effect is described [2] for the data $y_1 = 8.0 \pm 2\%$ and $y_2 = 8.5 \pm 2\%$, with a common (relative) normalisation error of $\varepsilon = 10\%$. Assuming that the true values for both data values are identical the mean value $\bar{y}$ calculated by χ^2 minimisation in the paper is

$$\bar{y} = 7.87 \pm 0.81 \qquad \text{i.e. } \bar{y} < y_1 \text{ and } < y_2$$

– this is apparently wrong. This result has been obtained by minimising

$$\chi^2 = \boldsymbol{\Delta}^T \boldsymbol{V}^{-1} \boldsymbol{\Delta} \qquad \text{with} \quad \boldsymbol{\Delta} = \begin{pmatrix} y_1 - \bar{y} \\ y_2 - \bar{y} \end{pmatrix}$$

using the covariance matrix

$$\boldsymbol{V} = \begin{pmatrix} \sigma_1^2 + \varepsilon^2 y_1^2 & \varepsilon^2 y_1 y_2 \\ \varepsilon^2 y_1 y_2 & \sigma_2^2 + \varepsilon^2 y_2^2 \end{pmatrix} , \qquad (1)$$

which should include the common normalisation error.

The discussion in papers attributes the problem to the least squares method; however the origin of the wrong result is in fact the above definition (1) of the covariance matrix: the contribution to $\boldsymbol{V}$ from the normalization error was calculated from the *measured* values, which were different; the result is a covariance ellipse with axis different from 45° and this produces a biased mean value, as can be seen in Figure 1. According to the assumption both true data values are identical and then the normalisation error contribution has to be $\varepsilon^2 \bar{y}^2$ for all elements, and the correct mean value is obtained (axis of the covariance ellipse at 45°). Another method leading to the correct result is the introduction of the normalisation factor α as an additional measured value and using the χ^2 definition

$$\chi^2 = \sum_k \frac{(y_k - \alpha \cdot \bar{y})^2}{\sigma_k^2} + \frac{(\alpha - 1)^2}{\varepsilon^2} .$$

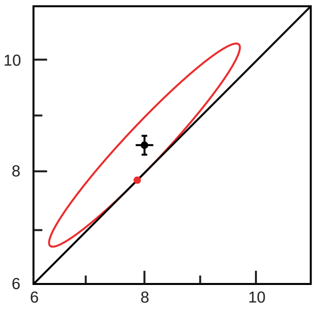


Figure 1: Measured point $(y_1, y_2) = (8.0, 8.5)$ and covariance ellipse according to the definition in equation (1). The slightly tilted ellipse touches the diagonal $y_1 = y_2$ at a point which is below both data points.

2. STANDARD METHODS

Least Squares. Doubts are raised in publications with χ^2 minimisation about the applicability of the method ("*the justification for using least squares lies in the assumption that the measurement errors are Gaussian distributed . . . it is doubtful that Gaussian errors are realistic.*"). Arbitrary factors are often applied to increase the parameter errors (see section 3).

For the standard least squares method the properties of the result can be derived from certain conditions. The linear least squares problem is denoted by $\boldsymbol{A}\boldsymbol{a} \cong \boldsymbol{y}$. Given a $n \times p$ matrix $\boldsymbol{A}$ and given a n vector $\boldsymbol{y}$ with covariance matrix $\boldsymbol{V}_y$ the problem is to find the p vector $\boldsymbol{a}$ of parameters which minimises $(\boldsymbol{W} = \boldsymbol{V}_y^{-1})$

$$S(\boldsymbol{a}) = (\boldsymbol{y} - \boldsymbol{A}\boldsymbol{a})^T \; \boldsymbol{W} \; (\boldsymbol{y} - \boldsymbol{A}\boldsymbol{a})$$

with respect to $\boldsymbol{a}$. The solution (from $\partial S/\partial \boldsymbol{a} = 0$) is a linear transformation of the data vector $\boldsymbol{y}$

$$\hat{\boldsymbol{a}} = \left[\left(\boldsymbol{A}^T\boldsymbol{W}\boldsymbol{A}\right)^{-1} \; \boldsymbol{A}^T\boldsymbol{W}\right] \boldsymbol{y} \qquad = \boldsymbol{B}\,\boldsymbol{y}\;,$$

the covariance matrix of vector $\hat{\boldsymbol{a}}$ is given by standard error propagation:

$$\boldsymbol{V}_a = \boldsymbol{B}\,\boldsymbol{V}_y\,\boldsymbol{B}^T = \left(\boldsymbol{A}^T\boldsymbol{W}\boldsymbol{A}\right)^{-1} \tag{2}$$

and this relation does not depend on the "*quality of the fit*". Properties of the solution are derived under certain conditions: the data unbiased, i.e. $E[\boldsymbol{y}] = \boldsymbol{A}\,\bar{\boldsymbol{a}}$ ($\bar{\boldsymbol{a}}$ = true parameter vector), and the covariance matrix $\boldsymbol{V}_y$ of the data known (and correct). Distribution-free properties of least squares estimates in linear problems are: the estimated parameters are unbiased, and in the class of unbiased estimates, which are linear in the data, the least squares estimates $\hat{\boldsymbol{a}}$ have the smallest variance (Gauß-Markoff theorem). The expectation of the sum of squares of the residuals is $\hat{S} = (n - p)$. However the distribution of $\hat{S}$ follows the χ^2 distribution with $(n - p)$ degrees of freedom *only* in the case of Gaussian distributed data. For non-linear problems the above properties are only approximately valid.

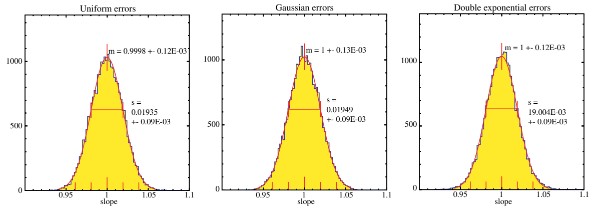


Figure 2: Distribution of the slope parameter in MC simulated least squares fits of straight lines. The data distributions were uniform (left), Gaussian (center) and double-exponential (right). The width of the parameter distributions are as expected from equation (2).

The Figure 2 shows the distribution of the slope parameter in 25 000 MC simulations of straight-line fits ($n = 20$ data points) with different data distributions: the uniform distribution (left), the Gaussian distribution (center) and the double-exponential distribution(right). In all these cases (*same* standard deviation of the data) a Gaussian distribution of the slope parameter (and the intercept) is observed (central limit theorem!), although the input data distribution are different and especially have very different tails. In addition the mean value of $\hat{S}$ is $(n-p) = 20-2$ in all three cases, but the distribution of the corresponding P-values (calculated from the observed χ^2 and $(n - p)$) is uniform only in the case of Gaussian-distributed data, as expected.

Likelihood function and Information. The maximum likelihood method can be used, if the details about the distribution of the data are known and the likelihood function $\mathcal{L}(\boldsymbol{a})$ of the problem can be constructed. In the case of several parameters $a_1, a_2 \ldots a_p$ a $p \times p$ symmetric information matrix $\boldsymbol{I}$ with elements determined by the expectation values

$$I_{jk} = E\left[\frac{\partial \ln \mathcal{L}}{\partial a_j}\frac{\partial \ln \mathcal{L}}{\partial a_k}\right] = -E\left[\frac{\partial^2 \ln \mathcal{L}}{\partial a_j \partial a_k}\right]$$

is defined, and it can be shown that the minimal variance $\boldsymbol{V}_{\hat{\boldsymbol{a}}}$ of an estimate $\hat{\boldsymbol{a}}$ is given by the inverse of the information matrix: $\boldsymbol{V}_{\hat{\boldsymbol{a}}} = \boldsymbol{I}^{-1}$. In practice the negative log likelihood function is defined as the objective function $F(\boldsymbol{a}) = -\ln \mathcal{L}(\boldsymbol{a})$ and the minimum w.r.t. $\boldsymbol{a}$ is determined by the condition $\boldsymbol{g} = \partial F/\partial a_j = 0$. In case of good statistic the Hessian $\boldsymbol{H}$, the matrix of second derivatives of $F(\boldsymbol{a})$, is almost constant in the region around the minimum and is a good estimate for the information matrix $\boldsymbol{I}$; the inverse $\boldsymbol{H}^{-1}$ is a good estimate of the covariance matrix $\boldsymbol{V}_{\hat{\boldsymbol{a}}}$ of the parameters $\hat{\boldsymbol{a}}$. This corresponds (like eq. (2)) to standard error propagation from the data errors to the parameter errors, especially there is no freedom to introduce additional factors. Objective functions from the maximum likelihood and the least squares method can be combined (e.g. $F(\boldsymbol{a}) + 1/2\, S(\boldsymbol{a})$).

The minimisation and the calculation of the covariance matrix require the inversion of the Hessian. The introduction of redundant parameters in χ^2 minimisation ("*. . . we found our input parameterisation was sufficiently flexible to accommodate data and indeed there is a certain redundancy evident . . .*") in a fit problem can be rather dangerous, because with redundant parameters the Hessian is singular and without special techniques neither the minimum of the objective function can be found nor the inverse can be calculated.

3. DATA AND PARAMETER ERRORS

Systematic errors. The statistical and systematic uncertainties of the data can only be correctly taken into account in a fit if there is a clear model describing all aspects of the uncertainties. For statistical errors the strategy is usually well-defined: they can be described by the standard deviations (uncorrelated data points), or by a covariance matrix. The latter is necessary if e.g. corrections for finite resolution are applied to the data. For contributions due to systematic errors there are two alternative models: multiplicative effects due to normalisations errors and additive effects due to offset errors have to be treated differently.

In general the normalisation uncertainty is given by a relative error, and in fits with data from more than one experiment the treatment of the normalisation error may be important. Instead of adding a contribution to the covariance matrix one additional parameter α can be introduced for each experiment (measured value $\alpha = 1 \pm \varepsilon$) and the expectation $f(x_i, \boldsymbol{a})$ for y_i is modified to $\alpha \cdot f(x_i, \boldsymbol{a})$, leaving the measured data point y_i unchanged:

$$S(\boldsymbol{a}) = \sum_i \frac{(y_i - \alpha \cdot f(x_i, \boldsymbol{a}))^2}{\sigma_i^2} \; + \; \Delta S^{\text{norm}}$$

with $\Delta S^{\text{norm}} = (\alpha - 1)^2 / \varepsilon^2$. The normalisation factor determined in an experiment is more the *product* than the sum of random variables. According to the *multiplicative* central limit theorem the product of positive random variables follows the log-normal distribution, i.e. the logarithm of the normalisation factor follows the normal distribution. For a log-normal distribution of a random variable α with $E[\alpha] = 1$ and standard deviation of ε the contribution to $S(\boldsymbol{a}, \alpha)$ is (from the likelihood function)

$$\Delta S^{\text{norm}} = \ln \alpha \left(3 + \frac{\ln \alpha}{\ln (1 + \varepsilon^2)} \right)$$

for each α, which reduces to the previous term for small deviations of the value α from 1.

An example for an additive error is the error of a calorimeter constant – a change of the constant will change *all* data values y_i, because events are moved between bins. Here one has to determine shifts s_i of data values y_i, for a one-standard deviation change of the calorimeter constant; the shifts s_i will carry a relative sign. The error could be taken into account by adding the rank=1 matrix $\boldsymbol{s}\boldsymbol{s}^T$ to the covariance matrix. Alternatively one additional parameter β can be introduced for each error contribution (measured value $\beta = 0 \pm 1$), and the expectation can be modified to $f(x_i, \boldsymbol{a}) + \beta \cdot s_i$:

$$S(\boldsymbol{a}) = \sum_i \frac{(y_i - (f(x_i, \boldsymbol{a}) + \beta s_i))^2}{\sigma_i^2} \; + \; \beta^2 \; .$$

The introduction of additional parameters α and β allow to see the effect of the systematic errors in the fit, including the correlation of the parameter to other parameters in the fit, and of the pull, which has an expected mean of zero and variance of 1. The pull is the ratio of the shift of a parameter value in the fit, divided by the standard deviation of the *shift* (not the original standard deviation). In the case of a parameter β introduced above the pull is $\hat{\beta}/\sqrt{1 - V_{\beta\beta}}$. Rather useful for checks of parameter correlations is the global correlation coefficient ρ_k,

$$\rho_k = \sqrt{1 - \frac{1}{(\boldsymbol{V})_{kk} \cdot \left(\boldsymbol{V}^{-1}\right)_{kk}}} \; ,$$

which is a measure of the total amount of correlation between the k-th parameter and *all* the other variables. It is the largest correlation between the k-th parameter and every possible linear combination of all the other variables.

Different expressions are used in publications using χ^2 minimisation. One example is called the offset method, where systematic errors are ignored in the fit ("*. . . forces the theory prediction to be as close as possible to the data . . .*"), but later added in quadrature in the error calculation. It is clear that the fit result must be biased, if incomplete error information is used.

Parameter errors. With the given definition of the fit expression and the error contributions there is no freedom in standard methods in the calculation of the parameter errors; they are the result of error propagation and the parameter covariance matrix is the inverse of the Hessian $\boldsymbol{H}$.

In publications using χ^2 minimisation the following statements are found: "*Notice that the covariance matrix*

$$V_{ij}^p = \langle \Delta_i \Delta_j \rangle = \Delta\chi^2 \cdot H_{ij}^{-1}$$

depends on the choice of $\Delta\chi^2$ which usually, but not always, is taken to be $\Delta\chi^2 = 1$. This choice

Table I The values of the parameter $\alpha_S(M_Z^2)$, obtained in different structure function analyses and the value of $\Delta\chi^2$, used in the error calculation. The column marked # gives the number of experiments used in the analysis.

Group	$\Delta\chi^2$	Ref.	#	Value of $\alpha_S(M_Z^2)$
H1	1	[4]	2	0.115 ± 0.0017 (exp) $^{+0.0009}_{-0.0005}$ (model) ± 0.005 (theory)
GKK	1	[5]	3	0.112 ± 0.001 (exp)
MRST02	20	[6]	many	0.1195 ± 0.002 (exp) ± 0.003 (theory)
ZEUS	50	[7]	several	0.1166 ± 0.0049 (exp) ± 0.0018 (model)
CTEQ6	100	[8]	several	0.1165 ± 0.0065 (exp)

. . . corresponds to the definition of the width of a Gaussian distribution." [9] "*Ideally $\Delta\chi^2 = 1$, but unrealistic.*" " *. . . and $\Delta\chi^2$ is the allowed deterioration in fit quality for the error determination.*" [6]. The freedom taken in this unconventional error definition is shown in the overview of table I, which shows the tendency to use a value $\Delta\chi^2$ larger than 1 in analyses where a large number of different experiments is combined. A value of $\Delta\chi^2 = 100$ is equivalent to multiplying all input errors by a factor of 10, and this procedure is not justified by the "χ^2-value" of the fit alone. In one of the cases this value is 2328 for 2097-15 = 2082 degrees of freedom; in standard methods all data errors would be increased by $\sqrt{\chi^2/n_{\mathrm{df}}} = 1.06$. The large, artificial and arbitrary magnification of errors points to severe problem of the whole data analysis.

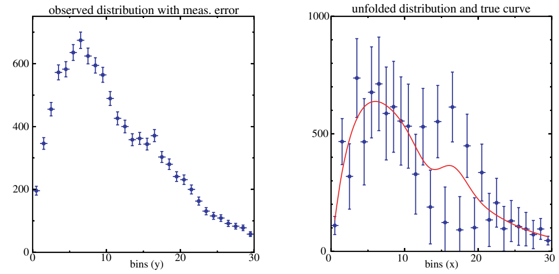


Figure 3: The measured distribution (left) and the result of unfolding by matrix inversion (right), showing large errors due to the negative correlations between adjacent data points.

4. STATISTICAL DATA PROPERTIES

Where is the origin of the problems apparent in the combined analysis of many experiments (see table I)? "*Indeed, we have always believed the theory, rather than experiment, will provide the dominant source of error.*" [6] However the origin seems to be on both the theoretical and the experimental side. Recent publications from experiments contain a lot of information on various types of errors, but this may be still insufficient for global fits.

The finite resolution in the measurement of kinematical variables requires in principle an unfolding procedure. Instead of measuring the true distribution $f(x)$ of a kinematical quantity x the distribution $g(y)$ of a quantity y is measured, which is related to the true variable x by a resolution function $A(y, x)$, known only implicitly by a sample of MC generated events:

$$g(y) = \int_\Omega A(y,x) f(x) \mathrm{d}x \qquad \text{or short } \boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} \ ,$$

The "correction factor"-method used by most of the experiments seems to introduce a hidden positive correlation between the data points; the method is usually not described by giving mathematical formulas, but in words like the following text taken from an early structure function measurement: " *. . . The main problem of the analysis is the correction for measurement errors (unsmearing corrections), which are large at large x where the structure functions vary rapidly with x. We proceed by assuming a true structure function and calculate by Monte Carlo simulation, on the basis of the known experimental resolution functions, the result to be expected in the apparatus. By iteration a true distribution which reproduces the experimental result is found. The unsmearing factor is the ratio of Monte Carlo events for any particular (x, Q^2) bin in the true distribution divided by those in the resolution smeared distribution. If this factor differs from unity by more than 30 %, the bin is not retained . . .*". [3] Often the result from a fit to a previous measurement is taken for the MC simulation, which may introduce a bias and this is certainly not a *blind analysis*. The unavoidable correlation between corrected data points is usually neglected, thus giving a too large weight to the experimental data in a later fit.

The method quoted above is usually applied because attempts to solve the problem by standard methods fail. This is illustrated in the simulated data of Figure 3. Figure 3a shows the 30-bins histogram of a distribution measured using 10 000 events, assuming a migration parameter $\varepsilon = 0.24$, which is the probability for the migration into both adjacent bins (the probability of measuring the entry in the correct bin is $(1-2\varepsilon)$). Using the symmetric migration matrix $\boldsymbol{A}$ the unfolded result $\boldsymbol{x}$ can be obtained by the solution of the equation $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$; the result in Figure 3b shows large fluctuations of the unfolded distribution.

An improved solution, derived from the properties

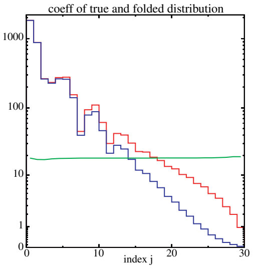


Figure 4: Absolute values of the elements of the transformed vectors $\boldsymbol{b} = \boldsymbol{U}^T\boldsymbol{x}$ and $\boldsymbol{c} = \boldsymbol{U}^T\boldsymbol{y}$ (without measurement errors) (left) and of the vector $\boldsymbol{c} = \boldsymbol{U}^T\boldsymbol{y}$ with measurement errors (right). The error level is shown as a line in both figures.

of the resolution matrix $\boldsymbol{A}$ alone, is based on the orthogonal decomposition $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^T$ of the migration matrix $\boldsymbol{A}$. Multiplying the original matrix equation

$$\boldsymbol{y} \cong \boldsymbol{A}\boldsymbol{x} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^T\boldsymbol{x}$$

by $\boldsymbol{U}^T$ from the left, one obtains ($\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{1}$)

$$\boldsymbol{c} = \boldsymbol{U}^T\boldsymbol{y} \cong \boldsymbol{D}\left(\boldsymbol{U}^T\boldsymbol{x}\right) = \boldsymbol{D}\boldsymbol{b} \qquad \boldsymbol{b} = \boldsymbol{D}^{-1}\boldsymbol{c}\,.$$

The transformed measurement vector $\boldsymbol{c} = \boldsymbol{U}^T\boldsymbol{y}$ allows one to calculate, by $\boldsymbol{b} = \boldsymbol{D}^{-1}\,\boldsymbol{c}$, the elements of the transformed result vector $\boldsymbol{b} = \boldsymbol{U}^T\boldsymbol{x}$ element by element, because the matrix $\boldsymbol{D}$ is diagonal (the diagonal elements are the eigenvalues of the symmetric matrix $\boldsymbol{A}$). Figure 4a shows the spectrum of elements for the true distribution before and after folding with the resolution matrix; the first elements are almost not affected by the resolution, but the *smoothing* effect of folding is clearly visible in the second half of the elements in a reduction of the size. The horizontal line shows the 1-standard deviation level of the measurement with the given number of events. It is clear that the second half of the elements can not be measured. The actual measurement is shown in Figure 4b. Only the first half of the elements represent a real measurement; for the second half of the elements the measured values corresponds to the average error level.

Taking the first 15 elements only and transforming the vector $\boldsymbol{b}$ back to the original bins by $\boldsymbol{x} = \boldsymbol{U}\boldsymbol{b}$ the result of Figure 5a with 30 bins is obtained, which is close to the curve representing the original dependence. Since this rather smooth result has been obtained from 15 measured elements, the rank of the covariance matrix can only be 15, and the correlations between adjacent bins are positive and large, like in the "correction factor"-method. A more acceptable result shown in Figure 5b is obtained by averaging each two (positively correlated) neighbor bins; the standard deviations of the data points are almost unchanged, but now the covariance matrix is of full-rank with small correlations. The severe problems observed in globals fits may be caused by retaining far too many data points with hidden large positive correlations from the "correction factor"-method, which appear to be more precise than they are.

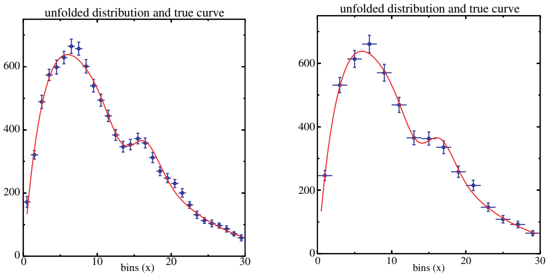


Figure 5: Result of unfolding with a cut-off after 15 elements. The left figure shows all 30 data points, which have a singular (rank-15) covariance matrix. The right figure show the result after combining pairs of two data points to one point; the covariance matrix has full rank.

References

[1] D. Lincoln, G. Morrow and P. Kaspar, A hidden bias in a common calorimeter calibration scheme, NIM **A 345**, 449 (1994)

[2] G. D'Agostini, On the use of the covariance matrix to fit correlated data, NIM **A 346**, 306 (1994)

[3] J. G. H. de Groot et al., Inclusive interactions of high-energy neutrinos and antineutrinos in iron, Zeitschrift für Physik **C** 1, 143 (1979)

[4] C. Adloff et al. (H1 Collaboration), Deep-inelastic inclusive ep-scattering at low x and a determination of α_s, Eur. Phys. J. C 21, 33 (2001)

[5] W. T. Giele and S. Keller, Implications of hadron collider observables on parton distributtion function uncertainties, Phys. Rev. D **58**, 094023 (1998), `hep-ph/9803393`

[6] A.D. Martin, R.G. Roberts, W.J. Stirling and R.S. Thorne, Uncertainties of predictions from parton distributions I: Experimental errors, Cavendish-HEP-2002/10, `hep-ph/0211080` (2002)

[7] S. Chekanov et al. (ZEUS Collaboration), ZEUS next-to-leading-order QCD analysis of data on deep-inelastic scattering, Phys. Rev. D **67**, 012007 (2003)

[8] J. Pumplin et al., New generation of parton distributions with uncertainties from global QCD analysis, JHEP 0207:012 (2002)

[9] M. Botje, Error Estimates on Parton Density Distributions, NIKHEF-01-014, `hep-ph/0110123` (2001)

A Software Toolkit for Statistical Data Analysis

G.A.P. Cirrone, S. Donadio, S. Guatelli, L. Lista, A. Mantero, B. Mascialino, S. Parlati, M.G. Pia
INFN
A. Pfeiffer, A. Ribon
CERN
P. Viarengo
IST, Genova, Italy

We present a project in progress to develop a software toolkit for statistical data analysis. The toolkit is based on advanced software technologies, integrating generic programming techniques with object oriented methods, and adopts a rigorous software process, to ensure a high quality of the product. Thanks to the component-based architecture and the usage of the standard **AIDA** interfaces, this tool can be easily used by other data analysis systems or integrated in experimental frameworks. The initial component of the system addresses goodness of fit tests; its applications include the comparisons of data distributions in a variety of use cases typical of HEP experiments: regression testing (in various phases of the software life-cycle), validation of simulation through comparison to experimental data, comparison of expected versus reconstructed distributions, comparison of different experimental distributions - or of experimental with respect to theoretical ones - in physics analysis, monitoring detector behavior with respect to a reference in online DAQ. The system will provide the user the option to choose among a wide set of goodness-of-fit tests (chi-squared, Kolmogorov-Smirnov, Anderson-Darling, Lilliefors, Kuiper, Cramer-von Mises, etc.), specialised for various types of binned and unbinned distributions. Its flexible design makes it open to further extension to implement other tests. This system would represent a significant improvement with respect to the current availability of comparison tests in HEP libraries, limited to the chi-squared and Kolmogorov-Smirnov algorithms. We present the architecture of the toolkit, the detailed design of the basic statistical testing component and preliminary results of its application, in particular concerning the physics validation of the Geant4 Simulation Toolkit. We discuss the openness of the project, welcoming contributions from experts and user requirements from experiments.

1. INTRODUCTION

Statistical methods play a significant role throughout the life-cycle of HEP experiments, being an essential component of physics analysis. In spite of this, only a few basic tools for statistical analysis were available in the public domain FORTRAN libraries for HEP. Nowadays the situation is unchanged even among the libraries of the new generation. The aim of this project is to build an open-source, up-to-date and sophisticated object-oriented statistical toolkit for HEP data analysis.

In this paper we will focus our attention on a specific component of the statistical toolkit, that is made-up by a collection of Goodness-of-Fit (**GoF**) [1] tests. Its aim is to provide a wide set of algorithms in order to test whether the distributions of two variables are compatible.

2. THE GOODNESS OF FIT STATISTICAL TOOLKIT

The applications of statistical comparisons of distributions in HEP are manyfold: regression testing (in various phases of the software life-cycle), validation of simulation through comparison to experimental data, comparison of different experimental distributions - or of experimental with respect to theoretical ones - in physics analysis, monitoring detector behavior with respect to a reference in online DAQ. From a mere statistical point of view, the problem consists in testing the non-parametric null hypothesis

$$\mathbf{H_0 : F = G}$$

against an alternative one

$$\mathbf{H_1 : F \neq G \quad or \quad F < G \quad or \quad F > G}.$$

Of course, in this kind of tests the acceptance of the null hypothesis $\mathbf{H_0}$ means that the researcher will be able to specify the distribution analyzed.

2.1. GoF statistical features

With the purpose of quantifying the measure of the deviation between the two distributions, many software toolkits for HEP data analysis solve the problem by means of the well known and wide-spread chi-squared test. This test is studied to describe discrete distributions, but it can be useful also in case of unbinned distributions. In this case the researcher is compelled to group data into classes, sacrificing in this way a good deal of the information conveyed by the distribution itself. In spite of the fact that this test has a general applicability, it must be noticed that the chi-squared asymptotic distribution is *not* valid if the theoretical frequencies involved in the computation are lower that 5. For these reasons, a powerful and up-dated statistical toolkit for HEP data analysis should supplement the chi-squared test with other statistical tests, involving individual sample values.

In order to compare unbinned distributions, the **GoF** toolkit includes a wide set of tests dealing with Kolmogorov's empirical distribution function (EDF). Using this toolkit the user is able to compare two EDFs selecting tests based on the supremum statistics:

- Kolmogorov-Smirnov test [2],
- Goodman approximation of Kolmogorov-Smirnov test [3],
- Kuiper test [4],

and together with tests based on the measure of integrated deviations of the two EDFs, multiplied by a weighting function:

- Cramer-von Mises test [5] [6],
- Anderson-Darling test [7].

Due to its mathematical formulation the Anderson-Darling test is favourable in case of fat-tailed distributions. A recent paper by Aksenov and Savageau [8] states that this last test statistic is suitable in case of any kind of distribution, independently on its particular skewness.

For these features, the **GoF** toolkit contains the generalization of these tests containing a weighting function to the case of binned distributions:

- Fisz-Cramer-von Mises test [9],
- k-sample Anderson-Darling test [10].

Dealing with a non-parametrical set of tests a *proper* evaluation about the power of these tests cannot be made. In general, the chi-squared test, for its simplicity, is the least powerful one because of information loss due to data grouping (binning). On the other hand, all the tests based on the supremum statistics are more powerful than the chi-squared one, focusing only on the maximum deviation between the two EDFs. The most powerful tests are undoubtedly the ones containing a weighting function, as the comparison is made all along the range of x, rather than looking for a marked difference at one point [11].

2.2. GoF toolkit architecture

The system has been developed following a rigorous software process (*United Software Development Process*), mapped onto the **ISO 15504** guidelines. With the aim of guaranteeing the quality of the product, the software development follows a spiral approach and the software life cycle is iterative-incremental, based on a User Requirements Document and providing Traceability.

The project adopts a solid architectural approach in order to offer the functionality and the quality needed by the user, to be maintainable over a large time scale and to be extensible, accommodating in this way future evolutions of the user requirements.

Both object-oriented techniques and generic programming allow a component-based design of the toolkit. This feature is very important as it facilitates the re-use of the toolkit as well as its integration in other data analysis frameworks.

Figure 1 represents the core components of the **GoF** toolkit. Its main features are summarized in two points:

- ○ the toolkit distinguishes input distributions on the basis of their type, as binned and unbinned data must be treated in different ways from a statistical point of view,
- ○ the whole comparison process is managed by one object (*ComparatorEngine*), which is templated on the distribution type and on the algorithm selected by the user.

The comparison returns to the user a statistics comparison result *object*, giving access to the computed value of the test statistics, the number of degrees of freedom and the quality of the comparison (p-value).

Figure 2 details all the algorithm implemented up to now: every algorithm is specialized for *only one* kind of distribution (binned or unbinned). In this way the user can access only those algorithms whose applicability conditions fit the kind of distribution he deals with.

The component-based design allow for an easy extension of the **GoF** toolkit to new algorithms without interfering with the existing code, employing the *Factory method* [12].

From the user's point of view, the object-oriented techniques adopted together with the standard **AIDA**(*Abstract Interfaces for Data Analysis*) [13] interfaces are able to shield the user from the complexity of both the architecture of the core components and the computational aspects of the mathematical algorithms implemented. All the user has to do is to choose the most appropriate algorithm (in practice writing one line of code) and to run the comparison. This implies that the user does not need to know statistical details of any algorithm, he also does not have to know the exact mathematical formulation of the distance nor of the asymptotic probability distribution he is computing. Therefore the user can concentrate on the choice of the algorithm relevant for his data. As an example, if the user tries to apply the Kolmogorov-Smirnov comparison to binned data, the **GoF** will not run the comparison, as the class *KolmogorovSmirnov-ComparisonAlgorithm* is defined to work only on unbinned distributions.

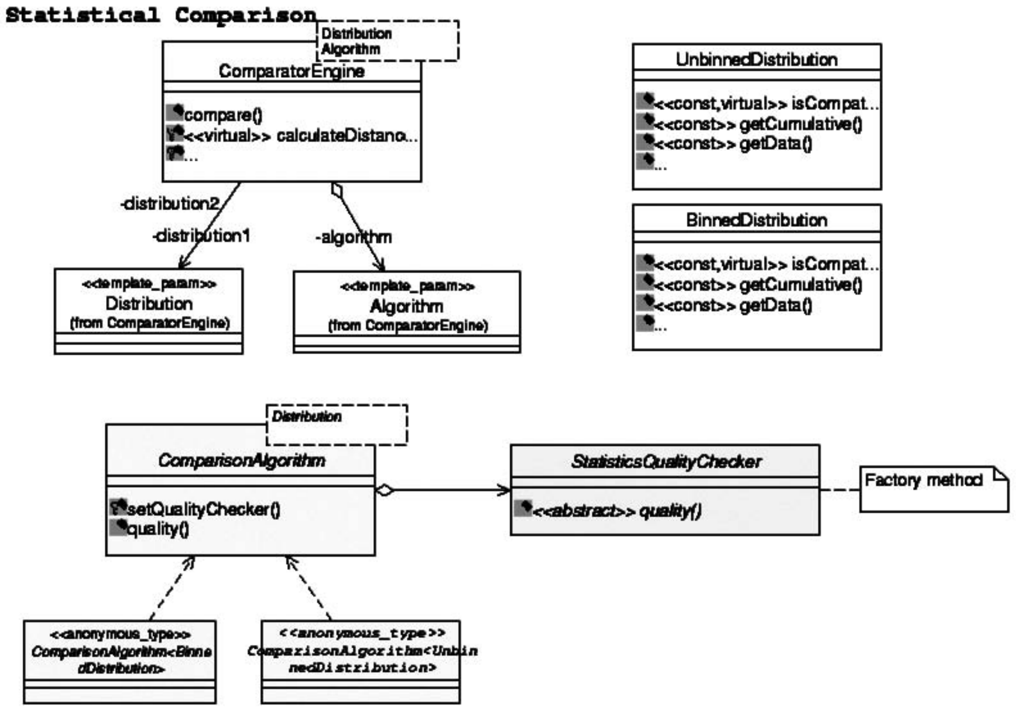


Figure 1: Statistical toolkit core design: one object (*Comparator Engine*) is responsible of the whole statistical comparison process.

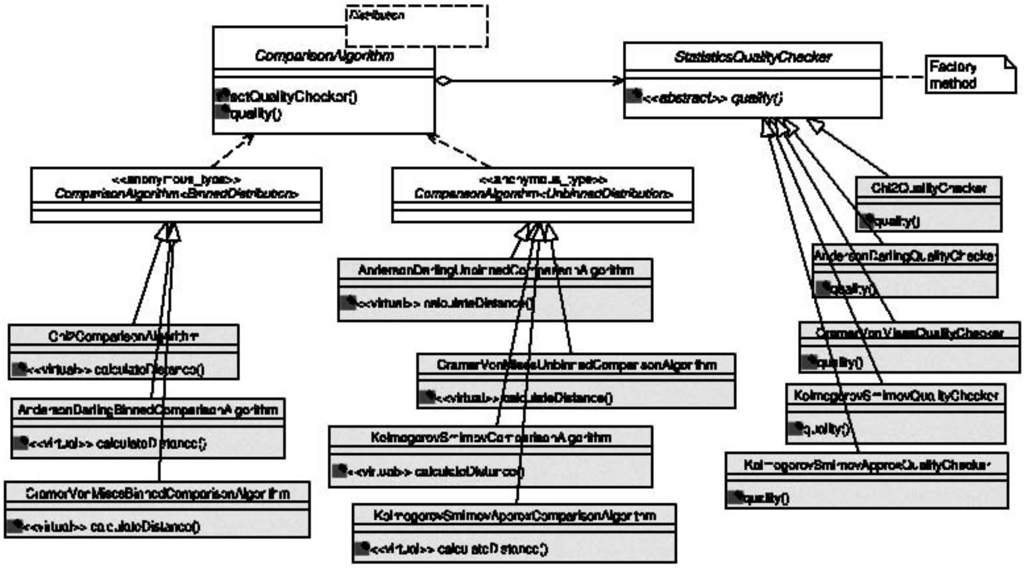

Distribution
ComparisonAlgorithm
setQualityChecker()
quality()
StatisticsQualityChecker
<<abstract>> quality()
Factory method
<<anonymous_type>>
ComparisonAlgorithm<BinnedDistribution>
<<anonymous_type>>
ComparisonAlgorithm<UnbinnedDistribution>
Chi2QualityChecker
quality()
AndersonDarlingQualityChecker
quality()
CramerVonMisesQualityChecker
quality()
KolmogorovSmirnovQualityChecker
quality()
KolmogorovSmirnovApproxQualityChecker
quality()
AndersonDarlingUnbinnedComparisonAlgorithm
<<virtual>> calculateDistance()
CramerVonMisesUnbinnedComparisonAlgorithm
<<virtual>> calculateDistance()
KolmogorovSmirnovComparisonAlgorithm
<<virtual>> calculateDistance()
KolmogorovSmirnovApproxComparisonAlgorithm
<<virtual>> calculateDistance()
Chi2ComparisonAlgorithm
<<virtual>> calculateDistance()
AndersonDarlingBinnedComparisonAlgorithm
<<virtual>> calculateDistance()
CramerVonMisesBinnedComparisonAlgorithm
<<virtual>> calculateDistance()

Figure 2: Detail of the statistical toolkit design: algorithms implemented for binned (Chi-squared, Fisz-Cramer-von Mises and k-sample Anderson-Darling tests) and unbinned (Kolmogorov-Smirnov, Goodman-Kolmogorov-Smirnov, Cramer-von Mises and Anderson-Darling tests) distributions.

3. EXAMPLES OF PRACTICAL APPLICATIONS OF THE GoF TOOLKIT

Thanks to the great variety of its sophisticated and powerful statistical tests, the **GoF** toolkit has been adopted by some projects, having as a crucial point the comparison of distributions of specific physical quantities. The three examples that follow have as a common denominator the essential need for an accurate validation of the simulations versus experimental data-sets. The field of applications are the following:

1. **Physics validation:** GEANT4 [14] decided to adopt the **GoF** toolkit for the microscopic validation of its physics (both Standard and Low Energies processes are involved) with a powerful statistical tool.

2. **Astrophysics:** ESA Bepi Colombo mission [15] decided to use it with the aim of comparing Bessy test beam experimental data with Geant4 simulations of X-ray fluorescence emission.

3. **Medical physics:** CATANA INFN [16], the unique Italian group performing hadrontherapy and treating patients affected by uveal melanoma, use the **GoF** toolkit in order to make comparison of physical quantities of interest (as Bragg peak, isodose distributions).

4. CONCLUSIONS

The **GoF** toolkit is an easy, up-to-date, and powerful tool for data comparison in physics analysis. It is the first statistical toolkit providing such a variety of sophisticated and powerful algorithms in HEP.

By employing a rigorous software process, using object-oriented techniques as well as generic programming, the toolkit features a component-based design. This facilitates the re-use of the toolkit in other environments. The adoption of AIDA interfaces simplifies the use of the toolkit further.

The code is downloadable from the web [1] together with all the documentation concerning the User Requirements Document and the Traceability Matrix.

Finally, for all the features described, the **GoF** toolkit constitutes a step forward in HEP data analysis quality and could be easily used by other experimental software frameworks.

References

[1] `http://www.ge.infn.it/geant4/analysis/HEPstatistics/`

[2] A.N. Kolmogorov, “Sulla determinazione empirica di una legge di distribuzione”, Giorn. Ist. Ital. Attuari, 4, 1933: 1-11.

[3] L.A. Goodman, “Kolmogorov-Smirnov tests for psychological research”, Psychol. Bull., 51, 1954: 160-168.

[4] N.H. Kuiper, “Tests concerning random points on a circle”, Proc. Koninkl. Neder. Akad. van. Wetenschappen A, 63, 1960: 38-47.

[5] H. Cramèr, “On the composition of elementary errors. Second paper: statistical applications”, Skand. Aktuarietidskrift, 11, 1928: 171-180.

[6] R. von Mises, “Wahrscheinliehkeitsrechnung und ihre Anwendung in der Statistik und theoretischen Physik”, Leipzig, F. Deuticke, 1931.

[7] T.W. Anderson, D.A. Darling “Asymptotic theory of certain goodness of fit criteria based on stochastic processes” Ann. Math. Statist., 23, 1952: 193-212.

[8] S.V. Aksenov, M.A. Savageau, “Mathematica and C programs for minimum distance estimation of the S distribution and for calculation of goodness-of-fit by bootstrap”, 2001, in press.

[9] M. Fisz, “On a result by M. Rosenblatt concerning the von Mises-Smirnov test”, Ann. Math. Statist., 31, 1960: 427-429.

[10] J.M. Dufour, A. Farhat, “Exact nonparametric two-sample homogeneity tests for possibly discrete distributions”, Cahier 23-2001, Universitè de Montrèal.

[11] M.A. Stephens, “Introduction to: Kolmogorov (1933) on the empirical determination of a distribution”, in S. Kotz, N.L. Johnson, “Breakthrougth in statistics”, vol II, Springer Verlag, New York, 1992.

[12] E. Gamma, R. Helm, R. Johnson, J. Vlissides, “Design Patterns”, Addison Wesley Professional Computing Series, 1994.

[13] `http://AIDA.freehep.org`

[14] `http://geant4.web.cern.ch/geant4/`

[15] `http://sci.esa.int`

[16] `http://www.lns.infn.it/catanaweb/`

Code-Testing of Statistical Test Implementations

F. James, A. Pfeiffer, A. Ribon
CERN
P. Cirrone, S. Donadio, S. Guatelli, A. Mantero, B. Mascialino, L. Pandola, S. Parlati, M.G. Pia
INFN,
P. Viarengo
IST

In this note we discuss in general how to test the implementation code of statistical tests, and then we treat in detail the case of the Kolmogorov-Smirnov test. It will be shown that some "obvious" expected properties, like the flatness distributions of p-values from repeating drawings from the same parent distribution, are not indeed reproduced even in absence of bugs in the code, due to either asymptotic approximations in the formulas used to compute the p-value, or to the discreteness of the distance distribution in the case of direct Monte Carlo evaluation of the p-value. This makes the code-testing more complicated. Some practical advice is presented anyhow.

1. INTRODUCTION

It is essential, before using any statistical test (χ^2, Kolmogorov-Smirnov, Cramer-von Mises, Anderson-Darling, etc.), to check whether its code implementation is correct. The obvious way to do so is to compare the results in some particular cases against either different implementations of the same statistical test, or against some formulas or tabulated values from the Statistics literature. This approach is quite limited (for example, to the best of our knowledge, there are no publicly available implementations of the Anderson-Darling statistical test), and even in the few situations when a formula or a table (for the p-value given the distance) is available, it is usually obtained under some assumptions, the most typical one being the asymptotic limit of the sample size, and it is difficult, in general, to know what is the bias on the p-value caused by such approximations. We present here a method which does not rely on external code, papers, books, or tables, to validate any implementation of any statistical test. The method is based on the expected mathematical properties that any statistical test should exhibit, which are checked using Monte Carlo trials. Although we believe that the method provides a reasonable and powerful way to detect bugs in code implementations, it cannot give an absolute guarantee that the code is completely bug-free.
We will consider uniquely 1-dimensional distributions.

2. TESTING STRATEGIES

We aim here to be quite general, so we will use some symbolic notation and our discussion will be somehow abstract, but we will be soon back to a concrete example in the next section. For the same reason, in the following we will not say anything on the continuous or discrete nature of the parent distribution and whether the sample data should be binned or unbinned.

Let S_1 and S_2 be two 1-dimensional samples of size N_1 and N_2 respectively, and $d(S_1, S_2)$ be a test statistic measuring the distance between the two samples. We can then calculate $T(S_1, S_2)$, the probability that $d(S_i, S_j)$ would not be smaller than $d(S_1, S_2)$, for any samples S_i and S_j of size N_1 and N_2 drawn randomly from the same parent distribution. T is called the *p-value.* Here are some simple properties of T:

i) $T(S_1, S_2) = 0$
when the two samples are in non-overlapping regions of the real axis.

ii) $T(S_1, S_1) = 1$
i.e. when the two samples are identical.

iii) $< T(S'_1, S'_2) > \; \geq \; < T(S''_1, S''_2) > \; \geq \; ...$
where $< T >$ means the average of the p-values obtained from some drawings from the same parent distribution, for both samples S_1 and S_2, but with increasing shifts. For example, let's consider as S'_1 and S'_2 the samples drawn from the same gaussian distribution $N(\mu, \sigma)$; then consider as S''_1 and S''_2 the samples drawn respectively from $N(\mu + \sigma, \sigma)$ and $N(\mu - \sigma, \sigma)$; and so on, for less and less overlapping gaussian parent distributions.

iv) $T(S_1, S_2) = T(f(S_1), f(S_2))$
for any monotonic function $f(x)$. Notice that this property is not rigorously valid in the case of binned distributions.

v) $T(S_1, S_2) = T(S_2, S_1)$
i.e. the test should not depend on the order of the two samples, that is on which one we label as "1" and "2".

vi) The above properties should be valid independently of the parent distribution from which we draw the samples. In practice, we think that some reasonable choices for the parent distribution can be the following: flat (uniform), gaussian, left-tailed and right-tailed exponential.

Suppose that the statistical test fulfills all the above requirements; then we can move to the next step, which is much more CPU demanding and trickier.

We define as *Pseudoexperiment* a random drawing of two samples, S_1 of size N_1 and S_2 of size N_2, from the same parent distribution (whatever it is). Given these two samples, we can calculate the *distance*, d, between them according to the statistical test we are considering, and then from that distance we can calculate the corresponding p-value, p. For each pseudoexperiment we thus have: $S_1^{(j)}, S_2^{(j)} \rightarrow d^{(j)}$, $p^{(j)}$ where $j = 1, 2, ..., N$, with N number of pseudoexperiments. Now, from the distribution of distances, $d^{(j)}$, $j = 1, 2, ..., N$ we can calculate the p-value directly from its definition: *the p-value of a given distance $\overline{d}$ between two samples S_1 and S_2 (with respect to a given statistical test) is the probability to get a distance $d \geq \overline{d}$ between two samples of the same size as S_1 and S_2 drawn from the same parent distribution (whatever it is).* In practice, the above probability is estimated as *the fraction of pseudoexperiments whose distance $d^{(j)} \geq \overline{d}$*. We call this operative definition of the p-value *Monte Carlo p-value*, p_{MC}.
Notice that it is important to include the *equal* case in $d \geq \overline{d}$: this of course would not matter for real continuous distributions, but in practice we are always dealing with discrete distributions of distances.
Concretely, one can consider either $N_1 = N_2$ or $N_2 >> N_1$; in the latter case, one could think also to draw the second sample, the one with higher statistics, only once instead of for each pseudoexperiment; as limiting case of $N_2 \rightarrow \infty$, one can make a 1-sample statistical test, comparing directly S_1 with the parent distribution, at least in the cases in which the analytic expression of its cumulant probability distribution is known. We will compare these possibilities in the example of the next section.

Naively, we would be tempted to require as necessary properties of the statistical test under consideration the following two:

a) In the limit of a large number of pseudoexperiments, N, the distribution of p-values (obtained from the statistical test), $p^{(j)}$, should be a flat (uniform) distribution between 0 and 1, hence, in particular, it should have: $\mu = \frac{1}{2}$, $\sigma = \frac{1}{\sqrt{12}}$ (where μ is the mean, and σ the rms).

b) Apart for tiny deviations due to finite numerical accuracy, the p-values determined from the statistical test have to coincide with the ones determined directly from Monte Carlo: $p^{(j)} = p_{MC}^{(j)}$.

Both properties are *not* true, for two independent reasons. The first one is due to the fact that p-value computed by the statistical test is usually valid under some "asymptotic" conditions, the most general one being the limit of the sample size (in our case above, N_1 and N_2) to ∞. The second one, more subtle, is due to the *discreteness of the distance distribution* $\{d_1, d_2, ...\}$, even in the limit of a very large number of pseudoexperiments, $N \rightarrow \infty$. As as consequence of this, even the following property, which is the analogous of a) for p_{MC}, does *not* hold:

a') In the limit of a large number of pseudoexperiments, N, the distribution of p-values calculated directly from Monte Carlo, $p_{MC}^{(j)}$, should be a flat (uniform) distribution between 0 and 1, hence, in particular, it should have: $\mu = \frac{1}{2}$, $\sigma = \frac{1}{\sqrt{12}}$.

It is even possible to find a very simple formula which predicts the mean value of $\left\{p_{MC}^{(j)} \; ; \; j = 1, 2, ..., N\right\}$, given the multiplicities of the various distances, i.e. the number of times that each different distance appears:

distance d_1 with multiplicity M_1 ;
...
distance d_K with multiplicity M_K ;

where K is the number of *different* distances, and $\sum_{i=1}^{K} M_i = N$:

$$< p_{MC} >= \frac{1}{2} + \frac{1}{2N} + \sum_{i=1}^{K} \frac{M_i\,(M_i - 1)}{2N^2} \qquad (1)$$

Notice that:

- $< p_{MC} > \; > \frac{1}{2}$ in all cases (with finite N);
- for a given N, the lowest value of $< p_{MC} >$, that is the closest to $\frac{1}{2}$, is reached when $K = N$, that is when all the distances are different: $M_1 = 1$, ..., $M_K = 1$: $< p_{MC} >= \frac{1}{2} + \frac{1}{2N}$
- in order to get $< p_{MC} > \rightarrow \frac{1}{2}$ not only $N \rightarrow \infty$ is necessary, but also $K/N \rightarrow 1$, i.e. only a finite number of distances can be repeated;
- $< p_{MC} >$ depends only on the multiplicities of the K different distances, but not on the explicit values of these distances.

The consequence of the above facts is that the task of checking the code implementation of a statistical test becomes much harder, because, for instance, discrepancies between computed p-values and direct Monte Carlo ones are expected even with no bugs in the code. However, these should decrease as the asymptotic conditions are approached. Notice that, in the case the remaining discrepancies are judged unacceptable but the implementation of the p-value is correct and a better formula for the p-value cannot be found, the direct Monte Carlo p-value can always be employed. The only drawback of this approach is that it is quite CPU intensive.

3. AN EXAMPLE: THE KOLMOGOROV-SMIRNOV TEST

For the Kolmorogov-Smirnov test we use the p-value formula given in [1]. As parent distribution we consider the flat (uniform) distribution between 0 and 1, because in this case the cumulant probability distribution is known ($F(x) = x$). $N = 100\,000$ pseudoexperiments have been generated, of four different types as defined by the way the distance has been determined:

d1 : in each pseudoexperiment we draw a single sample S_1 of size N_1, and then we consider the 1-sample Kolmogorov-Smirnov test against the parent distribution $F(x) = x$.
Hereafter we indicate with **d1** the corresponding distance.

d2a : in each pseudoexperiment we draw a single sample S_1 of size N_1, and then we consider the 2-sample Kolmogorov-Smirnov test against another sample, S_2, of very large size, $N_2 = 10\,000$, which is drawn from the same parent distribution, but only once (at initialization, not in each pseudoexperiment).
Hereafter we indicate with **d2a** the corresponding distance.

d2b : in each pseudoexperiment we draw two samples, S_1 of size N_1, and S_2 of very large size $N_2 = 10\,000$, and then we consider the 2-sample Kolmogorov-Smirnov test between them.
Hereafter we indicate with **d2b** the corresponding distance.

d2c : in each pseudoexperiment we draw two samples, S_1 and S_2, of the same size N_1, and then we consider the 2-sample Kolmogorov-Smirnov test between them.
Hereafter we indicate with **d2c** the corresponding distance.

As sample size N_1 we consider the following possibilities: $10\,,\, 50\,,\, 100\,,\, 500\,,\, 1000\,,\, 5000\,,\, 10\,000\,.$ For the sample size N_2, when not equal to N_1, we use $N_2 = 10\,000\,.$ Only for the case $N_1 = 1000$, to see what happens when N_2 is changed, we also consider $N_2 = 100\,000\,$, i.e. an increase of a factor ten.
For each of the above four types of distances (and, of course, for each pseudoexperiment) we determine two types of p-value: p, the analytic p-value, and p_{MC}, the p-value from the direct Monte Carlo method. The table on the right side reports the summary of our study. In the first column there is N_1, the size of the sample S_1; in the second column there is the type of distance; in the third column there is the number of different distances (i.e. what we have called "K" in the previous section); in the last two columns there are the mean values (over the N values obtained in the pseudoexperiments) of the two different types of p-values, p, and p_{MC} (in the latter case, such mean value agrees with the one predicted by (1)).

N1	Type	distances	$< p >$	$< p_{MC} >$
10	d1	99,768	0.6605	0.5000
	d2a	4,421	0.5118	0.5002
	d2b	4,467	0.5120	0.5002
	d2c	10	0.5509	0.6290
50	d1	99,906	0.5863	0.5000
	d2a	2,241	0.5103	0.5004
	d2b	2,261	0.5097	0.5004
	d2c	24	0.5304	0.5585
100	d1	99,864	0.5637	0.5000
	d2a	1,651	0.5101	0.5006
	d2b	1,654	0.5085	0.5006
	d2c	31	0.5237	0.5415
500	d1	99,712	0.5286	0.5000
	d2a	776	0.5129	0.5013
	d2b	802	0.5030	0.5013
	d2c	67	0.5112	0.5186
1,000	d1	99,547	0.5210	0.5000
	d2a	569	0.5185	0.5019
	d2b	595	0.5034	0.5018
	d2c	92	0.5076	0.5132
x10 S2 size	d2a	4,643	0.5028	0.5002
	d2b	4,646	0.5026	0.5002
5,000	d1	99,131	0.5087	0.5000
	d2a	288	0.5722	0.5039
	d2b	335	0.5020	0.5034
	d2c	193	0.5035	0.5059
10,000	d1	98,827	0.5056	0.5000
	d2a	208	0.6441	0.5056
	d2b	278	0.5027	0.5041
	d2c	268	0.5022	0.5042

From the table on the right side we can make the following observations:

- The number of different distances grows with:
  a) the number of pseudoexperiments;
  b) the sample size, in the case of two samples of equal size drawn in each pseudoexperiment;
  c) inversely with the sample size of the first sample, in the case that the second sample has a much larger size, and no matter whether is drawn once or each time;
  d) the sample size of the second sample, in the case the latter is much bigger than the first one, and no matter whether is drawn once or each time.

In the case of a single sample, that is when we compare the sample directly with the parent distribution, the number of different distances is almost always equal to the number of pseudoexperiments. In the case of two samples, but with the second of much higher size, the number of different distances is always slightly bigger (but very little) in the case of drawing of both samples in each pseudoexperiment, with respect to the case of a single drawing for the second sample.

- The mean of the Monte Carlo p-values, p_{MC}, depends only on the number of different distances, and their multiplicities, as predicted from (1);

- The means of the theoretically calculated p-values, p, have the following characteristics:

  a) they are systematically above $\frac{1}{2}$;

  b) for d1, d2b and d2c, p gives mean p-values which are closer to $\frac{1}{2}$ the larger the sample size N_1 is; however, in the case of d2b, a "saturation" sample size is reached for values around $N_1 = 500$;

  c) for d2a, p gives mean p-values which are not always getting closer to $\frac{1}{2}$ the larger the sample size gets, because N_1 gets closer to N_2 but we draw the second sample only once.

4. CONCLUSIONS

From the study we have presented it is possible to draw some useful practical suggestions on code-testing of statistical test implementations. Although we treated here explicitly only the case of the Kolmogorov-Smirnov test, we believe that such advices are valid in general, for any statistical test.

1) First of all, start by checking the properties i) ÷ vi) (see the section "Testing strategies"), and move on only when they are all satisfied.

2) Generate a very large number N of pseudoexperiments (e.g. 100 000), and for each pseudo-experiment do the following:
draw a sample S_1 of size N_1 (a fixed, arbitrary value, e.g. $N_1 = 100$), and a sample S_2 of size $N_2 >> N_1$ (e.g. $N_2 = 10\,000$), from the same parent distribution (whatever it is), and then calculate the distance and the p-value of the statistical test under consideration.
(Notice that a 2-sample, rather than 1-sample, statistical test is used because in general it is not possible to find an analytical expression for the cumulant probability distribution of a given parent distribution.)

3) From the distribution of distances, calculate the direct Monte Carlo p-value for each distance (i.e. pseudoexperiment).

4) Calculate the average of the direct Monte Carlo p-values: this has to coincide exactly with what is predicted by (1) (to apply this formula only the multiplicities of the different distances are needed). If this is not the case, then there is something wrong in the testing code itself (don't blame the statistical test implementation). Move on only when the two agree.

5) Calculate the average of the p-values returned by the statistical test, and the average and maximum absolute difference between these p-values and the corresponding direct Monte Carlo ones.

6) Repeat 2) ÷ 5) for few different values of N_1 (e.g. $N_1 = 100$, 500, 1000, 5000). You should observe a convergence, as N_1 grows, between the p-values returned by the statistical test and the direct Monte Carlo ones. If this is not the case, then there is something wrong in the statistical test implementation either with the distance calculation or with the p-value determination.
Finally, if such convergence is indeed observed, one should judge whether the average and maximum absolute difference of the p-values returned by the statistical test and by the direct Monte Carlo method, in the case of the highest N_1 value (e.g. 5000), look "reasonable" under the assumption that they are entirely due to the asymptotic approximations on which the p-value formula (or table) is based on. If this is not the case, then the implementation of such p-value should be first checked, and if it is fine, then a better formula should be used instead (if it can't be found, the direct Monte Carlo p-value can be employed; eventually, if it is too slow to do on the fly, one could store the Monte Carlo results on a table once for all).

References

[1] W.H Press, S.A. Teukolsky, W.T. Vetterling, B.P. Flannery *Numerical Recipes in C*, Cambridge (see chapter 14).

Highly-Structured Statistical Models in High-Energy Astrophysics

David A. van Dyk
Department of Statistics, University of California, Irvine, CA 92697-1250

In recent years, an innovative trend has been growing in applied statistics—it is becoming ever more feasible to build application-specific models which are designed to account for the hierarchical and latent structures inherent in any particular data generation mechanism. Such highly-structured models have long been advocated on theoretical grounds, but recently the development of new computational tools (e.g., hardware, software, and algorithms) for statistical analysis has begun to bring such model fitting into routine practice. In this paper, we describe these methods in the context of empirical high-energy astrophysics. A new generation of scientific instruments such as the *Chandra X-ray Observatory* are opening a whole new window to the study of the cosmos. Unlocking the information in the data generated with these complex high-tech instruments, however, requires sophisticated statistical models, methods, and computation. Here we discuss the techniques that the California-Harvard AstroStatistics Collaboration have been using to develop application-specific highly-structured statistical models to address these problems in high-energy astrophysics.

1. THE *CHANDRA X-RAY OBSERVATORY*

The *Chandra X-ray Observatory* took its place along side the *Hubble Space Telescope* and the *Compton Gamma-ray Telescope* as part of NASA's fleet of Great Observatories when it was launched by the Space Shuttle *Columbia* in July 1999. *Chandra* is by far the most precise X-ray telescope ever constructed; it is able to produce images over thirty times sharper than those available from previous X-ray telescopes. Although *Chandra* is a good example of a modern complex scientific instrument, it is but one of a host of such instruments. The complexity of these instruments along with the complexity of the objects that they study and the scientific questions they aim to answer demand sophisticated statistical methods. Off-the-shelf statistical techniques are simply not up to the inferential tasks involved in the scientific exploration of such data. In this paper we use *Chandra* as an example, to show how sophisticated application-specific statistical methods can be designed to meet the scientific challenges posed by modern instrumentation.

Chandra collects data on each photon that arrives at its active detector. The two-dimensional sky coordinates, energy, and time of arrival of each photon are recorded. Because of instrumental constraints, each of these quantities is rounded or binned into a discrete variable. Thus, in principle, the data can be represented by a four-way table of counts. Spectral analysis investigates the one-way marginal table of energy counts; image analysis focuses on the two-way marginal table of coordinates; and timing analysis studies the one-way table of arrival times. More sophisticated analysis might look at joint distributions to study, for example, how the spectrum varies across an extended source. In this paper we confine our attention to spectral analysis and image analysis. As we shall see, even these marginal analysis pose significant challenges.

A typical spectrum is modeled as a mixture of a smooth broad continuum term and a number of narrow emission lines. The continuum is formed by thermal (heat) radiation or by non-thermal processes in relativistic plasmas. The continuum is modeled using a smooth parametric form that includes emission across the entire width of the spectrum. Emission lines, on the other hand are narrow features in the spectrum that can be modeled with Gaussian distribution, Lorentzian distributions, or delta functions. When an electron jumps down from one quantum state of an atom to another, the energy of the electron decreases. This energy is radiated away from the atom in the from of a photon with energy equal to difference of the energies associated with the two quantum states. Unlike the emission that forms the continuum, the energies associated with these differences are discrete and from the emission lines in a spectrum.

Taken together, these features of the spectrum give subtle clues as to aspects such as the temperature and composition of the physical environment of the cosmological source. A stellar corona, for example, is made of numerous ions which can be recognized in a spectrum from their identifying emission lines. If the corona is relatively hot, the emission lines that correspond to more energetic quantum states will be relatively strong. Thus, the relative strength of the emission lines corresponding to a particular ion carries information as to the temperature of the source. Figure 1 shows an ultra high-resolution *Chandra* observation of the spectrum of the star Capella (α Aur). The spectrum is composed of a forest of spectral emission lines. Taken along with prior information obtained from detailed quantum mechanical computations and ground-based laboratory measurements, this data can be used to construct the physical environment of Capella's coronae. These calculations require sophisticated application-specific statistical methods. We do not discuss the details here. Instead we refer the interested reader to van Dyk *et al.* [2004] and detail a much simpler example in Section 3.

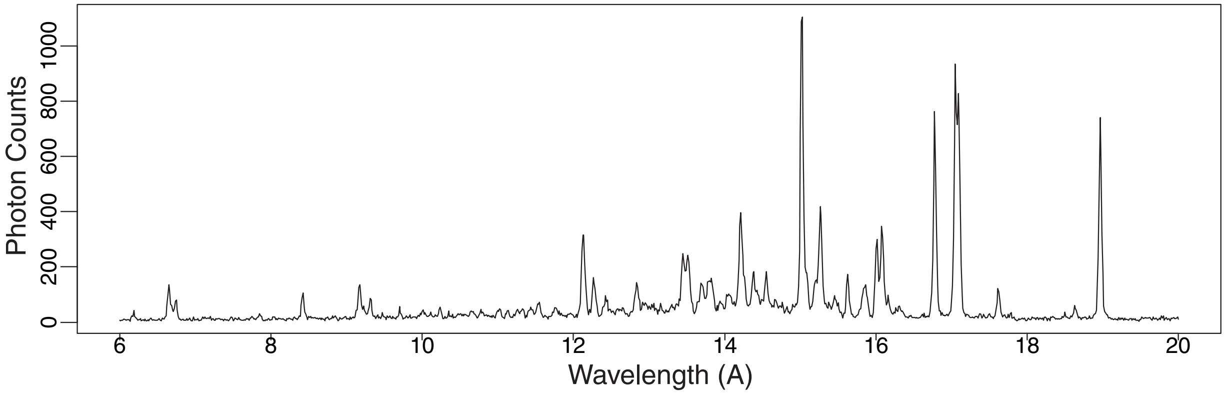


Figure 1: The Spectrum of Capella (α Aur). This high-resolution spectrum was collected using *Chandra*'s high-resolution camera along with its low-energy transmission grating spectrometer. Notice the numerous emission lines that compose the spectrum. A scientific goal is to use this forest of emission lines to reconstruct the composition and the distribution of the temperature of Capella's coronae, where the X-ray emission is produced.

Like spectra X-ray images can be composed of extended smooth features along with local bright features. At one extreme an image might reveal a smooth extended source such as a nebula without bright stars. There may also be a few bright point sources in the smooth extended emission, or the extended emission may be peppered with numerous point sources. Irregular and unpredictable structure is the rule rather than the exception when examining cosmological images. Figure 2 is a *Chandra* image of the central region of the galaxy NGC 6240, the product of the collision of two smaller galaxies. The center of this galaxy is dominated by two massive black holes; one is clearly visible as a white pixel in Figure 2. There appears to be additional structure in the extended source. A loop of hot gas appears toward the upper right of the image, and a larger fainter loop appears off to the right. Because of the high variability of the low-count per pixel data that typifies *Chandra*'s high-resolution images, however, it is difficult to distinguish features in the galaxy from artifacts of the statistical noise. As we shall discuss next, the situation is further complicated by a number of processes inherent in the data collection mechanism that degrade the quality of the data.

Both spectral and spatial characteristics of the data are degraded in a number of ways that must be accounted for in any principled data analysis. For example, the *effective area* of the detector varies with the energy of the photon. *Chandra* focuses X-rays with mirrors. Unfortunately, high-energy photons do not reflect uniformly and simply; some are absorbed and some pass right through the reflector, with a probability that is a function of their energy. A similar process occurs before the photon reaches the detector; lower energy photons are more likely to be absorbed by inter-stellar or inter-galactic media. Thus, the probability that an X-ray reaches the detector depends on the X-ray's energy. In statistical terms, we refer to the photons that are absorbed or undetected because of a relatively small effective area as *missing data*. Because ignoring the missing data mechanism would result in biased spectral analysis, it is called non-ignorable missing data Rubin [1976]. The likelihood that a photon is recorded also depends on where it lands on the detector. Photons landing near the boundary of the CCDs, for example, are less likely to be recorded. This effect is calibrated by the so called *exposure map*.

Because the focusing of the mirrors is not perfect the image of a point source is blurred; the character of the blurring is recorded in the *point spread function*. Another form of data degradation is due to a detector response, which results in a blurring of the photon energies. The recorded energy of a photon that arrives with a particular energy and location on the sky has a probability distribution. Finally, the source photons are generally contaminated by background counts. Common methods for handling data distortion can be quite ad hoc. For example, in spectral analysis a second data set is collected that is assumed to consist only of background counts. This background data is often *directly* subtracts from the source data and the result is analyzed as if it were a source observation free of background contamination. This procedure can lead to negative counts and estimates with questionable statistical properties.

The complexity of the cosmological sources, of the instrumentation, and of the scientific questions combine to result in sophisticated data analytic challenges. In the following sections we discuss how we propose to address these challenges using sophisticated application-specific highly-structured models. More details about *Chandra* and the analysis of *Chandra* data can be found in van Dyk *et al.* [2004]. The application of our methods to spectral analysis is the topic of van Dyk *et al.* [2001], Protassov *et al.* [2002], and van Dyk and Kang [2003]; image analysis is dis-

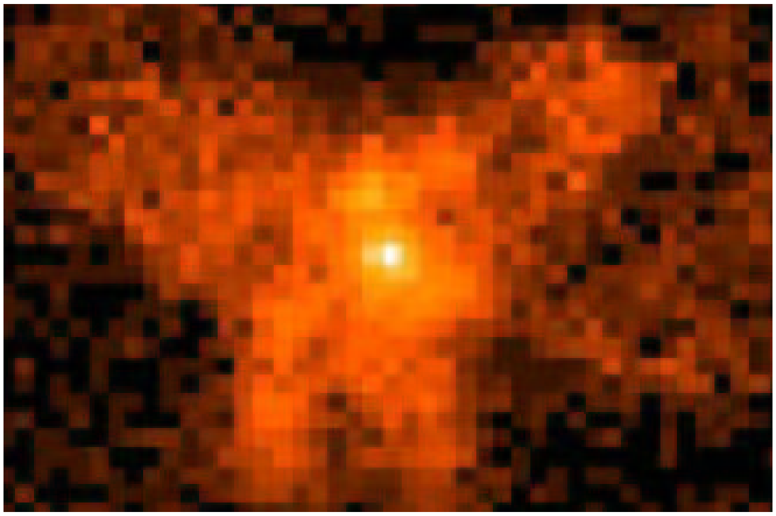


Figure 2: An X-ray Image of NGC 6240. The bright spot at the center of this galaxy is massive black hole. A second black hole appears above and a bit to the left of the brighter black hole. A loop of hot gas appears in the upper right quadrant of this image, and there appears to be a second larger loop off to the right. Whether these are actual features in this galaxy or artifacts of the highly-variable low-count Poisson data is an important astrostatistical question.

cussed in van Dyk and Hans [2002] and Esch *et al.* [2004].

2. APPLICATION-SPECIFIC STATISTICAL MODELS

Any principled analysis of *Chandra* data must account for the complexity of the data generation mechanism. Thus, we propose designing application-specific models that accounts not only for the spectra or images that are of primary scientific interest, but also for the complexities of the instrumentation. These models can be formulated via a hierarchical set of levels that allow us to separate a complex data generation mechanism into a number of well-understood components each of which on its own can be addressed via standard statistical techniques. When these levels are combined they form a highly-structured model that specifically addresses the complexities of the problem at hand. This multi-level approach not only allows us to formulate a highly-structured model using simple tools, but also gives us access to computational techniques that take advantage of the multi-level structure to combine a number of simple steps to fit a highly-structured model. In this article we outline the use of highly-structured multi-level models; a more thorough introduction can be found in Gelman *et al.* [2003].

One way to formulate these models in terms of *missing data.* In a spectral analysis, for example, we might consider the *ideal counts* to be a mixture of continuum and emission line counts that are unaffected by the effective area of the instrument, by photon absorption, by blurring of the photon energies, or by background contamination. These ideal counts are unobserved, and, in this sense can be regarded as missing data. Each ideal count can be modeled as a finite mixture of Poisson random variables, each of which corresponds to one of the emission lines or the continuum term. The key here is that this model can be formulated ignoring all of the mechanisms that degrade the data, thereby separating the complexity of the instrumentation from that of the cosmological sources. Thus, we are able to separate the task of modeling a sophisticated data generation mechanism into a sequence of simple tasks.

Adding a level to this model, we can account for photon absorption by modeling the ideal counts subject to absorption given the ideal counts. There is a parameterized absorption probability in each bin that depends on the energy corresponding to the bin. Because absorption operates independently on each photon, the counts in each bin after absorption are binomial given the corresponding ideal counts. In statistical terms, we use a generalized linear model that is akin to logistic regression for this level of the model. What is important here is that this is again a standard well-understood statistical model.

Similarly we can add levels to the model to account for the effective area of the instrument, the blurring of the phonon energies, and background contamination. Each of these levels incorporates what is known about the instrumentation (e.g., from instrument calibration) into a standard statistical model.

The discussion in this section is a broad overview of how multi-level models can be built for spectral analysis of *Chandra* data; details of this approach in this specific application can be found in van Dyk *et al.* [2001] and van Dyk and Kang [2003]. The key here, however, is the principle that a statistical model can be designed to incorporate specific features of any scientific data generation mechanism. The goal should be to formally model as much as this mechanism as possible. The mathematics of probability modeling ensures that properly modeled variability in the data generation mechanism will be reflected in the resulting uncertainty in the fitted model and the error bars on the model parameters. Preprocessing the data and ad hoc data manipulation do not properly account for variability and can have unpredicted consequences on the statistical properties (e.g., bias and coverage) of the resulting estimates. In practice, there is always some data preprocessing that must occur, but it is important to consider the effects on model uncertainty and generally to avoid preprocessing when possible.

A Simple Example. For the purpose of illustration, we consider a simple example; we emphasize that this example is not meant to illustrate the power of highly-structured models, but rather to show how they work. More sophisticated examples appear in the papers cited in this article.

Suppose we observe a single count, Y that is background contaminated, i.e., Y is a mixture of source and background counts. We also observe a second count, Z, that is a pure background count. We allow the exposure times for the two counts to be different and label them τ_S and τ_Z, respectively. The goal is to estimate the source count rate. We can easily formulate this problem in terms of missing data by supposing $Y = Y_S + Y_B$, where Y_S and Y_B are the source and background counts in the initial exposure. Clearly, Y_S and Y_B are unobserved quantities; we refer to these quantities as missing data. If the missing data were observed, it would be easy to estimate the source count rate, Y_S/τ_S. Likewise if the source and background count rates were known, we could easily split Y into Y_S and Y_B based on the relative intensities of the two rates. Thus, identifying Y_S and Y_B as missing data simplifies the relationships among the data and the quantities of scientific interest. Although this discussion is heuristic, it can be formalized to formulate algorithms for maximum likelihood fitting and Bayesian methods. Details in this particular example can be found in van Dyk [2003], a general discussion of these topics are the subject of Section 3 and 4.

3. STATISTICAL INFERENCE

Fitting a statistical model involves not only statistical computation, the subject of Section 4, but also the selection of a criterion for the fit. Common methods include χ^2 fitting, maximum likelihood, and Bayesian methods.

The method of χ^2 fitting ignores the variance structure inherent in the data by essentially making large sample Gaussian assumptions on the errors. As such, this method is especially inappropriate for high-resolution low-count data which exhibit Poisson errors.

Methods based on the likelihood are more appropriate in that they can explicitly account for error structures in the data. Bayesian methods take this one step further by allowing statistical inference that combines other scientific information with the data . The prior distribution is used to quantify information outside the data, the likelihood function quantifies information in the data, and these are combined via Bayes Theorem to form the posterior distribution. From a Bayesian perspective, the posterior distribution is a compete summary of the available information.

In practice, the prior distribution can be used to quantify information available from other data sources, from instrumental calibration, or from analytical physical calculations. Prior distribution are often used to quantify what is know about the values of parameters that are not of primary interest, and in some cases are used to quantify what is known about the likely values of parameters of direct scientific interest. Alternatively, prior distributions can be used to introduce structure on groups parameters. For example, it might be known that the values of a group of parameters are related to each other. This is sometimes the case with the wavelength of a group of emission lines associated with a particular ion. In image analysis, we can use prior distributions to encourage a smooth reconstruction of extended emission. Thus, we emphasizes that despite their reputation for being subjective and unscientific, prior distribution can be used in an objective manner to quantify model assumptions or concrete scientific information.

In Figure 3, we illustrate prior and posterior distributions under the simple background contamination example introduced in Section 2. The figure corresponds to a simulated data set with $Y = 1$, $Z = 48$, $\tau_S = 1$, and $\tau_B = 24$. The first plot illustrates two possible prior distributions on the source rate; one is flat and the other prefers values near three. The corresponding posterior distributions appear in the second plot. Since both the source and background rates are unknown parameters, the posterior distributions for the source rate are marginal distributions that result from integrating the joint posterior distribution over the background rate. An attractive feature of Bayesian methods is a simple principled prescription for handling nuisance parameters: They can be integrated out, leaving the marginal posterior distribution of the parameters of interest.

The posterior distributions in the second plot represent a compromise between the data and the prior distributions. The data from the background exposure alone, $Z = 48$ with $\tau_B = 24$ suggests a background rate of two. Given that Y is only one, direct background subtraction would result in a nega-

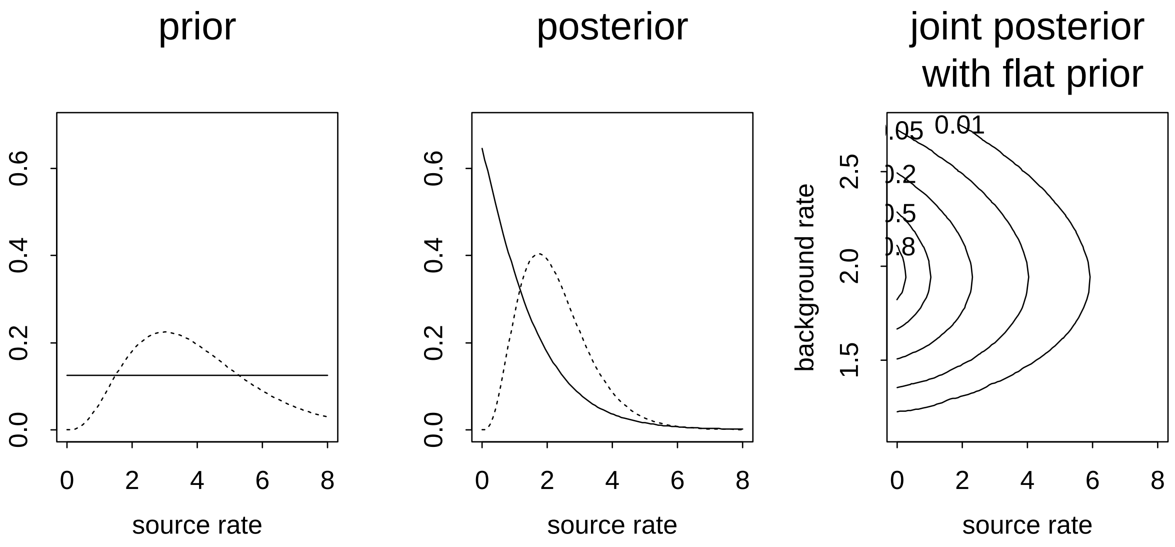


Figure 3: Sensitivity Analysis. The first plot shows two possible prior distributions on the source rate parameter. The second plot represents the resulting marginal posterior distributions on the same parameter. Notice that in this case the posterior distribution is sensitive to the choice of the prior distribution. The final plot shows the joint posterior distribution of the source and background rates under the flat prior distribution. The marginal posterior plotted with a solid line in the second plot is the integral of that in the third plot over the background rate.

tive source rate of -1.[1] Thus, the data favors small values of the source rate; the maximum likelihood estimate is zero. This is reflected in the solid posterior distribution, which has its mode at zero. The dotted posterior distribution, on the other hand, is a compromise between the dotted prior distribution which favors slightly larger values of the source rate and the data. The final plot illustrates the contours of the joint posterior distribution under the flat prior on the source rate. Integrating this joint posterior distribution over the background rate yields the solid posterior distribution in the second plot.

From a Bayesian perspective, the marginal posterior distribution of the source rate is a complete summary of the information available for this parameter. The posterior mode or mean are often used as the fitted values of the parameters, while some measure of the posterior variability is used to generate confidence intervals or error bars. Any such summary of the posterior distribution, however, is an imperfect representation and is less informative than the posterior distribution itself. Summaries of this sort are especially problematic when the posterior distribution is multi modal or highly skewed. This is illustrated by the marginal posterior distribution plotted by the solid line in the second plot of Figure 3. Although the mode of this distribution is zero, this value does not appear to be an adequate summary of the distribution. Thus, one of the primary advantages of the Monte Carlo methods described in Section 4 is that they summarize the entire posterior distribution.

4. STATISTICAL COMPUTATION

In this section we discuss two computational methods for posterior exploration: mode finders and Monte Carlo methods.

Although modes can be misleading summaries of likelihood functions or posterior distributions, mode finders can be useful for initial exploration. For example, a *Chandra* image can easily have tens of thousands of pixel intensities. When working in very high dimensional parameter spaces, algorithms that quickly find areas of high posterior probability are a valuable tool.

There are many well-known strategies for finding modes of high-dimensional posterior distributions; Newton's method, Fisher's scoring, and conjugate gradient are well-known examples. Here we discuss another method that is especially useful with highly-structured models. The EM algorithm (Dempster *et al.* [1977]) is a two-step iterative routine for computing posterior modes (maximum a posterior, MAP, estimates) in problems that are formulated in terms of missing data. Details can be found in van Dyk [2003] or McLaughlan and Krishnan [1997]; here we

[1] The fact that this ad hoc technique results in a negative count rate is an indication that such ad hoc methods can have unexpected and uninterpretable results. Thus, these methods should be avoided and model based methods such as maximum likelihood or Bayesian methods should be preferred.

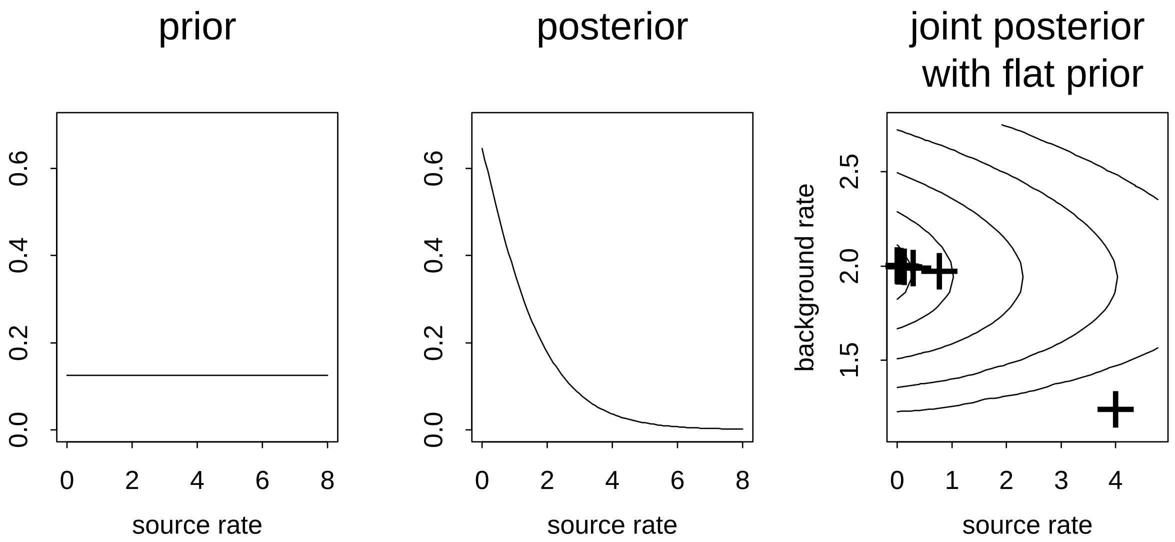


Figure 4: Using the EM Algorithm to Find the Posterior Mode. The first two plots are the same as in Figure 3 except that we consider only the flat prior distribution on the source rate. The final plot illustrates the contours of the corresponding joint posterior distribution along with the steps of an EM algorithm designed to compute the posterior mode. The skewed nature of the posterior distribution illustrates the potentially misleading nature of the modal estimates.

simply discuss how the EM algorithm works in the simple background contamination example described in Section 2. The first step of the EM algorithm, the *Expectation-Step*, replaces the missing values of Y_S and Y_B by their conditional expectation given Y, X, and the source and background rates. Simple probabilistic calculations show that this is accomplished by dividing the counts, Y into Y_S and Y_B using the relative size of the two corresponding rates. In the second step of this EM algorithm, the *Maximization-Step*, the rates are updated treating (Y, Y_S, Y_B, X) as data, i.e., we set the source and background rates equal to Y_S/τ_S and $(Y_B + X)/(\tau_S + \tau_B)$, respectively. The iterates of this EM algorithm using the simulated data set discussed in Section 3 are illustrated in Figure 4.

There are a variety of extensions to the EM algorithm that significantly broaden its application in the context of models formulated in terms of missing data (Meng and van Dyk [1997], van Dyk and Meng [2000], McLaughlan and Krishnan [1997]). As our example illustrates, EM-type algorithms are often easy to formulate even in highly-structured models. Another advantage of the EM algorithm is that it exhibits much more predictable and stable convergence than many other mode finders. In particular, it is guaranteed to increase the value of the posterior distribution at each iteration, i.e., it converges monotonically, see Figure 4. The primary disadvantage of the EM algorithm is that it sometimes can be slow to converge. Several of the extensions of EM, however, can be used to significantly improve its rate of convergence (van Dyk and Meng [2000], McLaughlan and Krishnan [1997]).

Although mode finders are useful for initial exploration of a posterior distribution, more sophisticated methods are required for thorough exploration. Figure 5 shows the same prior and posterior distributions as Figure 4, but includes a Monte Carlo sample from the posterior distribution. The Monte Carlo sample can be used to summarize the full posterior distribution and to easily represent marginal distributions of interest. Thus, the histogram of the Monte Carlo sample in the second plot of Figure 5 contains the same information as the plotted marginal posterior distribution. Likewise, the scatterplot of the sample in the third plot conveys the same information as the contours of the joint posterior distribution. The advantage of the Monte Carlo sample is clear when one considers high-dimensional parameter spaces. With a Monte Carlo sample from the joint posterior distribution one can easily plot histograms of the relevant marginal distributions even if the dimension of the joint distribution is in the tens, hundreds, thousands, or larger. Thus, with a Monte Carlo sample we can numerically integrate a distribution that would be impossible to integrate with any other numerical method.

There is a large statistical literature on methods to obtain a Monte Carlo sample from posterior distributions. One method that has proved to be very useful is to construct a Markov chain with stationary distribution equal to the target posterior distribution. Upon convergence, the Markov chain will deliver a (correlated) Monte Carlo sample from the posterior

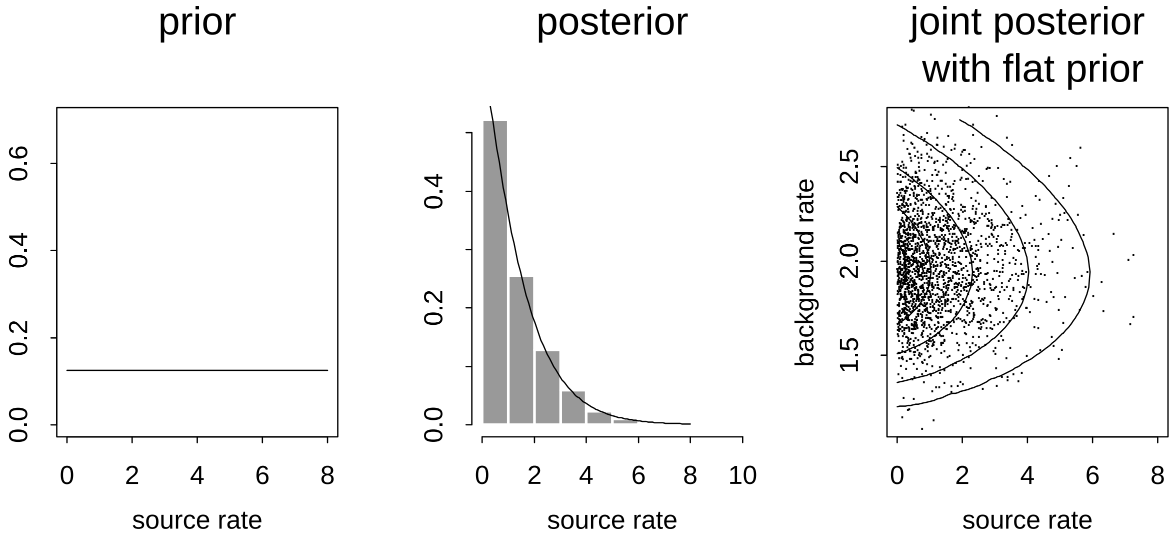


Figure 5: Monte Carlo Samples from the Posterior Distribution. The prior and posterior distributions in the three figures are the same as those in Figure 4. The second two plots illustrate Monte Carlo samples from the posterior distributions. The histogram conveys the same information as plotted marginal posterior distribution. Likewise, the scatter plot in the third plot contains information equivalent to that in the contour plot.

distribution. This technique is known as Markov chain Monte Carlo or MCMC. There are a number of technical issues that arise when using MCMC. It is more difficult to determine when a Markov chain has reached its stationary distribution than when a mode finder has reached a mode. Multi-modal posterior distributions pose extra challenges because Markov chains can easily be caught in one of the modes. This again highlights the advantage of identifying the modes using a mode finder *before* running a MCMC sampler. Severe correlation among the draws can also complicate Monte Carlo integration and evaluation of the posterior distribution. Thus, numerous strategies have been developed to improve the convergence and to reduce the autocorrelation of MCMC samplers. We do not attempt to address these issues here. Instead we point the interested reader to a number of references on the subject (Gelman *et al.* [2003], Gilks *et al.* [1996], van Dyk [2003], van Dyk and Meng [2001]) and describe how the Gibbs sampler can be used to construct a Markov chain with stationary distribution equal to the joint posterior distribution illustrated in Figure 5.

The Gibbs sampler constructs a Markov chain by partitioning the vector of unknown quantities (e.g., model parameters and missing data) into a number of subvectors. Each of these subvectors is updated by sampling from its conditional distribution given the most recent draw of the other subvectors and the observed data. In our simple example, we wish to sample from the posterior distribution of Y_S, Y_B, and the source and background rates given Y and X. We start by sampling Y_S and Y_B given the two rates and the observed data. That is, we stochastically separate Y into source and background counts. It can be shown that this distribution is a simple binomial distribution with probability determined by the relative sizes of the source and background rates and the number of trials equal to Y. In the second step, we sample the rates given Y_S, Y_B, Y, and X. Under this conditional distribution, the rates are independent and both follow gamma distributions, the posterior distribution of a Poisson rate parameter under the standard Bayesian prior distribution. Thus, we divide the unknown quantities into two groups: the missing data and the rate parameters. By iteratively sampling each from their corresponding standard conditional distributions, we construct a Markov chain with stationary distribution equal to the target posterior distribution. The result under our simulated data set is plotted in the histogram and scatter plot in Figure 5.

5. SUMMARY

In this article we have outlined a framework for statistical inference that designs application-specific highly-structured statistical models, uses a Bayesian paradigm for statistical inference, and utilizes sophisticated computational methods such as EM-type algorithms and MCMC. Although space does not permit us to illustrate the power of these methods in real problems, we hope interested readers will refer to the several papers cited in this article that use these methods to solve outstanding challenges in empirical high-energy astrophysics.

Acknowledgments

The author gratefully acknowledges funding for this project partially provided by NSF grant DM-01-04129 and by NASA Contract NAS8-39073 (CXC). This work is the product of the collaborative efforts of the California-Harvard AstroStatistics Collaboration (CHASC, URL: `http://www.ics.uci.edu/∼dvd/astrostat.html`) whose members include A. Connors, D. Esch, P. Freeman, H. Kang, V. L. Kashyap, X. L. Meng, T. Park, A. Siemiginowska, E. Sourlas, Y. Yu, and A. Zezas.

References

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B, Methodological* **39**, 1–37.

Esch, D. N., Connors, A., Karovska, M., and van Dyk, D. A. (2004). An image reconstruction technique with error estimates. *Submitted to The Astrophysical Journal* .

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003). *Bayesian Data Analysis*. Chapman & Hall, London, 2nd edn.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (1996). *Markov chain Monte Carlo in Practice*. Chapman & Hall, London.

McLaughlan, G. J. and Krishnan, T. (1997). *The EM Algorithm and Extensions*. John Wiley & Sons, New York.

Meng, X.-L. and van Dyk, D. A. (1997). The EM algorithm – an old folk song sung to a fast new tune (with discussion). *Journal of the Royal Statistical Society, Series B, Methodological* **59**, 511–567.

Protassov, R., van Dyk, D. A., Connors, A., Kashyap, V., and Siemiginowska, A. (2002). Statistics: Handle with care – detecting multiple model components with the likelihood ratio test. *The Astrophysical Journal* **571**, 545–559.

Rubin, D. B. (1976). Inference and missing data. *Biometrika* **63**, 581–592.

van Dyk, D. A. (2003). Hierarchical models, data augmentation, and Markov chain Monte Carlo with discussion. In *Statistical Challenges in Modern Astronomy III* (Editors: E. Feigelson and G. Babu), 41–56. Springer–Verlag, New York.

van Dyk, D. A., Connors, A., Esch, D. N., Freeman, P., Kang, H., Karovska, M., Kashyap, V., Siemiginowska, A., and Zezas, A. (2004). Deconvolution in high energy astrophysics: Science, instrumentation, and methods. *Bayesian Analysis* submitted.

van Dyk, D. A., Connors, A., Kashyap, V., and Siemiginowska, A. (2001). Analysis of energy spectra with low photon counts via Bayesian posterior simulation. *The Astrophysical Journal* **548**, 224–243.

van Dyk, D. A. and Hans, C. M. (2002). Accounting for absorption lines in images obtained with the Chandra X-ray Observatory. In *Spatial Cluster Modelling* (Editors: D. Denison and A. Lawson), 175–198. CRC Press, London.

van Dyk, D. A. and Kang, H. (2003). Highly structured models for spectral analysis in high energy astrophysics. *Statistical Science, to appear* .

van Dyk, D. A. and Meng, X.-L. (2000). Algorithms based on data augmentation. In *Computing Science and Statistics: Proceedings of the 31st Symposium on the Interface* (Editors: M. Pourahmadi and K. Berk), 230–239. Interface Foundation of North America, Fairfax Station, VA.

van Dyk, D. A. and Meng, X.-L. (2001). The art of data augmentation. *The Journal of Computational and Graphical Statistics* **10**, 1–111.

Definition and Treatment of Systematic Uncertainties in High Energy Physics and Astrophysics

Pekka K. Sinervo
Department of Physics, University of Toronto, Toronto, ON M5S 1A7, CANADA

Systematic uncertainties in high energy physics and astrophysics are often significant contributions to the overall uncertainty in a measurement, in many cases being comparable to the statistical uncertainties. However, consistent definition and practice is elusive, as there are few formal definitions and there exists significant ambiguity in what is defined as a systematic and statistical uncertainty in a given analysis. I will describe current practice, and recommend a definition and classification of systematic uncertainties that allows one to treat these sources of uncertainty in a consistent and robust fashion. Classical and Bayesian approaches will be contrasted.

1. INTRODUCTION TO SYSTEMATIC UNCERTAINTIES

Most measurements of physical quantities in high energy physics and astrophysics involve both a statistical uncertainty and an additional "systematic" uncertainty. Systematic uncertainties play a key role in the measurement of physical quantities, as they are often of comparable scale to the statistical uncertainties. However, as I will illustrate, the definition of these two sources of uncertainty in a measurement is in practice not clearly defined, which leads to confusion and in some cases incorrect inferences. A coherent approach to systematic uncertainties is, however, possible and I will attempt to outline a framework to achieve this.

Statistical uncertainties are the result of stochastic fluctations arising from the fact that a measurement is based on a finite set of observations. Repeated measurements of the same phenomenon will therefore result in a set of observations that will differ, and the statistical uncertainty is a measure of the range of this variation. By definition, statistical variations between two identical measurements of the same phenomenon are uncorrelated, and we have well-developed theories of statistics that allow us to predict and take account of such uncertainties in measurement theory, in inference and in hypothesis testing (see, for example, [1]). Examples of statistical uncertainties include the finite resolution of an instrument, the Poisson fluctations associated with measurements involving finite sample sizes and random variations in the system one is examining.

Systematic uncertainties, on the other hand, arise from uncertainties associated with the nature of the measurement apparatus, assumptions made by the experimenter, or the model used to make inferences based on the observed data. Such uncertainties are generally correlated from one measurement to the next, and we have a limited and incomplete theoretical framework in which we can interpret and accommodate these uncertainties in inference or hypothesis testing. Common examples of systematic uncertainty include uncertainties that arise from the calibration of the measurement device, the probability of detection of a given type of interaction (often called the "acceptance" of the detector), and parameters of the model used to make inferences that themselves are not precisely known. The definition of such uncertainties is often ad hoc in a given measurement, and there are few broadly-accepted techniques to incorporate them into the process of statistical inference.

All that being said, there has been significant thought given to the practical problem of how to incorporate systematic uncertainties into a measurement. Examples of this work include proposals to combine statistical and systematic uncertainties into setting confidence limits on measurements [2–6], techniques to estimate the magnitude of systematic uncertainties [7], and the use of standard statistical techniques to take into account systematic uncertainties [8, 9]. In addition, there have been numerous papers published on the systematic uncertainties associated with a given measurement [10].

In this review, I will first discuss a few case studies that illustrate how systematic uncertainties enter into some current measurements in high energy physics and astrophysics. I will then discuss a way in which one can consistently identify and characterize systematic uncertainties. Finally, I will outline the various techniques by which statistical and systematic uncertainties can be formally treated in measurements.

2. CASE STUDIES

2.1. W Boson Cross Section: Definitions are Relative

The production of the charged intermediate vector boson, the W, in proton-antiproton ($p\bar{p}$) annihilations is predicted by the Standard Model, and the measurement of its rate is of interest in high energy physics. This measurement involves counting the number of candidate events in a sample of observed interactions,

N_c, estimating the number of "background" events in this sample from other processes, N_b, estimating the acceptance of the apparatus including all selection requirements used to define the sample of events, ϵ, and counting the number of $p\bar{p}$ annihiliations, L. The cross section for W boson production is then

$$\sigma_W = \frac{N_c - B_b}{\epsilon L}. \tag{1}$$

The CDF Collaboration at Fermilab has recently performed such a measurement [11], as illustrated in Fig. 1 where the transverse mass of a sample of candidate $W \to e\nu_e$ decays is illustrated. The measurement is quoted as

$$\sigma_W = 2.64 \pm 0.01(\text{stat}) \pm 0.18(\text{syst})\text{ nb}, \tag{2}$$

where the first uncertainty reflects the statistical uncertainty arising from the size of the candidate sample (approximately 38,000 candidates) and the second uncertainty arises from the background subtraction in Eq. (1). We can estimate these uncertainties as

$$\begin{aligned} \sigma_{stat} &= \sigma_0/\sqrt{N_c} && (3)\\ \sigma_{syst} &= \sigma_0\sqrt{\left(\frac{\delta N_b}{N_b}\right)^2 + \left(\frac{\delta\epsilon}{\epsilon}\right)^2 + \left(\frac{\delta L}{L}\right)^2}, && (4) \end{aligned}$$

where the three terms in σ_{syst} are the uncertainties arising from the background estimate δN_b, the acceptance $\delta\epsilon$ and the integrated luminosity δL. The parameter σ_0 is the measured value.

In the same sample, the experimenters also observe the production of the neutral intermediate vector boson, the Z. Because of this, the experimenters can measure the acceptance ϵ by taking a sample of Z bosons identified by the two charged electrons they decay into, and then measuring ϵ from this sample. The dominant undertainty in this measurement arises from the finite statistics in the Z boson sample. Thus, one could equivalently consider $\delta\epsilon$ to be a statistical uncertainty (and not a systematic one). This means that the uncertainties could have just as well been defined as

$$\begin{aligned} \sigma_{stat} &= \sigma_0\sqrt{1/N_c + \left(\frac{\delta\epsilon}{\epsilon}\right)^2} && (5)\\ \sigma_{syst} &= \sigma_0\sqrt{\left(\frac{\delta N_b}{N_b}\right)^2 + \left(\frac{\delta L}{L}\right)^2}. && (6) \end{aligned}$$

resulting in a different assignment of statistical and systematic uncertainties.

Why would this matter? If we return back to our original discussion of what defines a statistical and systematic uncertainty, we normally assume a systematic uncertainty is correlated with subsequent measurements and it does not scale with the sample size.

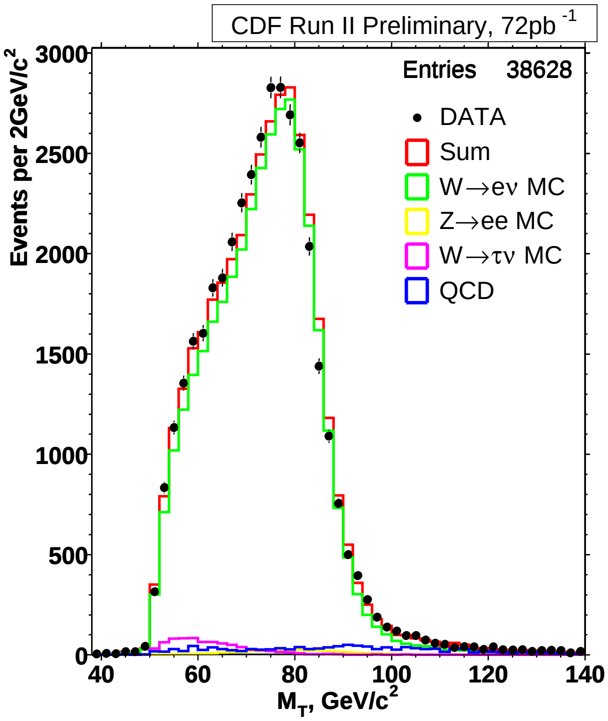


Figure 1: The transverse mass distribution for the W boson candidates as observed recently by CDF. The peak reflects the Jacobian distribution typical of W boson decays. The points are the measured distribution and the various histograms are the predicted distribution from W decays and the other background processes.

In this case, the uncertainty on ϵ does not meet these requirements. The acceptance is a stochastic variable, which will become better known with increasing Z boson sample size. It is therefore more informative to identify it as a statistical uncertainty. I will call this a "class 1" systematic uncertainty. Note that it would be appropriate to include in this category those systematic uncertainties that are in fact constrained by the result of a separate measurement, so long as the resulting uncertainty is dominated by the stochastic fluctuations in the measurement. An example of this could be the calibration constants for a detector that are defined by separate measurements using a calibration procedure whose precision is limited by statistics.

2.2. Background Uncertainty

The second case study also involves the measurement of σ_W introduced in the previous section. The estimate of the uncertainty on the background rate δN_b is performed by evaluating the magnitude of the different sources of candidate events that satisfy the criteria used to define the W boson candidate sample. In the CDF measurement, it turns out that the background is dominated by events that arise from the pro-

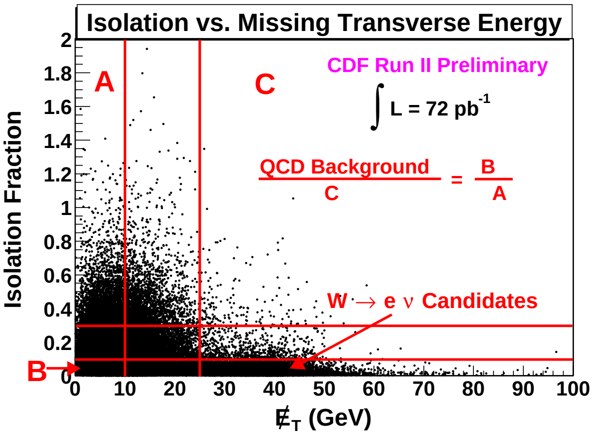


Figure 2: The distribution of the isolation fraction of the lepton candidate versus the missing energy in the event. Candidate leptons are required to have low isolation fractions (< 0.10), and QCD background events dominate the region with low missing transverse energy. The signal sample is defined by the requirement $\not{E}_T > 25$ GeV. The QCD background in the sample is estimated by the formula in the figure, which assumes that the isolation properties of the QCD events are uncorrelated with the missing transverse energy.

duction of two high-energy quarks or gluons (so-called "QCD events"), one of which "fakes" an electron or muon. A reliable estimate of this background is difficult to make from first principles, as the rate of such QCD events is many orders of magnitude larger than the W boson cross section, and the rejection power of the selection criteria is difficult to measure directly.

The technique used to estimate N_b in the CDF analysis is to take advantage of a known correlation: candidate events from QCD background will have more particles produced in proximity to the electron or muon candidate. At the same time, most of the QCD events will also have small values of missing transverse energy ($\not{E}_T$) compared with the W boson events where a high-energy neutrino escapes undetected. Thus, a measure of the isolation of the candidate lepton and the $\not{E}_T$ can be an instrument to extract an estimate of the background in the observed sample. This is shown in Fig. 2, where one sees in the bottom-right region the signal region for this analysis. The region with low missing transverse energy is populated by QCD background events.

The dominant uncertainty in the background calculation arises from the assumption that the isolation properties of the electron candidate in QCD events is uncorrelated with the missing transverse energy in the event. Any such correlation is expected to be very small, and this is consistent with other observations. However, even a small correlation in these two variables results in a bias in the estimate of N_b. This potential bias is difficult to estimate with any precision. In this case, the experimenters varied the choice of the isolation criteria to define the QCD background region and the signal region, and used the variation in the background estimate as a measure of the systematic uncertainty δN_b. This variation is shown in Fig. 3.

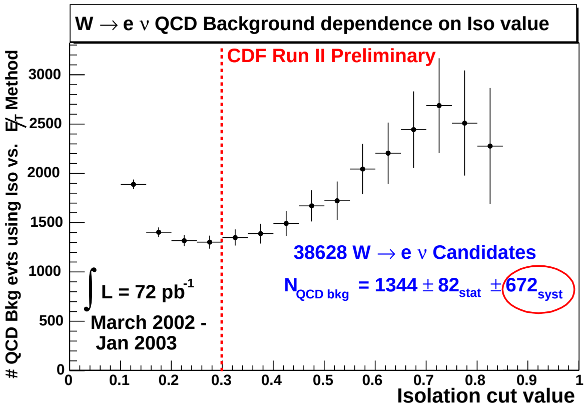


Figure 3: The variation of the background estimate as the isolation cut value is varied from 0.3 (the default value) up to 0.8. The isolation variable is a measure of the fraction of energy observed in a cone near the electron candidate normalized to the energy of the electron candidate.

This is an illustration of a systematic uncertainty that arises from one's limited knowledge of some features of the data that cannot be constrained by observations. In these cases, one often is forced to make some assumptions or approximations in the measurement procedure itself that have not been verified precisely. The magnitude of the systematic uncertainty is also difficult to estimate as it is not well-constrained by other measurements. In this sense, it differs from the class 1 systematic uncertainties introduced above.

I will therefore call this a "class 2" systematic uncertainty. It is one of the most common categories of systematic uncertainty in measurements in astrophysics and high energy physics.

2.3. Boomerang CMB Analysis

My third case study involves the analysis of the data collected by the Boomerang cosmic microwave background (CMB) probe, which mapped the spatial anisotropy of the CMB radiation over a large portion of the southern sky [12]. The data itself is a fine-grained two-dimensional plot of the spatial variation the temperature of part of the southern sky, as illustrated in Fig. 4. The analysis of this data involves the transformation of the observed spatial variation into a power series in spherical harmonics, with the spatial variations now summarized in the power spectrum as a function of the order of the spherical harmonic. The power spectrum includes all sources of uncertainty, including instrumental effects and uncertainties in calibrations.[1]

The Boomerang collaborators then test a large class of theoretical models of early universe development by determining the power spectrum predicted by each model and comparing the predicted and observed power as a function of spherical harmonic. These models are described by a set of cosmological parameters, each of them being constrained by other observations and theoretical assumptions. To determine those models that best describe the data, the experimenters take a Bayesian approach [13], creating a six-dimensional grid consisting of 6.4 million points, and calculating the likelihood function for the data at each point. They then define priors for each of the six parameters, and define a posterior probability that is now a function of these parameters. To make inferences on such key parameters as the age of the universe or its overall energy density, a marginalization is performed by numerically integrating the posterior probability over the other parameters. The experimenters can also consider the effect of varying the priors to explore the sensitivity of their conclusions to the priors themselves.

In this analysis, the lack of knowledge in the paradigm used to make inferences from the data is captured in the choice of priors for each of the parameters. A classical statistical approach could have equivalently defined these as sources of systematic uncertainty. Viewed from either perspective, the uncertainties that arise from the choice of paradigm are not statistical in nature, given that they would affect any analysis of similar data. Yet they differ from the two previous classes of systematic uncertainty I have identified, which arise directly from the measurement technique. I therefore define such theoretically-motivated uncertainties as “class 3” systematics. I also note that the Bayesian technique to incorporate these uncertainties has no well-defined frequentist analogue, in that one cannot readily identify an ensemble of experiments that would replicate the variation associated with these uncertainties.

The distinction between class 2 and class 3 systematics comes in part from the fact that one is associated with the measurement technique while the other arises in the interpretation of the observations. I argue, however, that there is an additional difference: In the first case, there is a specific piece of information needed to complete the measurement, the background yield N_b, and the systematic uncertainty arises from manner in which that is estimated. In the other case, the experiment measures the spatial variation in the CMB and summarizes these data in the multipole moments. The systematic uncertainties that are associated with the subsequent analysis of these data in terms of cosmological parameters is very model-dependent, and the systematic uncertainties arise from the attempt to extract information about a subset of the parameters in the theory (for example, the age of the universe or the energy density).

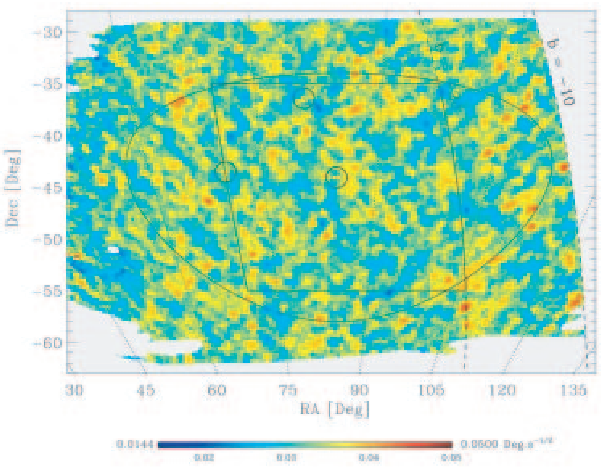


Figure 4: The temperature variation of the CMB as measured by the Boomerang experiment. The axes represent the declination and azimuth of the sky, and the contour is the region used in the analysis.

2.4. Summary of Taxonomy

In these case studies, I have motivated three classes of systematic uncertainties. Class 1 systematics are uncertainties that can be constrained by ancillary measurements and can therefore be treated as statistical uncertainties. Class 2 systematics arise from model assumptions in the measurement or from poorly understood features of the data or analysis technique that introduce a potential bias in the experimental outcome. Class 3 systematics arise from uncertainties in the underlying theoretical paradigm used to make inferences using the data.

The advantages of this taxonomy are several. Class 1 systematics are statistical in nature and will therefore naturally scale with the sample size. I recommend that they be properly considered a statistical uncertainty and quoted in that manner. They are not correlated with independent measurements and are therefore straightforward to handle when combining measurements or making inferences. Class 2 systematics are the more challenging category, as they genuinely reflect some lack of knowledge or uncertainty in the model used to analyze the data. Because of this, they also have correlations that should be understood in any attempt to combine the measurement with other

[1] The uncertainties associated with instrumental effects and calibrations are also systematic in nature, but we will not focus on these here.

observations. They also do not scale with the sample size, and therefore may be fundamental limits on how well one can perform the measurement. Class 3 systematics do not depend on how well we understand the measurement per se, but are fundamentally tied to the theoretical model or hypothesis being tested. As such, there is significant variation in practice. As is illustrated in the third case study and as I will discuss below, a Bayesian approach allows for a range of possible models to be tested if one can parametrize the uncertainties in the relevant probability distribution function and then define reasonable priors. A purely frequentist approach to this problem founders on how one would define the relevant ensemble.

3. ESTIMATION OF SYSTEMATIC UNCERTAINTIES

There is little, if any, formal guidance in the literature for how to define systematic uncertainties or estimate their magnitudes, and much of current practice has been defined by informal convention and "oral tradition." A fundamental principle, however, is that the technique used to define and estimate a systematic uncertainty should be consistent with how the statistical uncertainties are defined in a given measurement, since the two sources of uncertainty are often combined in some way when the measurement is compared with theoretical predictions or independent measurements.

Perhaps the most challenging aspect of estimating systematic uncertainties is to define in a consistent manner all the relevant sources of systematic uncertainty. This requires a comprehensive understanding of the nature of the measurement, the assumptions implicit or explicit in the measurement process, and the uncertainties and assumptions used in any theoretical models used to interpret the data. In any robust design of an experiment, the experimenters will anticipate all sources of systematic uncertainty and should design the measurement to minimize or constrain them appropriately. Good practice suggests that the analysis of systematic uncertainties should be based on clear hypotheses or models with well-defined assumptions.

In the process of the measurement, it is often typical to make various "cross-checks" and tests to determine that no unanticipated source of systematic uncertainty has crept into the measurement. A cross-check, however, should not be construed as a source of systematic uncertainty (see, for example the discussion in [7]).

A common technique for estimating the magnitude of systematic uncertainties is to determine the maximum variation in the measurement, Δ, associated with the given source of systematic uncertainty. Arguments are then made to transform that into a measure that corresponds to a one standard deviation measure that one would associate with a Gaussian statistic, with typical conversions being $\Delta/2$ and $\Delta/\sqrt{12}$, the former being argued as a deliberate overestimate, and the latter being motivated by the assumption that the actual bias arising from the systematic uncertainty could be anywhere within the interval Δ. Since it is common in astrophysics and high energy physics to quote 68% confidence level intervals as statistical uncertainties, it therefore is appropriate to estimate systematic uncertainties in a comparable manner.

There are various practices that tend to overestimate the magnitude of systematic uncertainties, and these should be avoided if one is to not dilute the statistical power of the measurement. A common mistake is to estimate the magnitude of a systematic uncertainty by using a shift in the measured quantity when some assumption is varied in the analysis technique by what is considered the relevant one standard deviation interval. The problem with this approach is that often the variation that is observed is dominated by the statistical uncertainty in the measurement, and any potential systematic bias is therefore obscured. In such cases, I recommend that either a more accurate procedure be found to estimate the systematic uncertainty, or at the very least that one recognize that this estimate is unreliable and likely to be an overestimate. A second common mistake is to introduce a systematic uncertainty into the measurement without an underlying hypothesis to justify the concern. This is often the result of confusing a source of systematic uncertainty with a "cross check" of the measurement.

4. THE STATISTICS OF SYSTEMATIC UNCERTAINTIES

A reasonable goal in any treatment of systematic uncertainties is that consistent and well-established procedures be used that allow one to understand how to best use the information embedded in the systematic uncertainty when interpreting the measurement. Increasingly, the fields of astrophysics and high energy physics have developed more sophisticated approaches to interval estimation and hypothesis testing. Frequentist approaches have returned to the fundamentals of Neyman constructions and the resulting coverage properties. Bayesian approaches have explored the implications of both objective and subjective priors, the nature of inference and the intrinsic power embedded in such approaches when combining information from multiple measurements.

I will outline how systematic uncertainties can be accommodated formally in both Bayesian and frequentist approaches.

4.1. Formal Statement

To formally state the problem, assume we have a set of observations $x_i, i = 1, n$, with an associated probability distribution function $p(x_i|\theta)$, where θ is an unknown random parameter. Typically, we wish to make inferences about θ. Let us now assume that there is some additional uncertainty in the probability distribution function that can be described with another unknown parameter λ. This allows us to define a likelihood function

$$\mathcal{L}(\theta, \lambda) = \prod_i p(x_i|\theta, \lambda). \qquad (7)$$

Formally, one can treat λ as a "nuisance parameter." In many cases (especially those associated with class 1 systematics), one can identify a set of additional observations of a random statistic $y_j, j = 1, m$ that provides information about λ. In that case, the likelihood would become

$$\mathcal{L}(\theta, \lambda) = \prod_{i,j} p(x_i, y_j|\theta, \lambda). \qquad (8)$$

With this formulation, one sees that one has to find a means of taking into account the uncertainties that arise from the presence of λ in order to make inferences on θ. I will discuss some possible approaches.

4.2. Bayesian Approach

A Bayesian approach would involve identification of a prior, $\pi(\lambda)$, that characterizes our knowledge of λ. Typical practice has been to either assume a flat prior or, in cases where there are corollary measurements that give us information on λ, a Gaussian distribution. One can then define a Bayesian posterior probability distribution

$$\mathcal{L}(\theta, \lambda)\ \pi(\lambda)\, d\theta d\lambda, \qquad (9)$$

which we can then marginalize to set Bayesian credibility intervals on θ.

This is a straightforward statistical approach and results in interval estimates that can readily be interpreted in the Bayesian context. The usual issues regarding the choice of priors remains, as does the interpretation of a Bayesian credibility interval. These are beyond the scope of this discussion, but are covered in most reviews of this approach [13].

4.3. Frequentist Approach

The frequentist approach to the formal problem also starts with the joint probability distribution $p(x_i, y_j|\theta, \lambda)$. There are various techniques for how to deal with the presence of the nuisance parameter λ, and I will outline just a few of them. I will note that there isn't a single commonly adopted strategy in the literature, and even the simplest techniques tend to involve significant computational burden.

One technique involves identifying a transformation of the parameters to factorize the problem in such a manner that one can then integrate out one of the two parameters [14]. This approach is robust and theoretically sound, and in the trivial cases results in a 1-dimensional likelihood function that now incorporates the uncertainties arising from the nuisance parameter. It has well-defined coverage properties and a clear frequentist interpretation. However, this approach is of limited value given that it is necessary to find an appropriate transformation.

I note that this approach is only of value in cases where one is dealing with a class 1 systematic uncertainty that is, as I have argued above, formally a source of statistical uncertainty. Class 2 and class 3 systematic uncertainties cannot be readily constrained by a set of observations represented by the $y_j, j = 1, m$.

A second approach to the incorporation of nuisance parameters is to define Neyman "volumes" in the multi-dimensional parameter space, equivalent to what is done in the case of a interval setting with one random parameter. In this case, one creates an infinite set of two dimensional contours defined by requiring that the observed values lie within the contour the necessary fraction of the time (say 68%). Then one identifies the locus of points in this two-dimensional space defined by the centres of each contours, and this boundary becomes the multi-dimensional Neyman interval for both parameters, as illustrated in Fig. 5. To "eliminate" the nuisance parameter, one projects the two-dimensional contour onto the axis of the parameter of interest. This procedure results in a frequentist confidence interval that over-covers, and in some cases over-covers badly. It thus results in "conservative" intervals that may diminish the statistical power of the measurement.

A third technique to take into account systematic uncertainties involves what is commonly called the "profile method" where one eliminates the nuisance parameter by creating a profile likelihood defined as the value of the likelihood maximized by varying λ for each value of the parameter θ [15]. This creates a likelihood function independent of λ, but one that has ill-defined coverage properties that depend on the correlation between λ and θ. However, it is a straightforward technique, that is used frequently and that results in inferences that are at some level conservative.

A variation on these techniques has been used in recent analyses of solar neutrino data, where the analysis uses 81 observables and characterizes the various systematic uncertainties by the introduction of 31 parameters [16]. They linearize the effects of the systematic parameters on the chi-squared function and then minimize the chi-squared with respect to each of the

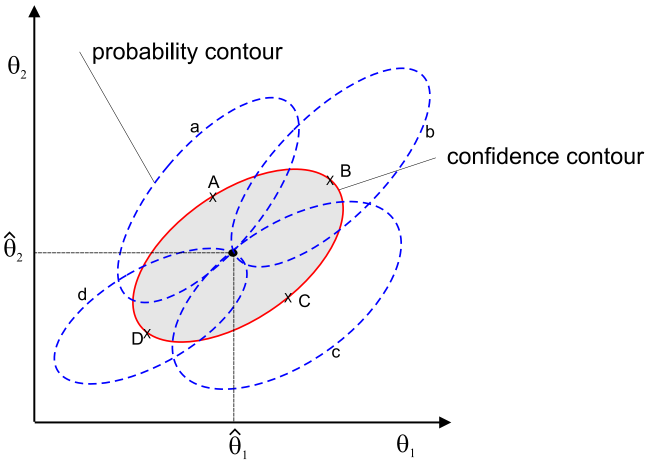


Figure 5: The Neyman construction for two parameters. The nuisance parameter is θ_2, and the shaded region is the interval defined by this construction. The four dashed contours are examples of the intervals that define the contour (from G. Zech).

parameters. In this sense, this analysis is a concrete example of the profile method.

Although most of these techniques are approximate in some sense, they have the virtue that they scale in an intuitively acceptable manner. In the limit where the statistical uncertainties are far larger than the systematic uncertainties, the former are the dominant effect in any inference or hypothesis test. Conversely, when the systematic uncertainties begin to compete with or dominate the statistical uncertainties, the results of any statistical inference reflect the systematic uncertainties and these become the limiting factor in extracting information from the measurement. I would argue that this should be a minimal requirement of any procedure used to take into account systematic uncertainties.

4.4. Hybrid Techniques

There has been one technique in common use in high energy physics to incorporate sources of systematic uncertainty into an analysis, first described by R. Cousins and V. Highland [2]. Using the notation introduced earlier, the authors argue that one should create a modified probability distribution function

$$p_{CH}(x|\theta) = \int p(x|\theta,\lambda)\pi(\lambda)\,d\lambda, \qquad (10)$$

which could be used to define a likelihood function and make inferences on θ. They argue that this can be understood as approximating the effects of having an ensemble of experiments each of them with various choices of the parameter λ and with the distribution $\pi(\lambda)$ representing the frequency distribution of λ in this ensemble.

Although intuitively appealing to a physicist, this approach does not correspond to either a truly frequentist or Bayesian technique. On the one hand, the concept of an ensemble is a frequentist construct. On the other hand, the concept of integrating or "averaging" over the probability distribution function is a Bayesian approach. Because of this latter step, it is difficult to define the coverage of this process [5]. I therefore consider it a Bayesian technique that can be readily understood in that formulation if one treats the frequency distribution $\pi(\lambda)$ as the prior for λ. I note that it also has the desired property of scaling correctly as one varies the relative sizes of the statistical and systematic uncertainties.

5. SUMMARY AND CONCLUSIONS

The identification and treatment of systematic uncertainties is becoming increasingly "systematic" in high energy physics and astrophysics. In both fields, there is a recognition of the importance of systematic uncertainties in a given measurement, and techniques have been adopted that result in systematic uncertainties that can be compared in some physically relevant sense with the statistical uncertainties.

I have proposed that systematic uncertainties can be classified into three broad categories, and by doing so creating more clarity and consistency in their treatment from one measurement to the next. Such classification, done a priori when the experiment is being defined, will assist in optimizing the experimental design and introducing into the data analysis the necessary approaches to control and minimize the effect of these systematic effects. In particular, one should not confuse systematic uncertainties with cross-checks of the results.

Bayesian statistics naturally allow us to incorporate systematic uncertainties into the statistical analysis by introducing priors for each of the parameters associated with the sources of systematic uncertainty. However, one must be careful regarding the choice of prior. I recommend that in all cases the sensitivity of any inference or hypothesis test to the choice of prior be investigated in order to ensure that the conclusions are robust.

Frequentist approaches to systematic uncertainties are less well-understood. The fundamental problem is how one defines the concept of an ensemble of measurements, when in fact what is varying is not an outcome of a measurement but ones assumptions concerning the measurement process or the underlying theory. I am not aware of a robust method of incorporating systematic uncertainties in a frequentist paradigm except in cases where the systematic uncertainty is really a statistical uncertainty and the additional variable can be treated as a nuisance parameter. However, the procedures commonly used to incorporate systematic uncertainties into frequentist statistical inference do have some of the desired "scaling" properties.

Acknowledgments

I gratefully acknowledge the support of the Natural Sciences and Engineering Research Council of Canada, and thank the conference organizers for their gracious hospitality.

References

[1] A. Stuart and J. K. Ord, *Kendall's Advanced Theory of Statistics,* Oxford University Press, New York (1991).

[2] R. D. Cousins and V. L. Highland, Nucl. Instrum. Meth. **A320**, 331 (1992).

[3] C. Guinti, Phys. Rev. D **59**, 113009 (1999).

[4] M. Corradi, "Inclusion of Systematic Uncertainties in Upper Limits and Hypothesis Tests," CERN-OPEN-2000-213 (Aug 2000).

[5] J. Conrad *et al.*, Phys. Rev. D **67**, 012002 (2003).

[6] G. C. Hill, Phys. Rev. D **67**, 118101 (2003).

[7] R. J. Barlow, "Systematic Errors, Fact and Fiction," hep-ex/0207026 (Jun 2002). Published in the Proceedings of the Conference on Advanced Statistical Techniques in Particle Physics, Durham England, Mar 16-22, 2002. See also, R. J. Barlow, "Asymmetric Systematic Errors," hep-ph/0306138 (Jun 2003).

[8] L. Demortier, " Bayesian Treatment of Systematic Uncertainties," Published in the Proceedings of the Conference on Advanced Statistical Techniques in Particle Physics, Durham England, Mar 16-22, 2002.

[9] G. Zech, Eur. Phys. J **C4:12** (2002).

[10] See, for example: A. G. Kim *et al.*, "Effects of Systematic Uncertainties on the Determination of Cosmological Parameters," astro-ph/0304509 (Apr 2003).

[11] T. Dorigo *et al.* (The CDF Collaboration), FERMILAB-CONF-03/193-E. Published in the Proceedings of the 38th Rencontres de Moriond on QCD and High-Energy Hadronic Interactions, Les Arcs, Savoie, France, March 22-29, 2003.

[12] B. Netterfield *et al.*, Ap. J. **571**, 604 (2002).

[13] E. T. Jaynes, *Probability Theory: The Logic of Science*, Cambridge University Press (1995).

[14] G. Punzi, "Including Systematic Uncertainties in Confidence Limits," CDF Note in preparation (Jul 2003).

[15] See, for example, N. Reid, in these proceedings.

[16] Fogli *et al.*, Phys. Rev. D **66**, 053010 (2002).

Evaluating And Constructing Features For Identification Of Tau Leptons

R. Vilalta, A. Bagherjeiran, C. Sun
University of Houston, 4800 Calhoun Rd., Houston TX 77204-3010, USA
B. P. Padley, S. J. Lee
Bonner Nuclear Lab, Rice University, 6100 Main Street, Houston, TX 77005, USA

In this paper we show the importance of choosing the right feature representation in attempting to improve the quality of a predictive model. We explain how to evaluate and construct new features using information-theoretic measures (information gain, gain ratio) and statistical tests (e.g., χ^2, G statistic). Our experiments use Monte-Carlo simulated data containing both τ lepton signals and background events. Results show how our evaluation process can identify a small set of relevant features that bear correlation with the class (τ signals). We also show how to construct new features by exploring the space of logical feature combinations using genetic algorithms; the set of newly constructed features can effectively improve the quality of the feature representation.

1. INTRODUCTION

The aim of this study is to construct good predictive models for the identification of τ leptons; we wish to identify clusters of energy obtained from particle detectors associated with jets that are characteristic of τ leptons (e.g., tightly collimated jets of energy). The problem is complex because the distribution of energy corresponding to τ decay overlaps with that corresponding to the fragmentation of quarks [8]. Our approach is to use multivariate data analysis techniques for classification to separate τ leptons from background events.

From a pattern classification view [2], searching for a good predictive model can be attained following two different paths. The most common path is to employ various multivariate data analysis techniques (e.g., neural network, decision tree, support vector machine), and to compare them by assessing their model performance (e.g., off-training set accuracy) measured via some re-sampling technique, such as 10-fold cross-validation. The most accurate model is then used for prediction. A second path is to keep the multivariate analysis technique fixed and instead work on improving the feature representation, by either selecting the most relevant features (e.g. calorimeter cluster parameters, number of tracks associated with the event, mass of τ tracks, etc.), or by constructing new features through a search in the space of possible feature combinations.

Our study follows the second path described above in the identification of τ leptons. This step is particularly important because most classification algorithms are highly sensitive to the quality of the feature representation. The presence of irrelevant features and/or features interacting in complex ways demands learning machines with low bias or high capacity (i.e., high flexibility in the decision boundaries), to capture the complex data distribution, at the expense of increasing the variance component in error [3, 5]. By selecting the most relevant features and joining together highly interacting features the data distribution is transformed to a form amenable to current classification techniques.

Moreover, experiments in particle physics often embed a complex characterization of event signals where evaluating and constructing new features can become highly instrumental. In general, attaining accurate classifiers depends to a great extent on the quality of the feature set characterizing the system under study. On the one side, high quality features convey much information about the system; in this case, a simple classifier suffices to produce good results. In contrast, complex features must be combined with many other features to unveil system structure; here features can interact in many ways and identifying the most relevant features is needed to discover important combinations.

2. MULTIVARIATE DATA ANALYSIS FOR CLASSIFICATION

We begin by giving a brief overview of the classification problem. We assume an n-component vector-valued random variable, $(A_1, A_2, \cdots, A_n)$, where each A_i represents an attribute or feature; the space of all possible attribute vectors is called the input space $\mathcal{X}$. Let $\{y_1, y_2, \cdots, y_k\}$ be the possible classes, categories, or states of nature; the space of all possible classes is called the output space $\mathcal{Y}$. A classifier receives as input a set of training examples $T = \{(\mathbf{x}, y)\}$, where $\mathbf{x} = (a_1, a_2, \cdots, a_n)$ is a vector or point in the input space and y is a point in the output space. We assume T consists of independently and identically distributed (i.i.d.) examples obtained according to a fixed but unknown joint probability distribution ϕ in the input-output space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. The outcome of the classifier is a function h (or hypothesis) mapping the input space to the output space, $h : \mathcal{X} \rightarrow \mathcal{Y}$. Function h can then be used to predict the class of previously unseen

	A	A'	
S	n_1^1	n_0^1	n_1
S'	n_1^0	n_0^0	n_0
	n^1	n^0	N

Probabilities:

For S and S': $P^1 = \frac{n^1}{N}$ $P^0 = \frac{n^0}{N}$

For A: $P_1^1 = \frac{n_1^1}{n_1}$ $P_1^0 = \frac{n_1^0}{n_1}$ $P_1 = \frac{n_1}{N}$

For A': $P_0^1 = \frac{n_0^1}{n_0}$ $P_0^0 = \frac{n_0^0}{n_0}$ $P_0 = \frac{n_0}{N}$

Figure 1: Cross-classification of feature values and classes with probabilities estimated from the data.

attribute vectors.

In our study each feature vector $\mathbf{x}$ stands for a jet of energy flow, characterized by features such as calorimeter cluster parameters, number of tracks associated with the event, mass of τ tracks, etc. The class variable is binary-valued: events either belong to a τ signal or a background event. Our problem formulation can be exploited by any classification algorithm including neural networks, decision trees, support-vector machines, etc. Rather than focusing on the classification algorithm, however, we focus on the feature representation as described next.

3. FEATURE EVALUATION

We first address the problem of selecting the most relevant features. To assess the value of potentially useful features we need an evaluation metric. Formally, an evaluation metric M is used to quantify the quality of the partitions induced by a feature A over a training set T, where $|T| = N$. For simplicity assume feature A is binary-valued, such that it divides T in two sets: $\{(\mathbf{x}, y) \mid A(\mathbf{x}) = 1\}$ and $\{(\mathbf{x}, y) \mid A(\mathbf{x}) = 0\}$; we say the former set is covered by A, whereas the latter set is covered by the complement A'. Similarly, set T can be divided according to the class label on each example; we assume two class values: 1 (for signal S) and 0 (for background S'). Figure 1 shows the cross-classification of classes and values of A. Let n^1 and n^0 be the number of examples in T of class 1 and 0 respectively, where $n^1 + n^0 = N$. Let n_1^1 and n_1^0 be the number of examples covered by A of class 1 and 0 respectively, such that $n_1^1 + n_1^0 = n_1$, and let n_0^1 and n_0^0 represent the corresponding numbers in A', such that $n_0^1 + n_0^0 = n_0$. In addition, Figure 1 defines probabilities as estimated from the data.

The quality of the partitions made by A is simply determined by the class-uniformity or purity of such partitions. Good attributes tend to split dataset T into subsets that are class-uniform. The following are traditional definitions of evaluation metrics where the goal is to maximize the output value.

Information Gain (IG) [11]

Let entropy $\mathrm{H}(x, y) = -x\log_2(x) - y\log_2(y)$

$$\mathrm{IG}(A) = \mathrm{H}(P^1, P^0) - \sum_{i=0}^{1}(P_i \ \mathrm{H}(P_i^1, P_i^0)) \qquad (1)$$

Gain Ratio (GR) [11]

$$\mathrm{GR}(A) = \frac{\mathrm{IG}(A)}{\mathrm{H}(P_0, P_1)} \qquad (2)$$

G Statistic (G) [13]

$$\mathrm{G}(A) = 2N \ \mathrm{IG}(A) \ \log_e 2 \qquad (3)$$

$\boldsymbol{\chi^2}$ [13]

$$\chi^2(A) = \frac{N \ (n_1^1 n_0^0 - n_0^1 n_1^0)^2}{n^1 n^0 n_1 n_0} \qquad (4)$$

Traditional evaluation metrics as the ones described above define the degree of class-uniformity in each new partition using the proportion of classes on the example subsets; the best result is attained if each example subset is class uniform.

To gain a better understanding of the nature of evaluation metrics, note each metric M is a function of the number of examples covered by feature A and of its complement A', $M : f(n_1^1, n_1^0, n_0^1, n_0^0)$ (Figure 1). Alternatively M could be defined as a function of the coverage of A and of the coverage of the whole set T, $f(n_1^1, n_1^0, n^1, n^0)$, since $n^1 = n_0^1 + n_1^1$ and $n^0 = n_0^0 + n_1^0$. For a given learning problem, n^1 and n^0 are fixed; by considering them as constants, we can simply express M as $f(n_1^1, n_1^0)$. For simplicity let's rename n_1^1 and n_1^0 as p and n (the positive –or signal– and negative –or background– examples covered by A), such that $M : f(p, n)$. Metric M extends above the plane defined by these two variables. We define this plane as the *coverage plane*. Each of the metrics defined above can be plotted above a coverage plane bounded by the total examples of class 1 and class 0 in T, n^1 and n^0. As an example, Figure 2 plots Information Gain when the number of positive and negative examples in T is the same ($n^1 = n^0 = 100$); each point (p, n) is evaluated according to Equation 1. The fact that the value of $f(p, n)$ takes into account both the coverage of A

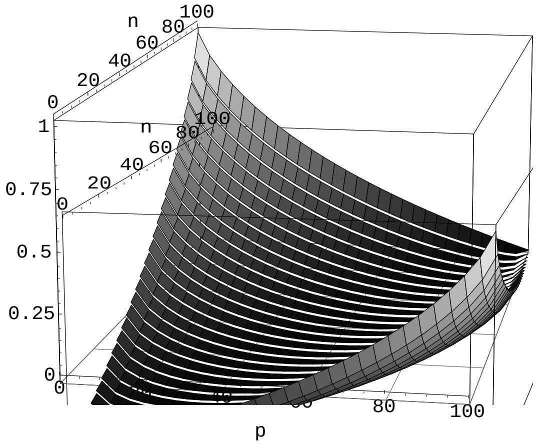


Figure 2: Information Gain as a function of the possible coverage (number of examples of class 1 and class 0) of a feature.

and of its complement A' is reflected by the symmetry about the axis-line $((0,0),(100,100))$. The maximum values are attained at the extreme points $(100,0)$ and $(0,100)$, when the induced example subsets are class uniform [12].

Although our focus here is on traditional metrics, the reader should be aware of additional families of metrics that quantify the ability of a feature to separate away examples of different class [6, 7, 9]. These metrics deserve particular attention because of their ability to address the high interaction problem, in which the relevance of a feature can be observed only in combination with other features.

3.1. Empirical Results

Our first set of experiments rank all available features using information gain (Equation 1) as the evaluation metric. We use Monte-Carlo simulated data with a skewed class distribution (about 5% of events belong to background and the rest to signal). Each feature is first discretized into intervals [1]. Out of twenty one features, each considered separately, only six produce information gain above 0.1. Figure 3 shows histograms corresponding to the posterior class distributions conditioned on the best two features. The first feature, *tautz* (Fig. 3, top), is the coordinate value in the direction of the beam (z-coordinate) of the most energetic daughter particle produced from a τ lepton decay at its closest approach to the z-axis. The second feature, *tauiso* (Fig. 3, bottom), is the normalized transverse component of τ lepton momentum measured based on the distribution of electromagnetic calorimeter cells. Though some overlap is present, both features effectively discriminate between both classes.

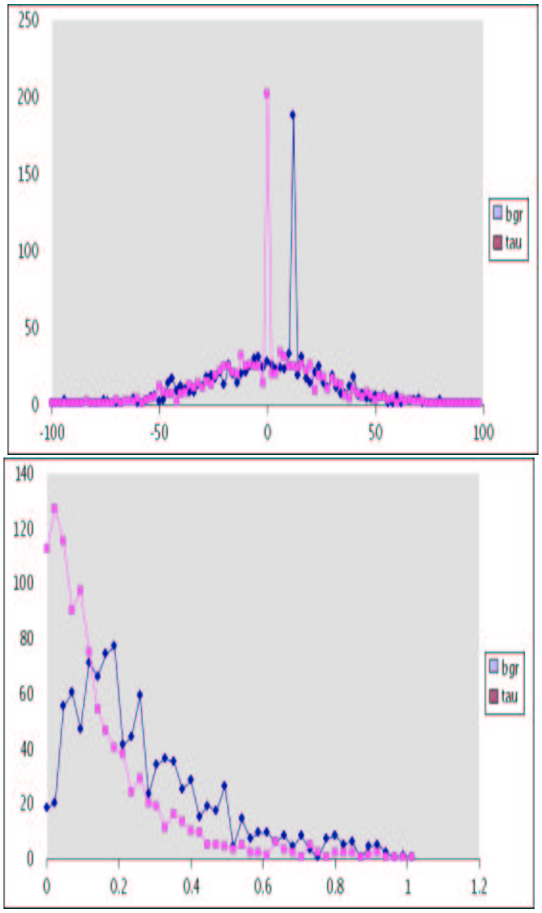


Figure 3: Posterior class distributions conditioned on the two best features, *tautz* (top) and *tauiso* (bottom).

Reducing the number of features is not only useful to speed up the training phase, but also eliminates noise. Additional experiments compare the accuracy of several classifiers using the top six features; our results show no significant difference in accuracy between these results and those obtained using all available features. We conclude feature evaluation can be very useful in identifying relevant features when discriminating signal events from background events in particle physics.

4. FEATURE CONSTRUCTION

A second approach to improve the data representation is to construct new features that could potentially capture important dependencies among original features. To begin, let us assume our problem characterization is made of Boolean features. This is attained by dividing numeric features into intervals [1], and by mapping each nominal value into a Boolean feature. One approach to feature construction is to build new features as the logical combination of Boolean features or their complements (e.g., $\bar{a_1} \wedge a_2$).

Our algorithm conducts a search over the space of all logical combinations of features. Each combination can be evaluated using any of the metrics described in Section 3. Since the size of the search space is frequently computationally intractable, we used genetic algorithms to find the highest-ranked feature combinations.

Genetic algorithms operate in a simple fashion: they work iteratively on a population of individuals or candidate solutions with highest fitness value. At each iteration individuals are ranked based on their fitness value (in our case best score output by the evaluation metric, e.g., information gain) and a new population is produced by applying genetic operators on individuals selected probabilistically. The process continues until an individual is found with a fitness value above a predefined threshold. Genetic algorithms are search mechanisms amenable to parallelization, effective in finding solutions in complex spaces [4, 10].

4.1. Empirical Results

Our second set of experiments use genetic algorithms to explore the space of logical feature combinations. We observe best results when using Gain Ratio as the evaluation metric (Equation 2). To assess the utility of our results we compare the accuracy of a decision tree with and without the new constructed features. We observe a significant difference in accuracy between the two resulting hypotheses, supporting our claim that constructing new features can improve the original data representation by capturing dependencies among attributes.

5. CONCLUSIONS

Choosing the right feature representation can influence the quality of the predictive model with equal or higher impact than choosing the right classification algorithm. This is because classification algorithms are highly sensitive to the quality of the feature representation. In this paper we show how feature evaluation can be used to systematically rank features according to their correlation with the class under prediction (i.e., their correlation with τ lepton signals).

Our experiments use Monte-Carlo simulated data with two types of events: τ lepton signals and background events. Our evaluation process shows how only a few features are necessary to produce a classifier. Such reduction in the size of the feature set reduces the time to train the classifier and often results on an improvement in the accuracy of the final hypothesis.

We also show how to automatically construct new features by exploring the space of logical feature combinations using genetic algorithms. Our results show an improvement in predictive accuracy when the newly constructed features are integrated into the pool of original features. Although feature construction can become computationally expensive, the resulting combinations may point to interesting relations about the physical processes involved in particle collisions and decay.

Acknowledgments

We thank the DØ collaboration for sharing the Monte-Carlo simulated data used in our experiments.

References

[1] J. Catlett (1991). “On Changing Continuous Attributes Into Ordered Discrete Attributes”, *Proceedings of the European Conference on Machine Learning* pp. 164-178. Springer-Verlag.

[2] R. O. Duda, P. E. Hart, and D. G. Stork (2001). *“Pattern Classification”*, John Wiley Ed. 2nd Edition.

[3] S. Geman, E. Bienenstock, and R. Doursat (1992). “Neural Networks and the Bias-Variance Dilemma”, *Neural Computation*, 4, pp. 1-58.

[4] D. Goldberg (1989). *“Genetic Algorithms in Search, Optimization, and Machine Learning”*, Addison Wesley.

[5] T. Hastie, R. Tibshirani, and J. Friedman (2001). *“The Elements of Statistical Learning, Data Mining, Inference, and Prediction”*, Springer-Verlag.

[6] S. J. Hong (1997). “Use of Contextual Information for Feature Ranking and Discretization”, *IEEE Transactions of Knowledge and Data Engineering.*

[7] K. Kira and L. Rendell (1992). “A Practical Aapproach to Feature Selection”, *Ninth International Workshop on Machine Learning*, pp. 249-256, Morgan Kaufmann.

[8] B. Knuteson and P. Padley (2003). “Statistical Challenges with Massive Data Sets in Particle Physics”, *Unpublished Manuscript.*

[9] I. Kononenko and S.J. Hong (1997). “Attribute Selection for Modeling”, *Future Generation Computer Systems.*

[10] T. Mitchell (1997). “*Machine Learing*”, McGraw-Hill.

[11] J. R. Quinlan (1994). *“Programs for Machine Learning”*, Morgan Kaufmann, San Francisco.

[12] R. Vilalta and D. Oblinger (2000). “A Quantification of Distance-Bias Between Evaluation Metrics in Classification”, *Proceedings of the 17th International Conference on Machine Learning*, pp. 1087–1094, Morgan Kaufman.

[13] A.P. White and W. Z. Liu (1994). “Bias in Information-Based Measures in Decision Tree Induction”, *Machine Learning*, 15, pp. 321-329, Kluwer, Boston, MA.

Application of Adaptive Mixtures and Fractal Dimension Analysis Technique to Particle Physics

S.J. Lee, P. Padley, B. Chase
Rice University, Houston, TX 77005, U.S.A.

In this paper, we examine the applicability of using Adaptive Mixtures to represent high energy physics data. We attempt to use Adaptive Mixtures to derive a discriminant variable to identify tau leptons in hadron collider data. In addition, we examine the applicability of Fractal Dimensions as a tool to search for physics signals.

1. INTRODUCTION

The discrimination of physics "signal" from "background" is one of the most important subjects in high energy physics analysis since this process usually governs the magnitude of measurement errors. Background suppression using kernel density estimation to estimate the parent distribution of a data sample appears to be an effective method. In this paper, Adaptive Mixtures [1] and Kernel Density Estimation with Likelihood Maximization (KDELM) are examined. In addition, the Fractal Dimension [2] of a data set is studied as a tool to find physics signals.

Adaptive Mixtures are designed to accept the strengths of the kernel estimates and that of the finite mixtures while discarding the weaknesses of these methods. In the kernel estimation method, one kernel is assigned to each datum. As a result, kernel density estimates converge asymptotically under very weak conditions. Due to the large number of kernels, however, there is computational disadvantage in the case that there are many data points. On the other hand, since finite mixture methods employ small fixed numbers of kernels, these approaches have an advantage in calculation time. However, very strong assumptions must be put on the underlying densities as well as on the initial states of kernels. In the adaptive mixture, the number of kernels is driven by data. At each data point, criteria for making the decision whether a new kernel is added are tested. When these criteria are satisfied, a new kernel is added. Otherwise, the parameters of the adaptive mixture kernels are updated to accommodate the new data point. In this way, only the necessary number of kernels are introduced in the fit while the characteristic of asymptotic convergence remains without strong assumptions on the underlying distributions.

KDELM uses the maximization of likelihood. The main difference between KDELM and Adaptive Mixtures is that, in KDELM, all data points are taken into account at the same time to calculate the likelihood of the kernel estimates. Only when addition of a kernel reduces the overall likelihood is the new kernel added to the kernel estimates. If the new kernel produces a worse overall likelihood, the process adding new kernels is ceased.

The fractal dimension [2] D, also called capacity dimension, is defined by

$$n(\epsilon) = \epsilon^{-D}, \tag{1}$$

where $n(\epsilon)$ is the minimum number of open sets of diameter ϵ a to cover the set. Fractal dimension quantifies the increase in structural definition that magnification yields. As an example, a good estimation of the length of a coast line can be obtained using a meter stick. However, if a centimeter stick is used in this measurement, a better and larger estimation of the coast line length will be obtained. Fractal dimension quantifies this increase in detail that occurs by magnifying or, in this case, by switching rulers. The motivation to use the fractal dimension in the signal discrimination process is that the fractal dimension would efficiently identify signal from background if the geometrical configurations of signal and background in the d-dimensional space are significantly different. Furthermore, since the calculation of fractal dimension does not contain a complicated logical algorithm, the fractal dimension calculation has an advantage in calculation time.

2. METHODS

2.1. Adaptive Mixtures

The kernel density estimate K consists of N_K kernels, and each kernel K_i has a weight w_i. In d-dimensional space, a kernel K_i is a multivariate product of d univariate Gaussians. As a result, K_i is characterized by d-dimensional mean and variance, μ_i and σ_i, respectively. The kernel density estimate K at a d-dimensional data point x is represented by

$$K(x) = \sum_{i=1}^{N_K} w_i K_i(x; \mu_i, \sigma_i). \tag{2}$$

In Adaptive Mixtures, the number of kernels is driven by data. The kernel density estimate K^n found by the data points $x_1, x_2, ..., x_n$ is given by

$$K^n = \sum_{i=1}^{N_K^n} w_i^n K_i(\mu_i^n, \sigma_i^n). \tag{3}$$

A new kernel is added at the next data point x_{n+1} when the Mahalanobis distance from the new data point to each kernel is greater than a predefined threshold T_c. If this criterion is not satisfied, then w_i^n, μ_i^n and σ_i^n are updated to w_i^{n+1}, μ_i^{n+1} and σ_i^{n+1} without addition of a new kernel. The update and creation rules are specified in reference [1]. The Mahalanobis distance M between a one-dimensional data point x and a kernel with mean μ and standard deviation σ is defined by

$$M = \frac{(x-\mu)^2}{\sigma^2}. \tag{4}$$

In this study, the threshold T_c is set to be 9.

2.2. Kernel Density Estimation with Likelihood Maximization

In this method, addition of a new kernel takes place only when the addition produces better fit result. The goodness of fit is estimated by examining the log-likelihood given by

$$\log\mathcal{L} = \sum_i K(x_i), \tag{5}$$

where the sum runs over data points and the kernel estimate $K(x_i)$ is defined in equation 2. The constraint that the kernel weights sums to one is applied.

2.3. Fractal Dimension Calculation

There are several techniques to calculate fractal dimension, yet all involve estimating the dimension from the slope of a log-log power law point. The technique used in this study is the box counting technique [2]. The specific process for the fractal dimension calculation is:

1. Grids or boxes of varying side lengths are placed over data sample.

2. A count of how many boxes contain data points is made for the power low plot.

3. From the least square fit of the slope of the power low plot, calculate the fractal dimension.

3. RESULTS AND CONCLUSIONS

3.1. Adaptive Mixtures

To test the performance of the adaptive mixture method, two one-dimensional data samples, which are randomly derived from $(1/\sqrt{2\pi})\exp(-x^2)$ and $\exp(-x)$ for $x \geq 0$, are used. Each data sample contains 100 data points and the fit results obtained using Adaptive Mixtures is shown in Fig. 1.

In the example of the Gaussian distribution shown in the left plot of figure 1, the data is over-fitted. Furthermore, the fit in the exponential example given in figure 1 is not consistent with the data points. It is found that these problems cannot be resolved by changing the value of the threshold T_c. As a result, a new algorithm may be necessary to obtain a better iteration result and to prevent over-fit.

Table I ϵ_S/ϵ_B comparison in various statistical tools

Method	ϵ_S/ϵ_B
KDELM	26.32 ± 4.51
Decision Tree [3, 4]	6.07 ± 1.24
Neural Network [4, 5]	5.98 ± 1.20
Support Vector [4, 6]	3.95 ± 0.42

3.2. Kernel Density Estimation with Likelihood Maximization

KDELM is considered as a solution to fix the problems with Adaptive Mixtures. The performance of KDELM for the Gaussian and exponential distributions is quite satisfactory, which is shown in the figure 2.

As another test, this technique is applied to tau lepton identification. A Monte Carlo (MC) sample for the decay mode $W^- \to \tau^-\nu_\tau$ (charge conjugation symmetry is assumed) generated for a hadron collider experiment and a generic Quantum Chromodynamics MC sample are used as "signal" and "background", respectively. A discriminant function defined by

$$D(x) = \frac{K_S(x)}{K_S(x)+K_B(x)}, \tag{6}$$

is considered to separate signal MC events from background. Here, $K_S(x)$ and $K_B(x)$ are the signal and background kernel estimates, respectively. Figure 3 shows the distribution of $D(x)$. To compare the signal discrimination power of various statistical tools, ϵ_S/ϵ_B's obtained using these tools are compared, where ϵ_S is the probability for a signal event to be identified as signal and ϵ_B is the probability for a background event to be identified as signal. In this ϵ_S/ϵ_B comparison, ϵ_S is fixed to be 50% and the results are shown in table I.

From these two tests, KDELM turns out to be (i) very robust in the fit, (ii) fast in computation, and (iii) good in signal discrimination. In the signal discrimination, KDELM is as good as the neural network technique, while with KDELM it is conceptually easier to understand the whole fit process.

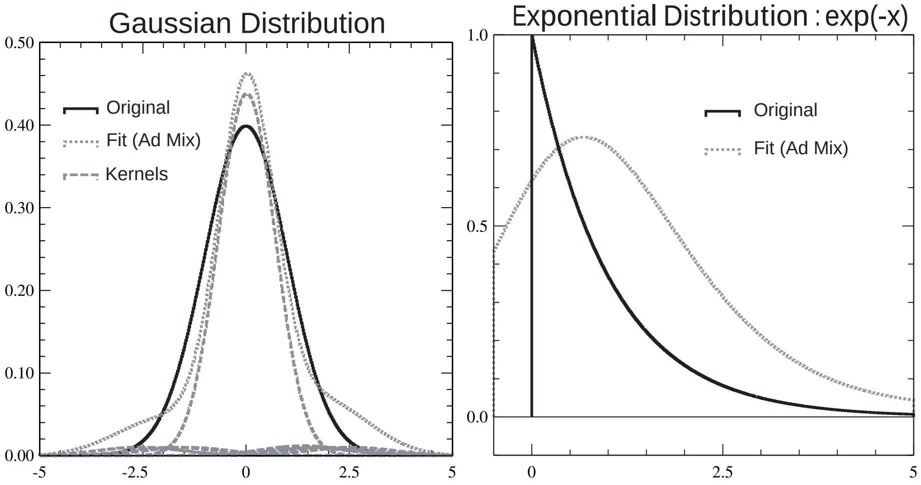


Figure 1: Fit results obtained using Adaptive Mixtures for data samples generated from (left)$(1/\sqrt{2\pi})\exp(-x^2)$ and (right) $\exp(-x)$ for $x \geq 0$.

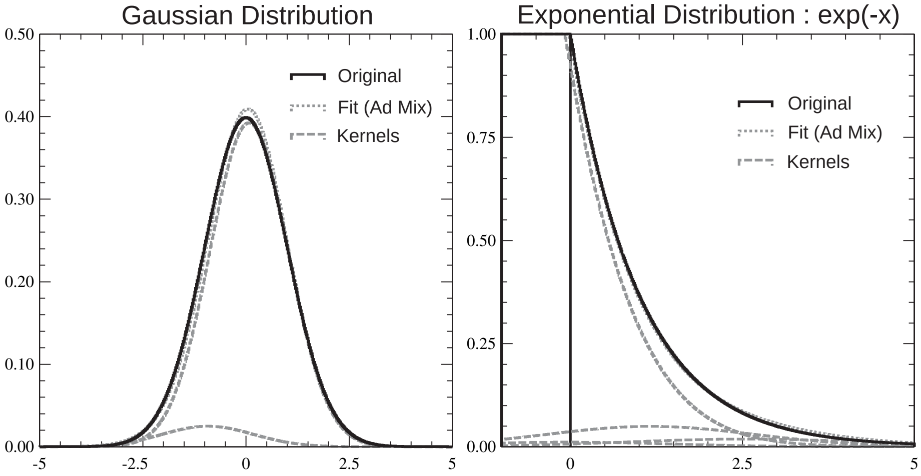


Figure 2: Fit results obtained using KDELM method for data samples derived from (right)$(1/\sqrt{2\pi})\exp(-x^2)$ and (left) $\exp(-x)$ for $x \geq 0$.

3.3. Fractal Dimension Calculation

To check whether fractal dimension calculation will be useful in high energy physics analysis, we use the same MC samples used in the second test of KDELM. First, using the fractal dimension technique, we find the combinations of physics variables which provide the best signal discrimination. Then, these variable combinations are compared with those found using KDELM. Due to lack of statistics of the MC samples, this comparison is taken only up to two-variable combinations. However, this comparison reveals that the best discriminant variables found using fractal dimension often disagree with those obtained using KDELM. As a result, the applicability of fractal dimension calculation as a discriminating feature in high energy physics analysis is skeptical.

Acknowledgments

The authors wish to thank professor David Scott (Rice University) and professor Bruce Knuteson (MIT) for their valuable discussion. Also, the authors would like to thank to professor Ricardo Vilalta (University of Houston) and Abraham Bagher for their calculation of ϵ_S/ϵ_B's in various statistical techniques other than KDELM.

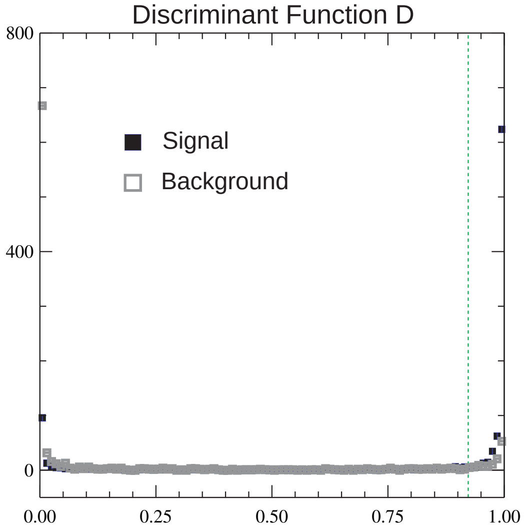


Figure 3: The distribution of $D(x)$ in the tau lepton identification study. The dashed line shows the location of $D_{1/2}(x)$. When we select an event as signal if its $D(x)$ is greater than $D_{1/2}(x)$, ϵ_S becomes 50%.

References

[1] C.E. Priebe, "Adaptive Mixtures", *Journal of the American Statistical Association*, **89**, 796 (1994).

[2] S.N. Rasband, "Fractal Dimension" Wiley, 1990.

[3] S. Russell and P. Norvig, "Artificial Intelligence: A modern approach", Prentice Hall, 1995.

[4] T.M. Mitchell, "Machine Learning", McGraw Hill, 1997.

[5] J. Hertz *et al.*, "Introduction to the Theory of Neural Computation", Addison-Wesley, 1991.

[6] N. Cristianini and J. Shawe-Taylor, "An Introduction to Support Vector Machines and Other Kernel-based Learning Methods", Cambridge University Press, 2000.

Binning-Free Unfolding Based on Monte Carlo Migration*

G. Zech and B. Aslan
University of Siegen, 57072 Siegen, Germany

An experimental data sample is compared to a Monte Carlo sample in the observation space. The Monte Carlo events migrate in the true variate space until the two observed samples coincide. The agreement is quantified using a Gaussian or logarithmic weight for the distance of the observations. The approach is binning-free and especially powerful in multidimensional applications where unfolding of histograms suffers from the "curse of dimensionality". The study is preliminary.

1. INTRODUCTION

In order to deconvolute experimental data which are distorted by measurement errors, physicists usually group the data in histogram bins [1]. The histogram of the true distribution can be regarded as a set of unknown parameters which are determined in fitting procedures. Statistically not significant bin-to-bin oscillations are damped by appropriate regularization schemes. Other methods regularize the Monte Carlo matrix which relates observed and true parameter values [2].

Less stringent than fitting methods, but simpler and quite effective is iterative unfolding introduced in 1982 [3] and reinvented in 1994 [4, 5]. Oscillations are suppressed by stopping the iteration process.

An introduction to unfolding and the related statistical problems is given in Ref. [6].

The energy approach can be combined with Monte Carlo methods to perform binning-free unfolding. Unfolding without binning offers several advantages.

- Arbitrary bin boundaries are avoided.
- Selection criteria can be applied to the data after unfolding.
- Variable transformations are possible after unfolding.
- The initial Monte Carlo simulation which is required to relate true and observed histograms can be rather crude.
- Low statistics data can be handled in arbitrarily high dimensions where histogramming is problematic.
- The unfolded sample represents the statistical precision of the measurement, while errors associated with histogram bins often depend strongly on the regularization strength.

Iterative unbinned unfolding has been presented in a previous paper [5]. Here we propose a new approach based on a simple idea: We start with a Monte Carlo sample of the same size as the observed data sample. We then let the Monte Carlo events migrate in the true variate space until the observed samples are compatible. When the process has converged, the true Monte Carlo sample represents the unfolding result.

The concept requires four ingredients, some of which are not as trivial as might seem at first sight:

1. To quantify the agreement between the experimental data sample and the Monte Carlo sample we need an appropriate test statistic. We use the *energy* concept[7].
2. The simulation scheme has to avoid additional statistical fluctuations. Each Monte Carlo true event is accompanied by a cloud of observations.
3. An efficient migration process should be used.
4. Regularization has to be provided.

2. OUTLINE OF THE METHOD

2.1. The test statistic

To compare an experimental sample $\mathbf{x}'_1, ..., \mathbf{x}'_N$ in the n-dimensional space $\mathcal{R}^n$ to the Monte Carlo sample $\mathbf{y}'_1, ..., \mathbf{y}'_M$, we use the following relation:

$$\phi = \frac{1}{N^2}\sum_{j>i} R(|\mathbf{x}'_i - \mathbf{x}'_j|) + \\ + \frac{1}{M^2}\sum_{j>i} R(|\mathbf{y}'_i - \mathbf{y}'_j|) + \\ - \frac{1}{NM}\sum_i \sum_j R(|\mathbf{x}'_i - \mathbf{y}'_j|)$$

The distance function R is either $R_{\log}$ and R_G:

$$R_{\log}(r) = -\ln(r + \varepsilon) \qquad (1)$$

$$R_G(r) = e^{-r^2/(2s^2)} \qquad (2)$$

*Supperted by Bunderministerium für Bildung und Forschung, Deutschland

The cut-off parameter ε is introduced to avoid divergencies. Its precise value is unimportant, it should be small compared to the distance of observations in the most dense regions. The test statistic ϕ is minimum in the limit $N, M \to \infty$ if and only if the two samples originate from the same parent distribution [7].

2.2. The Monte Carlo sample

We associate with the data sample $\{\mathbf{x}'_i\}$ of size N a Monte Carlo sample $\{\mathbf{y}_i, \mathbf{y}'_{ik}\}$ where i again runs from 1 to N and $\mathbf{y}'_{ik}$ is a set of K observed values of the true value $\mathbf{y}_i$. Thus each true value is accompanied by a subsample of K observations. The number K is typically of the order of 20. The statistical fluctuations of the simulation are smaller than those introduced by the experimental statistic by the factor $\sqrt{K}$. The starting values $\mathbf{y}_i$ are arbitrary. To avoid large computing times, they should not be too far off from their final position.

2.3. The migration process

So far we have spent little effort to optimize the migration process. We select randomly one Monte Carlo point i and modify $\mathbf{y}_i$ by a random uniform displacement $\mathbf{\Delta}$. At the new position K new observations are generated. The change of energy ϕ is computed and the move is accepted if the energy has decreased and rejected otherwise. The migration process continues until the minimum of ϕ is reached.

The process can be accelerated if those Monte Carlo points are selected preferentially which contribute strongly to the energy. The direction of the move could possibly be optimized moving along the energy gradient. Another possibility is to form substructures by grouping observations replacing them by a single replacement charge which is again dissolved at a later stage. We could also consider including only neighboring charges in the energy calculation for a first crude adjustment. The computing time would then be a linear function of the number of events. We have not investigated these possibilities. For up to a few thousand events the calculations can be performed without refinement on a standard PC.

So far we have also neglected the possible existence of more than one local minimum of ϕ. If this happens, one could try introducing an “annealing” term. Moves increasing the energy by $\Delta\phi$ are accepted with probability $p = 1/(1 + e^{\Delta\phi/T})$ with an appropriate choice of the parameter T.

2.4. The regularization

There are two different choices of regularization: i) The migration process can be stopped before the oscillations become intolerable. ii) The value of the parameter K can be adjusted. Large values provide high resolution but introduce oscillation. The point spread function σ_u after deconvolution for N experimental observations of a single point with resolution σ is expected to follow

$$\sigma_u = \sigma\sqrt{\frac{1}{N}(1 + \frac{\kappa}{K})}$$

where κ is a constant depending on the shape of the distance function.

3. EXAMPLES

Example 1: First we illustrate the new unfolding approach with a one-dimensional distribution. Even though the unfolding has been performed without binning, we present the result in form of histograms in Figure 1. Two Gaussians are superposed to a uniform distribution. The standard deviation of the Gaussians and the assumed experimental resolution were both $\sigma = 0.05$. The Monte Carlo enhancement factor was $K = 16$.

Example 2: In Figure 2 we present a simple example in two dimensions. The original picture of the face consisted of infinitely thin circular lines and of dots for the eyes. The picture contains 600 observations. Each true Monte Carlo point was accompanied by $K = 25$ observations. The unfolded picture was obtained after $20,000$ trials of random moves.

It would have been be quite difficult to convert the pictures of Figure 2 into histograms and to apply the standard deconvolution methods.

Example 3: Finally, we apply the unfolding to a toy PET measurement in two dimensions. A positron and an electron annihilate at rest and the tomograph registers the two back-to-back photons at a circular detector at angles α, β. The emission point has to lie on the line connecting the two positions where the photons are detected. For the simple case that the source consists of a single point, all observations are located on a curve in the two-dimensional α, β space. We have simulated the process for a source consisting of two source points located at $(x_1 = 1, y_1 = 0)$ and $(x_2 = 1, y_2 = 1)$ and a total of 500 observation pairs α_i, β_i. The initial position of the Monte Carlo sources was $(x_{MC} = 0, y_{MC} = 0)$.

Figure 3 shows the result of the deconvolution which was performed with a logarithmic distance function and a K factor of 25. The Monte Carlo source points have moved to the expected locations and their angle pairs ly on the corresponding curves.

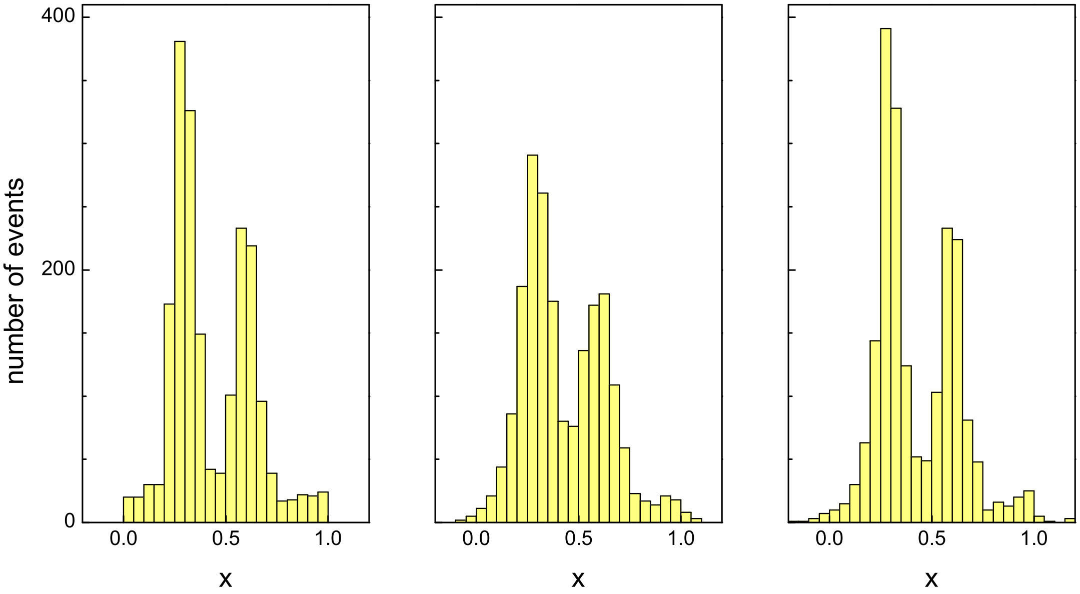


Figure 1: Unfolding of a one-dimensional distribution. The true distribution (left), the smeared distribution (center) and the unfolded distribution (right) are shown in form of histograms.

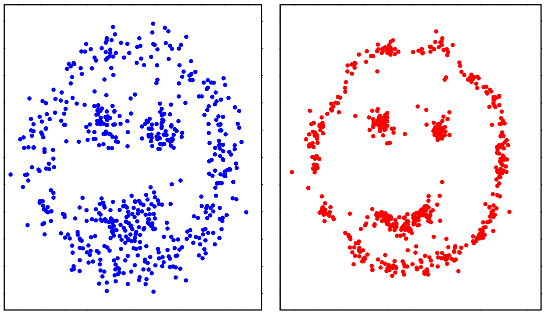


Figure 2: Unfolding of a simple picture.

4. TECHNICAL REMARKS

- The distance function used to compute the energy is only of technical importance. In the limit of large numbers it has no influence on the result, only the speed of convergence depends on it. We suggest using Gaussians with width similar to the resolution or logarithmic distance functions.
- The average migration steps should be larger than the resolution. We propose to generate the steps using uniform random numbers for each dimension. Again, the choice of the migration procedure influences only the speed but not the result.
- After each move the energy has to be recalculated. Only the charge combinations which contain the moving charge have to be evaluated.
- Acceptance losses can be included in our method by weighting the Monte Carlo events. The weight is set equal to the ratio T/K of number T of trials required to generate K observations, the inverse of the acceptance.
- If acceptance and resolution are independent of the location, the K observed Monte Carlo points can migrate together with the true point. Resimulation is not required in this case.

5. DISCUSSION

The performance of the various unfolding procedures on the market is very similar. Without regularization and constraints, histogram fitting methods, likelihood or χ^2, iterative unfolding and matrix in-

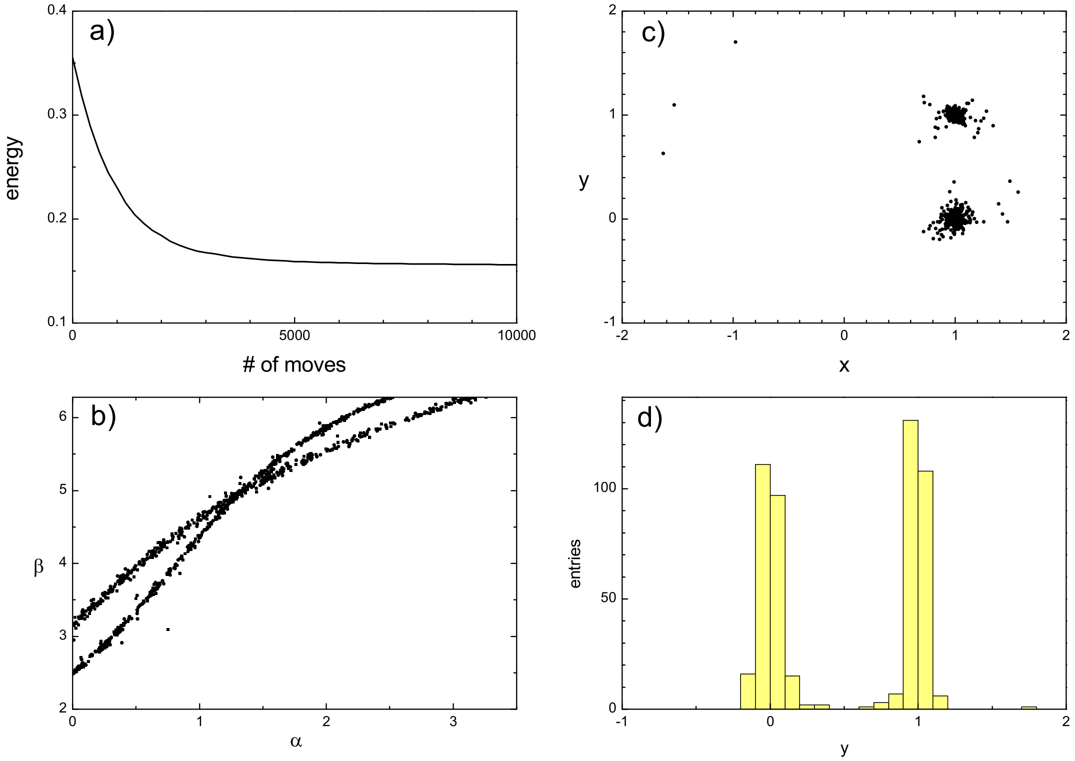


Figure 3: Unfolding of PET measurement: a) convergence of energy, b) observed angles, c) reconstructed source positions, d) corresponding y projection.

version give identical results[1]. The differences of the methods lie in the regularization schemes they apply. Also the binning free approach is exact in the limit of $K >> 1$ up to the necessary smoothing effects. There is no additional loss of information. This is obvious, because we can reproduce the original observed distribution from the unfolded sample up to the regularization effects.

The new approach opens the possibility to solve problems which are not accessible with the conventional unfolding methods. It is especially powerful in multidimensional applications with sharp structures.

In smooth, high statistics distributions, histogramming methods are preferable because they are faster.

Additional work is required to study the effect of local minima, to optimize the migration process and to study cases with a very small number of observations.

[1] There is a difference in rare cases: Matrix inversion can produce negative bin contents which are avoided in the other methods which then yield a biased result.

References

[1] A. N. Tikhonov, "On the solution of improperly posed problems and the method of regularization", Sov. Math. 5 (1963) 1035
V. B. Anykeev, A. A. Spiridonov and V. P. Zhigunov, "Comparative investigation of unfolding methods", Nucl. Instr. and Meth. A303 (1991) 350.

[2] V. Bobel, "An unfolding method for high energy physics experiments", Proceedings of Conf. Advanced Statistical Techniques in Particle physics, ed. M. R. Whalley and L. Lyons, Durham 2002.

[3] L. A. Shepp and Y. Vardi, IEEE trans. Med. Imaging MI-1 (1982) 113.,
A. Kondor, "Method of converging weights - an iterative procedure for solving Fredholm's integral equations of the first kind", Nucl. Instr. and Meth. 216 (1983) 177,
H. N. Mülthei and B. Schorr, "On an iterative method for the unfolding of spectra", Nucl. Instr. and Meth. A257 (1986) 371.

[4] G. D'Agostini, "A multidimensional unfolding method using Bayes' theorem", Nucl. Instr. and Meth. A362 (1995) 487.

[5] L. Lindemann and G. Zech, "Unfolding by weighting Monte Carlo events", Nucl. Instr. and Meth. A354 (1994) 516.

[6] G. Zech, "Comparing statistical data to Monte Carlo simulation - parameter fitting and unfolding", Desy 95-113 (1995).

[7] G. Zech and B. Aslan, "A Multivariate Two-Sample Test Based on the Concept of Minimum Energy", available in these Proceedings on page 97 and at http://arxiv.org/abs/math.PR/0309164.

Variational Methods in Bayesian Deconvolution

K. Zarb Adami
Cavendish Laboratory, University of Cambridge, UK

This paper gives an introduction to the use of variational methods in Bayesian inference and shows how variational methods can be used to approximate the intractable posterior distributions which arise in this kind of inference. The flexibility of these approximations allows us to include positivity constraints when attempting to infer hidden pixel intensities in images. The approximating posterior distribution is then optimised by minimising the Kullback-Leibler divergence between it and the true distribution. Unlike traditional methods such as Maximum Likelihood or Maximum-A-Posteriori methods, the variational approximation is immune to overfitting, since the sensitivity of the approximation is towards probability mass rather than probability density. The results show that the present algorithm is successful in interpolation and deconvolution problems.

1. INTRODUCTION

1.1. Measurement

Before the analysis of data is considered, it is worth discussing the measurement process by which the data is collected. Measurement involves an interaction between the instrument and the environment, possibly involving uncertainty in the obtained result. This uncertainty takes on many different forms including both systematic and random effects. The most general form of writing down a measurement is as follows:

$$D = R(\Theta) + \nu \tag{1}$$

where D is the obtained measurement, R is the response function of the instrument as a function of the parameters Θ we set out to measure, and ν is the noise or uncertainty introduced by the environment. Data analysis involves the treatment of the collected measurements to extract the required parameters.

The response function is often complicated, so that even without noise Equation (1) cannot be directly inverted to obtain the parameters. This means that an approximate method is required to form a suitable pseudo-inverse. Since many pseudo-inverses are possible this inversion process is ill-determined and a systematic way of choosing the correct one is required. Throughout this paper we employ a probabilistic solution to this general inference problem. R.A. Fisher discusses three aspects of valid inference: (1) model specification, (2) estimation of model parameters, and (3) estimation of precision. Model specification can be further subdivided into two main categories: the formulation of a set of candidate models, and the selection of a model (or small number of models) to be used in performing inference.

1.2. Model Selection

Choosing the best model typically requires a balance between minimising a cost function, the most ubiquitous one being the chi-squared statistic χ^2, and a regularising function, a common one being the entropy $-\sum_i p_i \log p_i$. The correct balance is deduced through the use of Bayes' theorem:

$$P(\Theta|\mathcal{D},\mathcal{H}) = \frac{\mathcal{L}(\mathcal{D}|\Theta,\mathcal{H}) \times P(\Theta|\mathcal{H})}{P(\mathcal{D}|\mathcal{H})} \tag{2}$$

where Θ are the parameters we are hoping to infer, $\mathcal{D}$ are the data we have collected and $\mathcal{H}$ is the model under scrutiny. The likelihood function, $\mathcal{L}(\mathcal{D}|\Theta,\mathcal{H})$, contains all the parameters we are seeking to infer and is a functional description of the relationship between the parameters and the data. The prior, $P(\Theta|\mathcal{H})$, contains all the knowledge available to the experimenter before the measurement is performed. It reflects any assumptions made by the experimenter and it can contain constraints on the range in which the data should be. The posterior distribution, $P(\Theta|\mathcal{D},\mathcal{H})$, quantifies the belief in the parameters inferred from the data. The denominator of Equation (2) is not merely a normalisation constant, as a further application of Bayes' theorem shows:

$$P(\mathcal{H}|\mathcal{D}) = \frac{P(\mathcal{D}|\mathcal{H}) \times P(\mathcal{H})}{P(\mathcal{D})} \tag{3}$$

The evidence ($P(\mathcal{D}|\mathcal{H})$) is crucial for model selection since Equation (3) is merely the posterior probability of the model given the data. Unfortunately, evaluation of the evidence is a hard problem, involving high-dimensional integrals over parameter space [6]. Although sampling methods have been more popular recently, this paper focuses on a variational method for evaluating a bound on the evidence value and using this bound to perform inference.

2. VARIATIONAL INFERENCE

Variational inference finds itself between the Laplace approximation and sampling methods. At one extreme, Laplace's method approximates the integrand by a Taylor expansion of the logarithm of the posterior around its peak. This method is unwieldy in high-dimensional problems since a large matrix of cross-derivatives is required. At the other extreme, we can approximate the evidence using numerical techniques such as Markov Chain Monte Carlo methods (MCMC) [4]. These are however very computationally intensive.

Variational inference attempts to approximate the integrand until the integral becomes tractable. The general idea, following [7], is to bound the integrand from above or below, reducing the integration problem to an optimisation problem, i.e. making the bound as tight as possible. No parameter estimation is required and the quality of the integral is optimised directly. The Kullback-Leibler cross-entropy is used as a measure of the disparity between the true and approximate posterior ($Q(\Theta)$), and quantifies the loss of information incurred through the approximation.

$$D_{KL}(Q||P) = \int_{\Theta} Q(\Theta) \log \left[\frac{Q(\Theta)}{P(\Theta|\mathcal{D},\mathcal{H})} \right] d\Theta \qquad (4)$$

This can be re-arranged to:

$$\begin{aligned} C_{KL}(Q||P) &= D_{KL}(Q||P) - \log P(\mathcal{D}|\mathcal{H}) \\ &= \int_{\Theta} Q(\Theta) \log \left[\frac{Q(\Theta)}{P(\mathcal{D}|\Theta,\mathcal{H})P(\Theta|\mathcal{H})} \right] d\Theta \\ &\geq -\log P(\mathcal{D}|\mathcal{H}) \qquad (5) \end{aligned}$$

so that the minimum of C_{KL} corresponds to the optimum approximating distribution which provides a lower bound for $\log P(\mathcal{D}|\mathcal{H})$.

This approach allows flexibility in specifying the prior and provides a deterministic way of obtaining a bound on the evidence value. A strength of this approximation lies in its sensitivity to probability mass rather than probability density.

3. INTERPOLATION

Consider the problem of interpolating a curve through a set of points, so that the generative model for our data is:

$$\mathcal{D}_i = \sum_{n=1}^{N} w_n f_{ni} + \nu_i \qquad (6)$$

where $\mathcal{D}$ is the observed data, $\mathbf{w}$ is a vector of parameters and $\mathbf{f}$ is a matrix of basis functions. If we assume our noise model to be gaussian with inverse variance γ, we can immediately write down the likelihood:

$$\mathcal{L}(\mathcal{D} \mid w, \gamma, \mathcal{H}) = \prod_{i=1}^{I} \mathcal{G}\left(\mathcal{D}_i \mid \sum_{n=1}^{N} w_n f_{ni}, \gamma \right) \qquad (7)$$

Bayes' theorem now demands we specify prior distributions for the parameters we are trying to infer, namely $\mathbf{w}$ and γ. The variational method provides the freedom to choose any analytical form for our priors, and we choose priors conjugate to the posterior distribution to provide a suitable analytical approximation. Furthermore, the resulting approximation will have the same form as the prior, so that a set of posterior distributions from one data set could be used as a prior distribution for a new set of data. The resulting prior distributions for $\mathbf{w}$ and γ are [5]:

$$\begin{aligned} p(\mathbf{w}|\mathcal{H}) &= \mathcal{G}(\mathbf{w}|0, a^{(w)}\mathbf{I}) \\ p(\gamma|\mathcal{H}) &= \mathrm{Gamma}(\gamma|a^{(\gamma)}, b^{(\gamma)}) \qquad (8) \end{aligned}$$

Both Maximum Likelihood and Variational methods are now used to perform inference on this problem.

3.1. Maximum Likelihood

By differentiating the likelihood function with respect to the parameters we are attempting to infer, we have:

$$\begin{aligned} \frac{\partial \mathcal{L}(\mathcal{D} \mid w, \gamma, \mathcal{H})}{\partial \mathbf{w}} &= 0 \\ \Rightarrow \mathbf{w} = [\mathbf{f}\mathbf{f}^T]^{-1}\mathbf{f}\mathcal{D} & \qquad (9) \end{aligned}$$

and similarly we obtain

$$\gamma^{-1} = \frac{1}{J} \sum_{i=1}^{J} \left(\mathcal{D}_j - \sum_{n=1}^{N} w_n f_{nj} \right)^2 \qquad (10)$$

where J is the number of data points in the time series.

3.2. Variational Learning

With the variational method, the optimal approximate posterior distributions are found to be [5]:

$$Q(\mathbf{w}) = \mathcal{G}(\mathbf{w}|\widehat{\mathbf{w}}, \widetilde{\mathbf{w}}) \qquad (11)$$

$$Q(\gamma) = \mathrm{Gamma}(\gamma|\overline{a}^{(\gamma)}, \overline{b}^{(\gamma)}) \qquad (12)$$

where the parameters obey

$$\widetilde{\mathbf{w}} = a^{(w)}\mathbf{I} + \langle\gamma\rangle_Q \, \mathbf{f}\mathbf{f}^T \qquad (13)$$

$$\widetilde{\mathbf{w}} = \widetilde{\mathbf{w}}^{-1} \langle\gamma\rangle_Q \, \mathbf{f}\mathcal{D} \qquad (14)$$

$$\overline{a}^{(\gamma)} = a^{(\gamma)} + \frac{1}{2} \sum_{j=1}^{J} \left\langle \left(\mathcal{D}_j - \sum_{n=1}^{N} w_n f_{nj} \right)^2 \right\rangle_Q \qquad (15)$$

$$\overline{b}^{(\gamma)} = b^{(\gamma)} + \frac{J}{2} \qquad (16)$$

where $\mathbf{I}$ is the identity matrix.

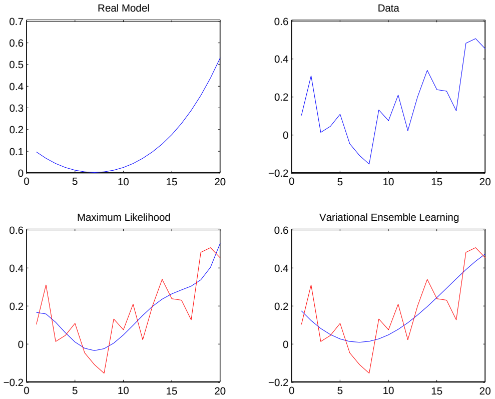


Figure 1: The Maximum Likelihood and Variational Learning algorithms for the interpolation model.

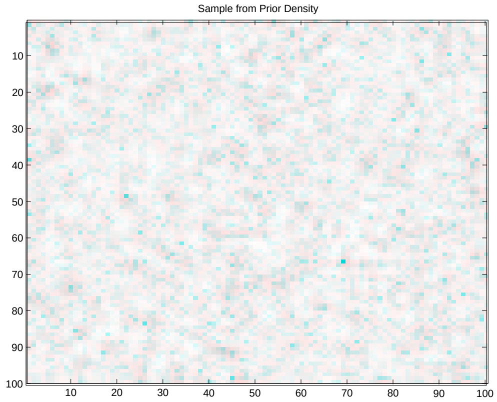


Figure 2: Samples from a mixture of Laplacians. This prior distribution favours sparse images.

3.3. Results

The results displayed in the figure were obtained by trying to approximate the curve $y(x) = \exp(x-0.3)-1$ after gaussian noise has been added to it. The true curve and the data are displayed in the top two panels, while the inferred curves are plotted in blue in the bottom two panels.

The two inference algorithms are similar in computational complexity, though it is clear from the bottom left panel that the Maximum Likelihood method is over-fitting the data, whereas the Variational method seems to approximate the true curve accurately. Also the variational method returns a distribution over possible interpolation curves, unlike the maximum likelihood method which just returns one.

4. DECONVOLUTION

Very often, Equation 1 can be written as a convolution process, where the response function is convolved with the true source distribution, such that Equation 1 becomes linear and of the form:

$$\mathcal{D} = \mathcal{A} * s + \nu \tag{17}$$

where $\mathcal{A}$ is the response or beam function of the apparatus, s is the true source-distribution (pixel intensity in an image case) and ν is the noise system. This means that we now need to specify prior distributions over the source distribution of the pixels in our image. In the next section we discuss the case in which the beam function is also unknown.

Pixel intensity is a positive quantity and our prior distribution should reflect this fact [3]. A suitable distribution is the Laplace distribution. To provide further flexibility, we can model the intensities as a mixture of Laplacian distributions given by:

$$p(s_{ij}) = \begin{cases} \sum_{\alpha=1}^{N_\alpha} \frac{\pi_\alpha}{b_\alpha} \exp\left(-\frac{s_{ij}}{b_\alpha}\right) & s_{ij} \geq 0 \\ 0 & s_{ij} < 0 \end{cases} \tag{18}$$

where ij represents the ij^{th} pixel in the image. Priors over the hyper-parameters π_α and b_α need to be specified in order to ensure sufficiently-broad priors over the pixel intensity. For b_α, a scale-invariant prior is selected so that:

$$p(\ln b_\alpha) = 1 \tag{19}$$

while, since π_α represents the fraction of mixture α present, we use a Dirichlet prior of the form:

$$p(\pi_\alpha) \propto \delta\left(\prod_{\alpha=1}^{N_\alpha} \pi_\alpha - 1\right) \prod_{\alpha=1}^{N_\alpha} \pi_\alpha \tag{20}$$

If we now assume the additive noise is gaussian, we can immediately write down the likelihood function:

$$\mathcal{L}(\mathcal{D}|s, \beta_\sigma) = \prod_{ij} \mathcal{G}(\mathcal{D}_{ij}|\hat{\mathcal{D}}_{ij}; \beta_\sigma^{-1}) \tag{21}$$

where β_σ is the inverse variance of the noise. Since we do not know this quantity, we must also assign it a prior. A scale invariant prior of the form:

$$p(\ln \beta_\sigma) = 1 \tag{22}$$

is used. We can now easily form the posterior distribution over the required parameters, namely s_{ij} and β_σ. Following the variational method, we now suppose that the posterior distributions are separable, so that we can write the approximate tractable distributions as:

$$Q(\Theta|\mathcal{H}) = Q(s, \beta_\sigma|\mathcal{H}) = \prod_{ij} (Q(s_{ij})) \times Q(\beta_\sigma|\mathcal{H}) \tag{23}$$

This assumption allows us to separate out the terms in the cost function so that it can be written as a sum of simple individual terms. Again, a specific form of these posterior distributions is not required, since the forms which optimise the cost function subject to the separable form and normalisation conditions can be found via the variational method. Following [6], each distribution can then be updated in turn using the current estimates for all the other distributions.

4.1. Example

As an example we consider the toy problem of deconvolving some text. Starting with the source distribution displayed in the bottom left panel, we convolve it with a Gaussian beam function and add some gaussian noise to it to obtain the panel in the bottom right of the figure.

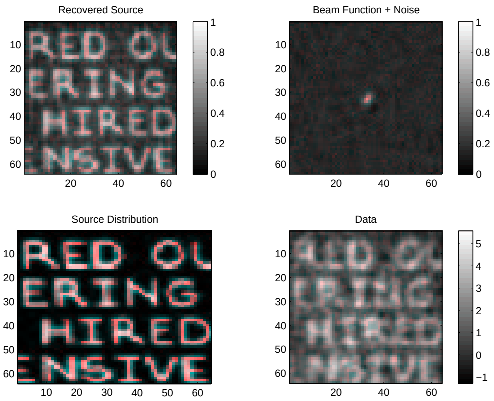


Figure 3: Deconvolution of a noisy image using the variational method with a mixture of Laplacians as a prior.

By learning the source distribution through the variational method, we obtain the result in the top left panel of the figure. However, in astrophysical problems it is sometimes the case that the beam function too is unknown. This problem is common in optical interferometry, where the incoming wavefronts are affected by atmospheric turbulence.

5. BLIND DECONVOLUTION

If the beam function is unknown too, we need to include it in the set of parameters we are trying to infer. In order to tackle the blind problem, we assume (following [2]) that (a) the beam function is smaller than the underlying source distribution and, more importantly, that (b) the beam function is **independent** of the source distribution. We now write down the blind deconvolution problem, following [6], such that:

$$\mathcal{D}_{ij} = \sum_{k=-K}^{K} \sum_{l=-K}^{K} A_{kl} s_{i-k,j-l} + \nu_{ij} \tag{24}$$

where A_{kl} is an element of the beam matrix, which we have assumed to be square and of side $2K$. In evaluating the above sum we assume that the source distribution outside the defined extent of the image is zero. In addition to our deconvolution priors we must now specify a prior over the elements of the beam matrix. In order to respect positivity a Laplacian prior is selected so that:

$$p(A_{kl}) = \begin{cases} \beta_a \exp\left(-\beta_a A_{kl}\right) & A_{kl} \geq 0 \\ 0 & A_{kl} < 0 \end{cases} \tag{25}$$

As in the previous section we use a scale invariant prior for β_a. If we again approximate the posterior distribution by separable distributions for each parameter, we can derive the update equations for the approximate posteriors, following [5], and then iteratively update the posterior distributions. Below is an example in which the beam matrix is unknown and is inferred using the above method.

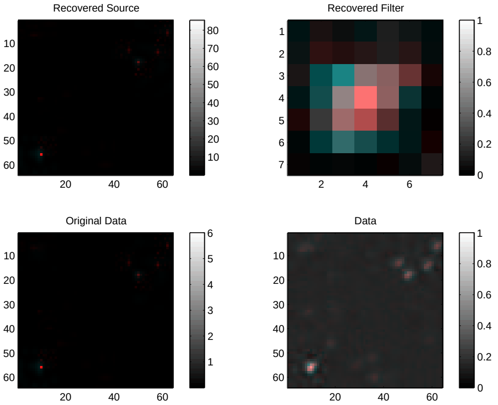


Figure 4: Blind deconvolution of a noisy image using the variational method.

As seen in the figure above, we can use the variational method to successfully infer both the beam function and the underlying source distribution. Throughout the previous two sections, we have assumed that pixel intensities are independent and have neglected any intrinsic correlations which may exist. As a further step, one might model an image as an independent set of pixels convolved with another unknown beam matrix, and priors over this beam matrix could be specified and inferred.

6. CONCLUSIONS

This paper has shown how variational methods can be used to perform valid inference by approximating the true posterior distribution by tractable solutions. The strength of variational methods lies in the reduction of a high-dimensional integration problem to a high-dimensional optimisation problem. The results presented in this paper demonstrate that the variational method is useful in both deconvolution and blind deconvolution problems as well as other inference problems. Work in combining the variational method with MCMC methods is continuing with the aim of using the variational approximation to speed up MCMC navigation through posterior space.

Acknowledgments

The author thanks ST Microelectronics for financial support of this project as well as Steve Gull, David Mackay and Michael Hobson for useful discussions.

References

[1] H.Attias, “Blind Source Separation and Deconvolution: The dynamic component analysis algorithm.”, Neural computation **10**, pp 1373-1424, 1998.

[2] H.Attias, “Independent factor analysis.”, Neural computation **11**, pp 803-805, 1998.

[3] S.F. Gull and G.J. Daniell, “Image reconstruction from incomplete and noisy data.”, Nature **272**, pp 686-690, 1978.

[4] J. Skilling, “Bayesys3 Users’ Manual.”, 2003.

[5] J. Miskin, “Ensemble Learning for Independent Component Analysis”, Ph.D. Thesis, University of Cambridge, December 2000.

[6] J. Miskin, D.J.C. Mackay, “Ensemble Learning for Blind Image Separation and Deconvolution”, Chapter 8 Independent Components Analysis: Principles and Practice, Cambridge University Press

[7] T.P. Minka,“Using lower bounds to approximate integrals”, Technical Report, Medai Lab, Massachusetts Institute of Technology, June 2001

A Comparison of Methods for Confidence Intervals

A.Bukin
Budker INP, 630090 Novosibirsk, Russia

Comparisons are carried out of the confidence intervals constructed with Neyman's frequentist method and with the $\Delta L = 1/2$ likelihood method, using the example of low-statistics life time estimates.

1. P.D.F. FOR LIFE TIME ESTIMATORS

For a given value τ of the true life time, the P.D.F. of a measurement is

$$\frac{\mathrm{d}W}{\mathrm{d}t} = \frac{1}{\tau}\exp\left(-\frac{t}{\tau}\right),$$

and so for an experiment with n measurements

$$\mathrm{d}W = \frac{1}{\tau^n}\cdot\prod_{k=1}^{n}\mathrm{d}t_k\cdot\exp\left(-\frac{t_k}{\tau}\right). \quad (1)$$

The negative log likelihood function is

$$L = n\ln\tau + \frac{1}{\tau}\cdot\sum_{k=1}^{n}t_k. \quad (2)$$

The maximum likelihood estimator of the lifetime can easily be found minimizing L

$$\hat{\tau} = \frac{1}{n}\sum_{k=1}^{n}t_k;\quad \min L = L_0 = n + n\ln\hat{\tau}, \quad (3)$$

so the probability (1) can be transformed to

$$\frac{\mathrm{d}W}{\mathrm{d}\hat{\tau}} = \frac{1}{(n-1)!}\cdot\left(\frac{n\hat{\tau}}{\tau}\right)^{n-1}\cdot\frac{n}{\tau}\cdot\exp\left(-\frac{n\hat{\tau}}{\tau}\right). \quad (4)$$

Given some true value τ then for any algorithm that defines a confidence interval $\hat{\tau}^{+\Delta\tau^{(+)}}_{-\Delta\tau^{(-)}}$ we can evaluate the coverage P:

$$P = \frac{1}{(n-1)!}\cdot\frac{n}{\tau}\cdot\int_{\hat{\tau}_1}^{\hat{\tau}_2}\left(\frac{n\hat{\tau}}{\tau}\right)^{n-1}\cdot\exp\left(-\frac{n\hat{\tau}}{\tau}\right)\mathrm{d}\hat{\tau}, \quad (5)$$

where $\hat{\tau_1} + \Delta\tau^{(+)} = \tau;\quad \hat{\tau_2} - \Delta\tau^{(-)} = \tau.$

2. LIKELIHOOD FUNCTION CONFIDENCE INTERVAL

The conventional Likelihood function method for finding a 68% confidence interval [1, 2] is to find the values of τ for which

$$\Delta L = L - L_0 = \frac{1}{2}.$$

In our case

$$\Delta L = n\cdot\left(\frac{\hat{\tau}}{\tau} - 1 + \ln\frac{\tau}{\hat{\tau}}\right) = \frac{1}{2}. \quad (6)$$

For example, for $n = 5$ the limits are

$$[\tau_1, \tau_2] = [0.6595\hat{\tau}, 1.6212\hat{\tau}]. \quad (7)$$

The coverage of this interval, from Equation (5), is

$$\frac{1}{4!}\cdot\frac{5}{\tau}\cdot\int_{\hat{\tau}_1}^{\hat{\tau}_2}\left(\frac{5\hat{\tau}}{\tau}\right)^4\cdot\exp\left(-\frac{5\hat{\tau}}{\tau}\right)\mathrm{d}\hat{\tau} = 0.6747,$$

where the integration limits, corresponding to (7), are

$$[\hat{\tau_1}, \hat{\tau_2}] = [0.6168\tau, 1.5163\tau].$$

The coverage is close to, but significantly different from, the nominal value of 0.6827.

Examples of confidence intervals obtained by this means are shown in Table I, as the values in parentheses. The 95% confidence interval was obtained using the rule $\Delta L = 2$, and 90% upper limit using a one side interval for which

$$\Delta L = \left[\mathrm{erf}^{-1}(2\cdot 0.9 - 1)\right]^2 \approx 0.821.$$

The coverage given by such intervals is shown in Fig. 1, evaluated using a Monte Carlo method.

3. BAYESIAN CONFIDENCE INTERVAL

For comparison we can estimate a Bayesian confidence interval for the same example of $n = 5$. In the Bayesian approach [3–5], the likelihood function is considered to be a probability density for the true parameter τ. Assuming a flat prior distribution for τ this is

$$\frac{\mathrm{d}W}{\mathrm{d}\tau} = P(\tau) \sim \mathcal{L} = \frac{1}{\tau^n}\cdot\exp\left(-\frac{n\hat{\tau}}{\tau}\right).$$

After normalization (for $n \geq 2$) this becomes:

$$P(\tau) = \frac{(n\hat{\tau})^{n-1}}{(n-2)!\cdot\tau^n}\cdot\exp\left(-\frac{n\hat{\tau}}{\tau}\right),$$

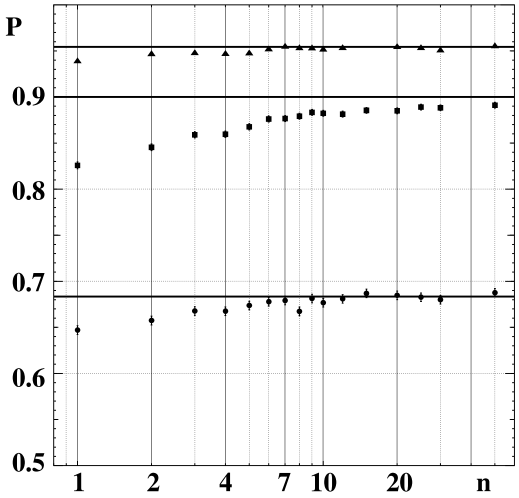


Figure 1: Coverage for likelihood function confidence intervals, evaluated by Monte Carlo. Statistical errors are shown when they exceed the size of polymarker. ▲ — 95% Conf.Interv., ■ — 90% Upper limit, • — 68% Conf.Interv.

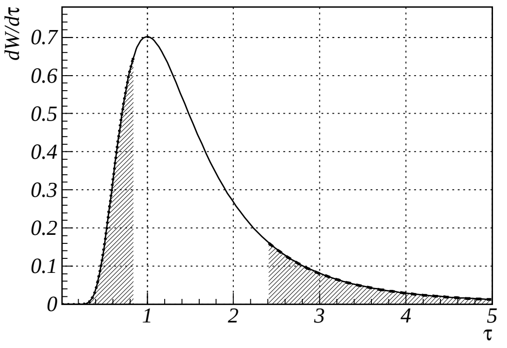


Figure 2: Probability density function for the true value of the parameter τ in a Bayesian approach (Equation (8) with $n = 5$ and $\hat{\tau} = 1$). The shaded regions are the 16% "tails".

which for $n = 5$ gives

$$\int_0^\tau \mathrm{d}W = \left[1 + \frac{5\hat{\tau}}{\tau} + \frac{1}{2}\left(\frac{5\hat{\tau}}{\tau}\right)^2 + \frac{1}{6}\left(\frac{5\hat{\tau}}{\tau}\right)^3\right] \cdot e^{-\frac{5\hat{\tau}}{\tau}}. \qquad (8)$$

The 68% central confidence region for this distribution is (see Fig. 2):

$$\tau = \hat{\tau} \cdot \left(1^{+1.3974}_{-0.1552}\right)$$

The coverage of this region is actually not 68.27% but 64.31%.

4. NEYMAN'S CONFIDENCE INTERVAL

Neyman [5–7] proposed a frequentist construction of a confidence zone (or confidence belt) as follows (see Figure 3):

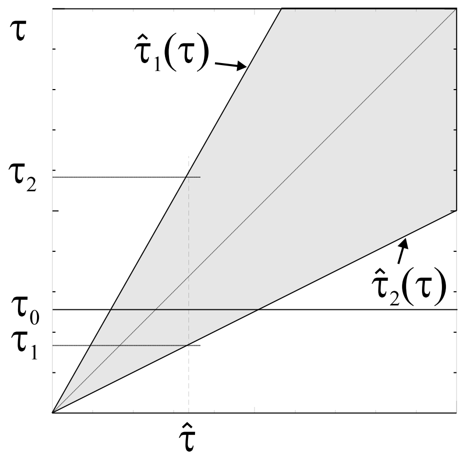


Figure 3: Illustration of the construction of a confidence zone (or confidence belt).

1. One obtains functions $\hat{\tau}_1(\tau)$ and $\hat{\tau}_2(\tau)$ of the true parameter τ such that

$$\int_0^{\hat{\tau}_1(\tau)} \frac{\mathrm{d}W(\hat{\tau};\tau)}{\mathrm{d}\hat{\tau}}\,\mathrm{d}\hat{\tau} = \frac{1-\beta}{2};$$

$$\int_{\hat{\tau}_2(\tau)}^{\infty} \frac{\mathrm{d}W(\hat{\tau};\tau)}{\mathrm{d}\hat{\tau}}\,\mathrm{d}\hat{\tau} = \frac{1-\beta}{2},$$

where β is the confidence level required, here $\beta = 0.6827$. For $n = 5$ these are simply $\hat{\tau}_1(\tau) = 0.568\tau$, $\hat{\tau}_2(\tau) = 1.433\tau$, as shown in Figure 3.

2. One defines the inverse functions

$$\tau_1(\hat{\tau}) = \hat{\tau}_2^{-1}(\hat{\tau}); \quad \tau_2(\hat{\tau}) = \hat{\tau}_1^{-1}(\hat{\tau})$$

which, for a given value of $\hat{\tau}$, define the borders of the confidence interval for τ, with coverage β.

In our example, there are $\tau_1(\hat{\tau}) = 0.698\hat{\tau}$, $\tau_2(\hat{\tau}) = 1.760\hat{\tau}$.

Thus the result of a lifetime experiment of this type can be written

$$\tau = \hat{\tau} \cdot \left(1^{+0.760}_{-0.302}\right).$$

The coverage evaluated is 0.6826 — the difference of 0.0001 is purely due to rounding errors.

Table I shows these intervals for several values of n, with the likelihood approximation shown in parentheses for comparison.

Table II compares the coverage of all three methods for the $n = 5$ case.

Table I Lifetime confidence intervals obtained by Neyman's method for various values of n, the number of measurements, and confidence levels.

n	68% C.L.		95% C.L.		90% C.L.
	$\frac{\Delta\tau^{(-)}}{\hat{\tau}}$	$\frac{\Delta\tau^{(+)}}{\hat{\tau}}$	$\frac{\Delta\tau^{(-)}}{\hat{\tau}}$	$\frac{\Delta\tau^{(+)}}{\hat{\tau}}$	upper limit
1	0.457 (0.576)	4.789 (2.314)	0.736 (0.778)	42.45 (18.06)	$9.49\hat{\tau}$ $(8.49\hat{\tau})$
2	0.394 (0.469)	1.824 (1.228)	0.648 (0.682)	7.690 (5.305)	$3.76\hat{\tau}$ $(2.76\hat{\tau})$
3	0.353 (0.410)	1.194 (0.894)	0.592 (0.621)	4.031 (3.164)	
4	0.324 (0.370)	0.918 (0.725)	0.551 (0.576)	2.781 (2.314)	
5	0.302 (0.341)	0.760 (0.621)	0.519 (0.541)	2.159 (1.858)	
6	0.284 (0.318)	0.657 (0.550)	0.492 (0.513)	1.786 (1.571)	
7	0.270 (0.299)	0.584 (0.497)	0.470 (0.489)	1.538 (1.374)	
8	0.257 (0.284)	0.529 (0.456)	0.452 (0.469)	1.359 (1.228)	
9	0.247 (0.271)	0.486 (0.423)	0.435 (0.451)	1.225 (1.116)	
10	0.237 (0.260)	0.451 (0.396)	0.421 (0.436)	1.119 (1.027)	
20	0.182 (0.194)	0.285 (0.261)	0.331 (0.341)	0.654 (0.621)	
50	0.124 (0.129)	0.164 (0.156)	0.232 (0.237)	0.356 (0.346)	

Table II Coverage of all three methods for $n = 5$

Method	Negative error $\Delta\tau^{(-)}/\hat{\tau}$	Positive error $\Delta\tau^{(+)}/\hat{\tau}$	Coverage, %
Likelihood	0.341	0.621	67.47
Bayesian	0.155	1.397	64.31
Neyman's	0.302	0.760	68.26

5. CONCLUSION

- Neyman's method for confidence intervals provides exact coverage, by construction.
- The intervals from $\Delta L = 1/2$ agree well with the Neyman intervals for large n, but differ for small n, as can be seen in Table I. In such cases they undercover, i.e. the interval is smaller than the true one.
- Bayesian confidence intervals give very different results, and can undercover or overcover.

References

[1] *Derek J. Hudson.* Lectures on elementary statistics and probability, CERN, Geneva, 1963

[2] *A.G. Frodesen, O. Skjeggestad, H. Toffe.* "Probability and statistics in particle physics". Universitetsforlaget, Oslo, 1979

[3] *H. Jeffreys.* "Theory of probability", 2nd ed., Oxford Univ. Press, 1948.

[4] B.P. Carlin and T.A. Louis. "Bayes and empirical Bayes methods for data analysis", Chapman & Hall, London, 1996

[5] *M. Kendall and A. Stuart.* "The advanced theory of statistics", vol. 2, "Inference and relationship". Macmillan Publishing Co., New York, 1978.

[6] *J. Neyman.* "Outline of a theory of statistical estimation based on the classical theory of probability". Phil. Trans. A, 236 (1937) 333.

[7] *R.J. Barlow.* "Statistics. A guide to the use of statistical methods in the physical sciences." John Wiley & Sons ltd., Chichester, England, 1989

Maximal Information Analysis: I - Various Wayne State Plots

G. Bonvicini
Wayne State University, Detroit, MI 48201, USA

Data analysis using all moments of the likelihood $L(\alpha)$ is presented. The relevant plots for various data fitting situations are presented. The goodness of fit parameter (currently the χ^2) is redefined as the isoprobability level in a multidimensional space. Fundamental properties of statistical analysis are described for the first time.

In 1987 I co-wrote a paper that reanalized narrow resonance data in e^+e^- collisions[1].The analysis was made necessary by the widespread use in the community of incorrectly calculated radiative corrections. Those resonance data were obtained from fits of the experimental resonance, a bell-shaped curve with three free parameters. Somehow large, shape-like changes in the radiative corrections were re-absorbed by the normalization parameter Γ_{ee}, producing a large bias in that quantity, while producing acceptable χ^2 results. At the end of the study the world average of Γ_{ee} for the various Υ resonances changed by up to three standard deviations. During our reanalysis it was noticed that the only trace of a biased fit was in the abnormally large uncertainty in the Γ_{ee} error, when data were compared with a toy Monte Carlo.

In 1992 I analyzed in detail various 17-keV neutrino experiments[2]. At the time a majority of experiments found null results, but five found results consistent with heavy neutrino mixing ($\sin^2\theta \sim 0.8\%$). In that analysis, it was pointed out that most experiments, including some of those which turned out to be "correct", had abnormally large uncertainties in the $\sin^2\theta$ error, indicating that there were biases, and therefore that the limits could be broader than published (thus, the discrepancy between the groups of experiments would significantly decrease).

In 1997 Jean Dubosq analyzed the end point of the decay $\tau \to 5\pi\nu$ in the CLEO data[3], following a similar analysis by ALEPH[4], for the purpose of extracting the neutrino mass. While CLEO had nearly 40 times the statistics of ALEPH, and a better energy and mass resolution, the ALEPH limit was a factor of 1.5 better. The limit is taken, roughly, as the average μ plus twice the error σ, and an anomalously low σ will produce an underestimated limit.

It is not a surprise that the σ is sensitive to bias (and therefore be included somehow in the definition of goodness-of-fit). The Cramer-Frechet-Rao limit[5] states as much. The variance (defined as σ^2) when bias is absent is (one parameter only)

$$\sigma_0^2 = \Sigma_i((\frac{\partial y_i}{\partial\alpha})^{-1}\delta_{yi})^2 \tag{1}$$

and is changed by a factor

$$\sigma^2 = \sigma_0^2(1+\frac{\partial b}{\partial\alpha})^2 \tag{2}$$

in the presence of a bias b.

With this work the usage of higher moments is introduced, and several, apparently previously unobserved, fundamental properties of statistical analysis are discussed. We consider the likelihood function which is the product of the probabilities for each data point y_i (in a 1-dimensional plot, the ordinate), given the fit parameter(s) α,

$$L(\alpha) = \Pi_i P(y_i|\alpha).$$

The data points are to be fitted with a function $f(x_i,\alpha)$, (the x variable(s) are, in a 1-dimensional plot, the abscissa). The probability $P(\alpha)$, proportional to the likelihood, is also defined, so that

$$P(\alpha) = \frac{L(\alpha)}{\int L(\alpha)d\alpha}.$$

In the cases discussed below, where population histograms are to be analyzed, one uses the binned likelihood with Poissonian statistics

$$L(\alpha) = \Pi_i \frac{e^{-f(x_i,\alpha)}f(x_i,\alpha)^{y_i}}{y_i!}.$$

If the fitting function is truly unbiased, or so close to an unbiased fit that it can be considered unbiased, then only a very restricted region of Hilbert space will be occupied by fits which were generated by the statistics described by the fitting function. The new, generalized, goodness of fit parameter is

$$G = \int P(A|N)P(\alpha)d\alpha. \tag{3}$$

N is the total number of events. In case of a continuous measurement, $N \to \delta$, the set of errors associated with the set of data points (that is, each data point is $y_i \pm \delta_i$). Because of the way G is constructed, the statistics to be used in allocating events to bins is multinomial. In the longer version of this paper, it is shown that the two statistics are equivalent. The procedure to determine G is a two-step procedure, first one determines the likelihood, and then one generates the likelihood-dependent plots described below and finds G.

A detailed discussion of the motivation for the choice of G as a goodness-of-fit statistic will be included in a longer version of this paper. Three points are made here. The first is that goodness of fit is an internal consistency check for the hypothesis H that the data y were generated by $f(x, \alpha)$. It seems appropriate that all moments of the likelihood be used. Second, suitable combination of the moments describe completely the likelihood in Hilbert space. If one assumes that all the information is contained in the likelihood, then the method retrieves maximal information about H.

Third and most important, the convolution over the experimentally obtained $P(\alpha)$ is done to take into account the knowledge of the true value of α. Equally prepared experiments will have slightly different plots because of the different results μ. It is important to show that the plots maintain their power when α_{true} is varied.

The set A of statistical quantities depends on the type of fit (how many parameters, and how many of them nuisance) and also on N (the smaller the statistics, the more important the skewness parameter M_3 will be). In the case of one parameter fits discussed below

$$A = (\chi^2, \sigma, (M_3)). \qquad (4)$$

These quantities (plus the estimator μ) are used below and are defined as

$$\begin{aligned} \chi^2 &= \ln(L_{max}), & (5)\\ \mu &= \int P(\alpha)\alpha d\alpha, & (6)\\ \sigma &= (\int P(\alpha)\alpha^2 d\alpha - \mu^2)^{1/2}, & (7)\\ M_3 &= (\int P(\alpha)\alpha^3 d\alpha - 3\sigma^2\mu + 2\mu^3)^{1/3}. & (8) \end{aligned}$$

The extra set of parentheses around M_3 indicates that it might, or might not, be of use. If the error is truly gaussian (or the statistics is truly very large), then M_3 will generally be devoid of information. It is the asymmetric nature of Poissonian statistics that generates a significant M_3. Two sample fitting functions $f(x, \alpha)$ are listed in Table 1. They are used to introduce the properties of the two-dimensional subspaces. Ten data (y) points, each corresponding to a bin centered at $x = 0.05, 0.15, ..., 0.95$, were generated by toy Monte Carlo with total number of events, summed over the ten bins, $N = 500$, and with the true parameter $\alpha_{true} = 1.0$. One such generation is called an "experiment". The number of experiments for each function was 3×10^5. The experiments were then fitted with the same function (by construction, an unbiased fit) and the quantities in Eqs.(5-8) recorded for plotting.

The various plots, obtained by plotting any two of the quantities in Eqs.(5-8), for a sufficient number of experiments, are called the Wayne State plots (plots containing μ as one of the axes are generally not usable in a real life experiment. They are useful at this stage to elucidate some hidden properties of statistical analysis). The population of the plots is equal to the number of experiments attempted in the simulation, in this case 3×10^5. Once a plot is produced, the real experimental result needs to be compared against the plot to evaluate the internal consistency of the analysis.

Table I Functions used to produce the plots. The integral is over the bin width, and the sum over the ten bins described in the text equals one.

Name	Function
f1	$\frac{2}{\alpha+2}\int_{min}^{max}(1-x)^{\alpha/2}$
f2	$\frac{1}{1-e^{-\alpha}}\int_{min}^{max}e^{-\alpha x}$

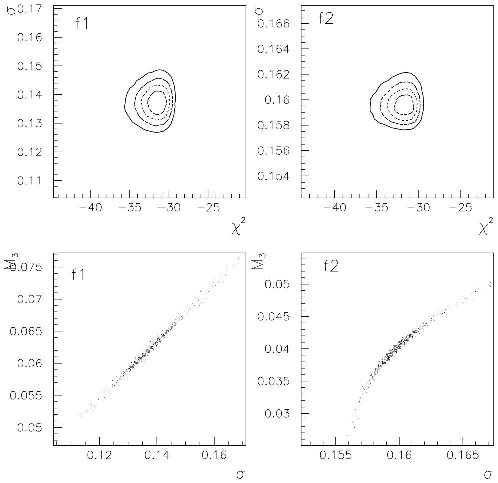


Figure 1: Top rows: first (χ^2, σ) plot for f1 and f2. Bottom row: second (σ, M_3) plot for f1 and f2.

The fit was done by raster scan over a large α interval, then over a more finely segmented, narrower interval within 10σ around the best-fit lattice point. As each point was probed, cumulative quantities useful for the determination of moments were computed.

In Fig. 1, top row, the first Wayne State plot (the (χ^2, σ) plot) are shown. In the limit of a meaningless M_3, this plot recovers maximal information about H. The contour levels shown are isoprobability curves, or rather, iso$-G$ curves. The confidence level for a given fit can be taken to be the fraction of fits lying outside that curve.

There are two interesting properties of the first plot. Along the σ axis, the first plot is narrow. Along the χ^2 axis, the plot is very broad. Generally, σ is far more sensitive to bias than the χ^2, and if one insists on using a single goodness of fit raw parameter, it should be σ and not χ^2. In the case of multi-parameter fits, the σ retains its parameter specific function (the goodness of fit for *that* parameter), whereas the χ^2 could be dominated by nuisance parameters.

The apparently very low correlation between σ and χ^2 is a general property of the first plot, except in the limit of very low statistics where a correlation exists (sorry, no space for these figures). Its meaning is that both of these quantities are needed to assess the quality of a fit, because they are generally independent and both have some sort of sensitivity to bias.

Fig. 1, bottom row, shows the two (σ, M_3) plots. It is unclear whether these should be called the second plot. First, these plots will be meaningless in case of truly gaussian errors, second, in the case of a fit with one true parameter and one nuisance parameter (the next simplest case of interest), the (σ, ρ) plot, with ρ the correlation coefficient, is more important than this plot.

Fig.1, bottom row, is shown because of the extreme correlation between the two quantities. It is unclear at this point how useful this might be, but a fit that falls only slightly off the strip is bound to be biased. Much more important than that, the nearly complete correlation shows the rapid onset of information replication. G can be defined in a space with limited dimensionality, and considering further moments adds nothing to G. This is important in view of the fairly complex software that will have to be developed to make full use of the method described here. If one uses only the first plot, the fraction of recovered information can be estimated as the global correlation coefficient between the first plot quantities and M_3.

Fig. 2, top row, shows the two (μ, σ) plots. There is clearly extreme correlation between μ and σ. The Particle Data Book model of "lottery winning experiments" (experiments with comparable or inferior statistics and/or resolution, which manage to obtain superior results or limits) can be put to the test here. Assume that ten equal, unbiased experiments be performed, and then the "lottery winning" one be chosen as the best estimate. From the plots one can see that in more than 50% of the cases the very worst experiment will be chosen and the best one (the one closest to the true value) will be picked with about 0.1% probability.

The second reason to show Fig. 2, top row, is to point out a further positive property of using σ as part of the goodness of fit parameter. σ is very sensitive to purely statistical fluctuations, which can bias the final result just as much as uncorrected bias. The σ value may provide further constraints on μ in those cases where the bias is known to be very low.

Fig. 2, bottom row, shows the first two plots, for the first function of Table 1, when one varies α_{true} by 1.5σ. There is little variation in the first plot, whereas in the second plot the plot the variation happens to be along the strip described by the main plot. The plots have clearly semi-invariant properties.

In conclusion, with the arrival of maximal information analysis, statistics will undergo a profound transformation. No longer bound by rules which are over 80 years old, we can actually extract much of the information available in a given fit. The method proposed here is valid at low and high statistics, for gaussian and non-gaussian errors, for single and multiple parameter fits, and independent of the definition of likelihood. While the phenomenology of maximal information analysis is fairly clear, the technology is going to be challenging. Certain applications (e.g., which set of nuisance parameters to use in a given fit) require relatively fast software which does not exist yet. I thank F. Porter for many useful suggestions.

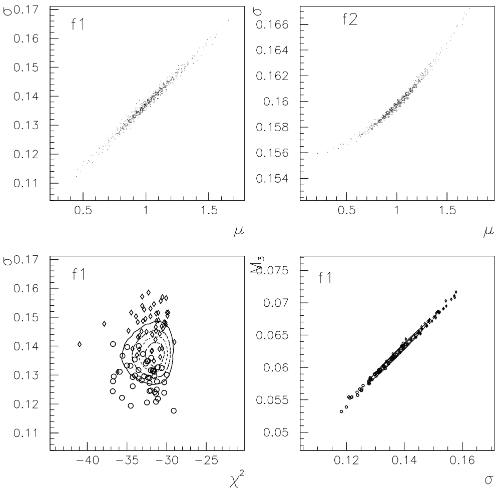


Figure 2: Top row: (μ, σ) plot for f1 and f2. Bottom row: first and second plot for f1 (50 experiments only), when α_{true} is varied. Solid squares: α_{true} =1.0; empty circles: $\alpha_{true} = 1.0 - 1.5\sigma$; empty diamonds: $\alpha_{true} = 1.0 + 1.5\sigma$.

References

[1] J. Alexander *et al.*, Nucl. Phys. B320: 45, 1989.

[2] G. Bonvicini, Z. Phys. A345: 97-117, 1993

[3] R. Ammar *et al.*, (CLEO Collaboration), Phys. Lett. B431: 209-218, 1998

[4] D. Buskulic *et al.*, (ALEPH Collaboration), Phys. Lett. B349:585-596,1995

[5] H. Cramer, Mathematical Methods of Statistics, Princeton Univ. Press, New Jersey (1958).

Event Selection Using Adaptive Gaussian Kernels

A. Askew, H. Miettinen, B. Padley
Rice University, Houston, TX 77271, USA

The Probability Density Estimation method is a technique that uses Kernel Density Estimation techniques to derive a discriminate function which an be used for event selection. This approach has the advantage of handling complex dependencies in data without using the 'black box' approach of neural networks. We present a new variant of the Probability Density Estimation method that allows the use of adaptive kernels. Studies comparing the performance of this method to that of neural networks are presented and prospects for use in physics analysis are described.

1. INTRODUCTION

Neural networks have been used in experimental particle physics analysis with increasing frequency in recent years. However the black box nature of such approaches is worrisome to some. Adaptive Probability Density Estimation was developed at Rice University as an alternative multi-variate approach which provides the flexibility of neural networks with an easily understood and visualizable method. Adaptive PDE is a method of kernel density estimation which uses the distribution of the training data to vary the width of the kernels such that outlying points may be dealt with without discontinuities in the resultant multivariate space.

2. THEORY

2.1. Ordinary Probability Density Estimation

The Probability Density Estimation (PDE) method was implemented by researchers at Rice University for the top quark search. In that analysis, the efficiency of the fixed kernel estimation was found to be similar to that of neural networks. The standard (fixed kernel) probability density estimation method of multivariate data analysis has been previously documented [1]. Unlike neural networks, this method has few free parameters, allowing for less complicated optimization. Given a training sample of data, consisting of sets of signal and background, two functions of the n input variables are formed of the data set. This is accomplished by forming a product of kernel functions in the space of the input variables for each data point. The complete function for the entire sample of events is the sum of all of these product kernels in the training data, normalized by the number of training events used to form the function. This is known as the feature function, because it is an estimate of the important features in the data. Mathematically this function is given by:

$$f(\mathbf{x}) = \frac{1}{N_{tr} h_1 \ldots h_d} \sum_{i=1}^{N_{tr}} [\prod_{j=1}^{d} K(\frac{x_i - x_{ij}}{h_j})]. \quad (1)$$

The $\mathbf{x}$ are the set of variables this analysis is being performed on after transformation into a set in which correlations are zero (the new set of variables is a linear combination of the original variables). This transformation is done so that the kernel structure will match the covariance structure of the data, and thus give a better representation of the data points. Here, K is the kernel chosen to suit the data. In this analysis a Gaussian kernel has been chosen:

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}. \quad (2)$$

The remaining items in the feature function are N_{tr} the number of training events, and $h_j = h^0 \times \sigma_j$, where σ_j is the standard deviation of variable j (in the space in which the correlations vanish), d is the number of variables used in the analysis and h^0 is a tunable parameter which must be optimized for the data set. h^0 is chosen such that the functions formed by the sum of the product kernels are smooth representations without losing information about the data.

Using a linearly independent set of testing data, and the feature functions (f_s, f_b), a discriminant function value ($D(\mathbf{x})$) can be found for each event in the testing sample, representing a measure of how well the event matches the feature function for signal ($D(\mathbf{x})$ close to one) or background($D(\mathbf{x})$ close to zero). This discriminant function is simply;

$$D(\mathbf{x}) = \frac{f_s}{f_s + f_b}, \quad (3)$$

where f_s is the feature function for the signal and f_b is the signal function for the background. Then a single cut on the value of $D(\mathbf{x})$ can be made, selecting the signal from the background by the value calculated for each event.

2.2. Adaptive Probability Density Estimation

The PDE adaptive kernel builds on this method with one modification. An additional parameter α is used to further fit the gaussian kernels to the data set. A pilot $f(x)$ is found using h_j and then the analysis is performed using;

$$h_j = h^0 \times \sigma_j \times \left(\frac{f_{pilot}(\overline{x})}{f_{pilot}(x)} \right)^{\alpha}, \tag{4}$$

where α is another free parameter which must be optimized along with h^0 for the data set, f_{pilot} is the feature function formed using h_j=$h^0 \times \sigma_j$, and $\overline{x}$ is a vector of the mean values of each of the variables chosen for the analysis. This new choice of the width of the product kernels ties the functional form closer to the actual distribution of the data. For sets of data which may have a few outlying events, this now selects a wider gaussian, so that the feature function is smoother for these areas.

2.3. Implementation

The PDE method, as used in this analysis is almost exactly as described in the above section. The only approximation made in the use of the technique is in the method used to diagonalize the covariance matrix of signal and background. For this procedure, a set of Jacobian rotations of the covariance matrices are made to find the eigenvalues and eigenvectors. These rotations are made until the sum of the absolute values of the off diagonal elements of the matrix are sufficiently small. The floating point precision of the machine determines how small the sum of the off diagonal elements must become. For a detailed description of these rotations and the algorithms the reader is referred to [2]. Once the eigenvalues are found, then one can form a transformation matrix which can be used to rotate the covariance matrices such that the correlations in the signal and background covariance matrices vanish. It can be shown that the matrix that achieves this transformation is;

$$A = v^T \times M, \tag{5}$$

$$M = (U \Lambda^{-\frac{1}{2}} U^T), \tag{6}$$

where U is the matrix of the eigenvectors of σ_b (the background covariance matrix), $\Lambda^{-\frac{1}{2}}$ is a diagonal matrix with one over the square root of the eigenvalues of σ_b, and v is the matrix of the eigenvectors of $M \times \sigma_b \times M$. As a consequence of this rotation, the diagonal elements of the background covariance matrix become ones. A proof of how this rotation may be performed can be found elsewhere [1]. This rotation may then be performed on the inputs, which results in the PDE analysis being performed in a space that is a linear combination of the input variables. The PDE algorithms have been implemented in C++ code, and prepared in a shared library form for use in ROOT (object oriented data analysis framework). In this analysis ROOT version 2.25/00 was used.

3. OPTIMIZATION

Using particle ID Monte Carlo courtesy of the D0 experiment, we examine the case of identification of tau leptons with background from QCD multijet production. The optimization of the two PDE methods is detailed.

3.1. Fixed Kernel Optimization

For the fixed kernel PDE method, there is only one parameter to be determined for each data set, h^0. To determine the optimum value for this parameter, the analysis was performed a number of different times using all of the values for h^0 between (0,1] in increments of 0.05. At each value of h^0 the purity times signal efficiency (for a discriminant value of $D(\mathbf{x}) = .5$) was computed, and then the purity times signal efficiency versus the value of h^0 was graphed. The maximum purity times signal efficiency was taken to be the optimum value of h^0 for that data set. The graph is shown below (Figure 1). For the test case of discriminating tau leptons from QCD background, a value of $h^0 = .55$ was found.

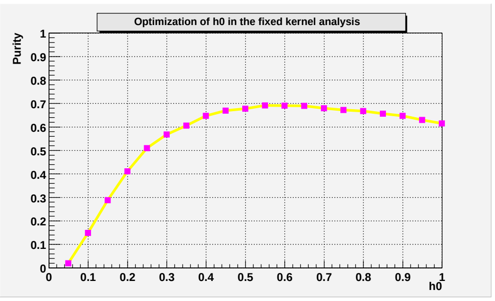


Figure 1: Optimization of h^0 parameter for fixed kernel analysis of τs

3.2. Adaptive Kernel Optimization

The PDE adaptive kernel has two parameters h^0 and α. The optimization of the performance for these two parameters was carried out in a similar way to the fixed kernel. In this case instead of a linear graph, a surface in the space of h^0, α and purity times signal efficiency was formed, in increments of 0.05 in h^0 and α (for the same discriminant function value as the fixed kernel case). The maximum of this surface was taken to correspond to the optimum values of h^0 and α for the analysis. The maximum of the surface was found to be at $(h^0 = 0.65, \alpha = 0.25)$.

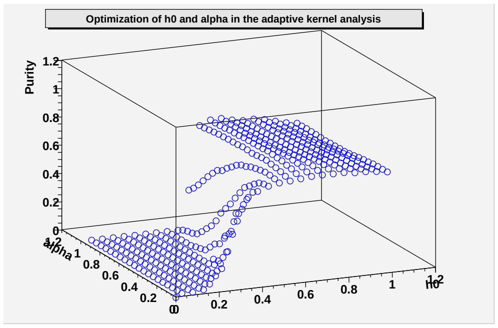


Figure 2: Optimization of h^0 and α: Purity times signal efficiency versus h^0 and α at $D = 0.5$ for adaptive kernel analysis of τs versus QCD jets

4. RESULTS

The above optimized PDE methods were compared with that of a neural network optimized on the same parameter (purity times signal efficiency) and the same signal and background training and testing sets. Figure 3 presents the resultant signal efficiency and background efficiency as a function of a cut on neural network output and PDE discriminant function. The optimum signal efficiencies as determined by the maximum signal efficiency times purity for each method are summarized in Table I.

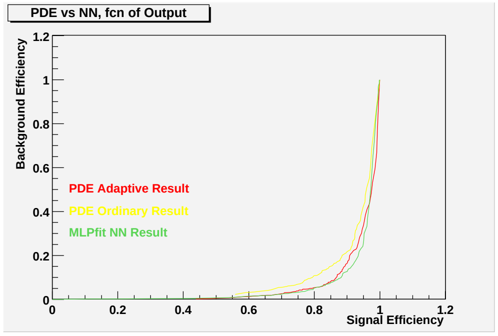


Figure 3: Signal versus Background efficiency for τ versus QCD jets

Table I Summary of Multivariate Performances

Method	Max. Purity $\times \epsilon_s$	ϵ_s	ϵ_b	$\frac{\epsilon_s}{\epsilon_b}$
τ versus QCD jets				
Fixed Kernel PDE	.731	.879	.178	4.94
Adaptive Kernel PDE	.781	.861	.0886	9.72
Neural Network	.793	.883	.100	8.83

Acknowledgments

Work supported by DOE Grant DE-FG05-97ER41031. The authors would like to thank the D0 Experiment for providing the Monte Carlo data used in this study.

References

[1] L. Holmstrom, S. Sain, H. Miettinen“ A new multivariate technique for top quark search”, Computer Physics Communications 88, 1995, 195-210.

[2] J. Gentle, *Numerical Linear Algebra for Applications in Statistics.* Springer-Verlag (1998).

Adaptive Piecewise-constant Modeling of Signals in Multidimensional Spaces

J. Scargle
NASA Ames Research Center, Moffett Field, CA 94035, USA
B. Jackson
Department of Mathematics, San Jose State University; Center for Applied Mathematics and Computer Science
J. Norris
NASA Goddard Space Flight Center, Greenbelt, MD, 20771, USA

This contribution describes an analysis method for the class of problems in which data elements – *e.g.* measurements, event detections, *etc.* – are distributed over some region of space and/or time, or other coordinates (*e.g.*, energy, redshift, category), with the goal of estimating the variation of some physical quantity. The nonparametric model is simply that the physical variable is constant over a finite set of segments of the data space. A dynamic programming algorithm implements such modeling of 1D data by yielding the optimal partition of an interval. Any fitness function that is additive on the partition elements can be used, but the Bayesian posterior probability distribution over partitions—marginalized over all but the geometrical parameters defining the partition—has proved particularly effective. The resulting *maximum a posteriori* piecewise constant model is readily extended to data spaces of higher dimension.

1. SIGNAL AND DENSITY ESTIMATION

A goal of most astronomical observations and particle physics experiments is to describe the variation of some physical quantity as a function of time, space, energy, or other independent variable. We call such a function a *signal.* The experimental procedure is to measure the quantity at a finite number of points in the corresponding *data space.* This paper outlines a way to characterize signal variability using a simple nonparametric model of such data.

Related work on cluster detection in point data in the form of 2D catalogs can be found in [5, 8, 11]. These authors use the distribution of areas of Voronoi cells to establish a threshold of cell density below which lies background and above which are over densities, or clusters.

In subsequent sections we discuss the data (defining the key concept of *data cells*), segmented models and the corresponding fitness functions, the prior distribution for the number of blocks, the algorithm implementing the optimization, and finally application of the method to the large scale distribution of galaxies.

2. DATA CELLS

The data may be in any of a number of forms, such as points (*e.g.* galaxy positions), counts (*e.g.* particle events), measurements (*e.g.* spectral energy density), *etc*, as long as the measurement errors are independent. The current formalism cannot deal with dependent errors nor deconvolve the effects of dispersion—e.g., the *point spread function* affecting GLAST photon data.

The data points are to be associated with *data cells.* A simple example of a data cell is a bin—commonly used to estimate the density of points on some measurement axis. The complete description of the cell corresponding to a given bin requires specification of the number of samples in the bin, plus the bin's boundaries. More generally a data cell is a data structure representing an individual measurement within the *data space*—the set of all values that the measured quantity can possibly take on. For our segmented models, the cells must contain whatever information is needed to compute the model fitness function (§4).

In most cases it is natural to define the data cells to be in one-to-one correspondence to the measurements. But in a specific application it may be preferable to do otherwise—for example, if two or more events have the same time-tag, it may be reasonable to assign them to the same data cell. Similarly, in most cases it will be natural that the data cells partition the entire data space, with no overlap or gaps between cells; and typically the data cells contain information on adjacency to other cells. But in specific applications any of these conditions may be violated.

3. PIECEWISE CONSTANT MODELS

We consider only segmented, *piecewise constant* models. That is to say the data space, whatever its dimension, is partitioned into a finite number of *blocks* within which the measured variable is represented as constant. The complete model consists of the partition, specified by the number of blocks N_b, a list of data cells in each blocks, plus the corresponding levels (*e.g.* event rates) in the blocks.

Boundaries separating blocks can be arbitrary: in 1D, points anywhere in the interval; in 2D, arbitrary line segments; in dimension ν, arbitrary surfaces of dimension $\nu - 1$. Optimization over all possible partitions then involves hugely infinite search spaces.

However, a simple restriction on the class of allowed boundaries yields finite search spaces that are good approximations to the true ones, and turns the problem into a comparatively simple combinatorial optimization. The underlying idea is that two partitions differing only in a small distortion of a block boundary are not significantly different from each other. Construct a bounded volume element around each of the N data points, say consisting of that part of the data space closer to the point than to any other. In only a slight abuse of terminology, we associate these volumes with the data cells discussed above. Further, *blocks* are defined as sets of these cells. Correspondingly a partition of the whole data space is defined by collecting the N cells into distinct blocks. The set of all possible such assignments is finite, but represents an approximation to the hugely infinite set of all possible partitions.

If the cells are defined as above, they form what is called the *Voronoi tessellation*, a geometric partition easily computed in spaces of any dimension [10]. The partition elements, here called *blocks*, are simply sets of cells, with the condition that each cells belongs to one block, and not more than one. There are two important cases: the cells in a block must all be adjacent to each other, or one may not insist on this condition. Think of the blocks as analogous to level surfaces for an unknown function; the two cases correspond to distinguishing or identifying the disconnected parts of a given level surface.

Such step functions comprise the simplest class of nonparametric models, are very easy to interpret, and allow easy computation of summary physical quantities. In visualizations the choppiness due to discontinuities in the *block representation* can be ameliorated, *e.g.* by smoothing, if desired.

We want the model to be sensitive to any and all true variations, but insensitive to apparent variations produced by the inevitable observational errors.[1] We would like to preserve all features in the signal, on all scales supported by the data. But of course all analysis schemes—even those using nonparametric models—involve choices which restrict the questions that can be addressed. Our approach favors local structures over global ones. Because we want to be sensitive to features on fine as well as coarse scales, we do not use smoothing for noise suppression, but rather rely on the accuracy of the statistical model of the observational noise to effect **denoising without smoothing**.[2] A subsidiary goal is to implement an objective procedure suitable for automatic analysis of large data sets (data mining) such as those generated by modern particle physics and astrophysics projects.

The setting just described is more general that it perhaps first appears, and the methodology given here applies to a variety of seemingly different problems, and with a variety of distinct data types. The former include detection of signals and upper limits thereof, density estimation (usually for point data), detection and characterization of clusters, unsupervised classification, and others – including multivariate versions of any of these problems. Essentially any data mode can be treated, as long as one can compute a suitable fitness function for the block model. Fitness functions for point, binned count, and measurement data are readily computed, and categorical data can certainly be dealt with too. Distortions such as data gaps, variable instrumental sensitivity, and (at least in 1D) convolution with an instrumental point-spread function, can also be treated in very natural ways. Perhaps most useful of all is the ready treatment of data in any dimension.

4. FITNESS FUNCTIONS: POSTERIOR PROBABILITIES

A key element in implementing the modeling procedure is a function to measure goodness-of-fit for partitions. The standard Bayesian model estimation method yields convenient expressions valid for a variety of data modes. The simplicity of the block model makes such computations very easy. In particular, we need only compute the posterior for a single block, since statistical independence of the observational errors insures that the posterior for the whole data space is the product of that for each of the partition elements. Indeed, our algorithm requires additivity: the fitness of a partition must be the sum of the fitnesses of its blocks. This condition is achieved by using logarithms of posteriors.

Here is an outline of the procedure. The full posterior probability for the piecewise constant model depends on the block edges and signal level for all blocks. Treating the levels as *nuisance parameters*, and marginalizing them, reduces the full problem into a much more tractable *combinatorial optimization* task—in a nutshell, finding the optimal number of blocks and their edges.

The posterior probability of model M, given data D, is $P(M, \phi, \theta|D)$, where the model parameters have

[1] Of course we sharply distinguish between noise in the sense of random variations inherent in the source and random observational errors.

[2] We adopt the slogan: *the Statistically Significant Structure, the whole Statistically Significant Structure, and nothing but the Statistically Significant Structure.*

been divided into two types: nuisance parameters, denoted by θ, and the others, denoted ϕ. Marginalization of the nuisance parameters is effected simply by carrying out the integral

$$P(M,\phi|D) = \int P(M,\phi,\theta|D)d\theta \ . \tag{1}$$

Bayes' theorem allows this to be written

$$P(M,\phi|N,V) \propto \int P(N,V|M,\phi,\theta)P(M,\phi,\theta)d\theta \ , \tag{2}$$

where we have replaced D with the two relevant parameters (N and V, defined below) and eliminated the factor $P(D)$, irrelevant for model comparison since it is independent of the model. We choose the parameters ϕ to be those specifying the edges of the model segments, leaving all others to be treated as nuisance parameters—the most important of which is the parameter representing the constant value of the signal in the block under consideration.

A useful example is the case where the data comprise events, or counts of events, at various locations in the data space, modeled as Bernoulli or Poisson point processes. Marginalizing the event rate parameter characteristically yields a posterior that depends on two *sufficient statistics*: N, the number of events in the block, and V, the size of the block. For event data the posterior of the block model (abbreviated B), marginalized and conditional on the data, is

$$P(B|N,V) = \frac{\Gamma(N+1)\Gamma(V-N+1)}{\Gamma(V+2)} \tag{3}$$

This quantity can be thought of as the weight[3] which the data gives to model B. The product over the blocks making up a partition gives its weight relative to other partitions. For binned data

$$P(B|N,V) = \frac{\Gamma(N+1)}{(V+1)^{N+1}} \tag{4}$$

The reader is referred to [13] for the details of this computation, including a discussion of the prior distribution for the signal strength and the units in which V needs to be expressed, and details of the fitness functions for several data modes. Applications are discussed in [14–16].

Nothing in the derivation of the above fitness functions depends on the dimensionality of the data space. For event data in a space of dimension ν, *e.g.*, all that matters is that the expected number of events in an elementary ν-dimensional volume element is equal to a constant (the Poisson rate) times the volume. Hence Eqs. (3) and (4) are valid in any dimension.

5. PRIOR ON THE NUMBER OF BLOCKS

One parameter not marginalized, namely the number of blocks, N_b, has a special status, since its value determines the number of other parameters in the complete model. That the value of N_b is automatically found in the optimization is one of the advantages of the dynamic programming algorithm over most cluster analysis methods, in which finding the number of clusters is a vexing problem. One approach is to introduce a term in the fitness function that applies a larger penalty to more complex models. There are various justifications for particular forms of such a penalty term, *e.g.* based on the Minimum Description Length principle [12]. In the Bayesian formalism, there is no need to introduce a penalty term *ad hoc*, since the marginalization of the nuisance parameters yields a built-in effective complexity penalty—sometimes described as the *Occam factor*. But we do need to prescribe a prior distribution for this parameter.

We have adopted a *geometric distribution* for this prior:

$$P(n_b) = C\gamma^{-n_b} \tag{5}$$

(for $n_b \geq 0$) advocated in [2]. This form yields the following contribution to the log-posterior (ignoring an overall constant):

$$log[P(n_b)] = -n_b \ log(\gamma) \ . \tag{6}$$

Note that Eq.(6), since it corresponds to subtracting the constant $log(\gamma)$ from the fitness function for each block, trivially maintains block additivity of the fitness function. We are investigating how the strategy of the algorithm might be modified to allow the use of other functional forms for this prior.

6. THE OPTIMIZATION ALGORITHM

The next step is to optimize the model by maximizing a measure of its goodness of fit over all possible partitions. In [7] we presented a way to find the global optimum of any block-additive fitness function, over all 2^N possible partitions of a 1D interval containing N data points, in time $O(N^2)$. This section is a brief description of this inductive 1D algorithm and its extension to higher dimensions.

Suppose we have the optimal partition of the first n data points, and the corresponding optimal fitness value. Now add one data point, and seek the optimal partition of the first $n+1$ points. Let j be an arbitrary index between 1 and $n+1$, and consider the partition consisting of two parts: (a) the optimal partition of the the first $j-1$ data points, followed by (b) a single block from j to $n+1$. Part (a) and its fitness were found and saved earlier, at iteration number $j-1$, and the fitness of (b) is easily computed. A simple

[3] The ratio of such weights for two models, called the *Bayes factor*, gives the models' relative probabilities.

argument shows that the optimal partition for $n + 1$ data cells corresponds to the value of j that maximizes the combined fitness of (a) and (b).

This algorithm can be extended to a data space of any dimension. We continue to take partitions to be sets of blocks containing data cells (*e.g.* Voronoi cells defined by the data points), but relax the constraint that blocks be simply connected.[4] If an optimal block turns out to be not simply connected, it is straightforward to identify its simply connected parts. Relaxing the connectedness constraint has the effect that a few isolated data cells may be assigned to the wrong block. For example a data point with unusually close (far) nearest neighbors, due to a rare statistical fluctuation, may be assigned to a higher (lower) density block than the one that it actually belongs to. Clearly the locations of the data cells are now irrelevant to the optimization. This permits us to arrange the cells in a 1D array so that the algorithm described above can be used. Ordering by cell density—$\rho(c)$ is the number of events in cell c (usually 1) divided by the volume of c—is reasonable, because the piecewise constant model obviously tries to collect together cells with similar densities. This idea is made rigorous by the *intermediate density property*: given three cells c_1, c_2, c_3 ordered by density, $\rho(c_1) \leq \rho(c_2) \leq \rho(c_3)$, if both c_1 and c_3 are in block B_k, an element of an optimal partition, then c_2 is also in B_k. We have proven that this result follows from a certain convexity property possessed by many fitness functions.

7. AN EXAMPLE AND OTHER WORK

We have applied this methodology to a variety of density estimation problems in 1D (mainly time series and the construction of adaptive histograms), 2D (e.g., data from sky surveys), 3D (e.g. data from redshift surveys) and higher dimensions. Space does not permit more than brief mention of one example. Figure 1 shows the Bayesian block analysis of a data set consisting of three dimensional rectangular coordinates of the galaxies with measured redshifts in the first data release from the Sloan Digital Sky Survey. These data are confined to a relatively narrow range of declination, and thus represent a fairly thin slice, here shown in a view perpendicular to the slice. We are developing visualization methods for this block representation, to provide an intuitive picture of the galaxy distribution, free of assumptions about the existence of "clusters" with various symmetry properties. Also, directly from the block representation or by transforming it, one can compute a large variety of derivative statistical quantities describing the 3D galaxy distribution and its topology—correlation and clustering statistics, biasing, genus and genus-related statistics, Minkowski functionals, etc. ([3, 4])

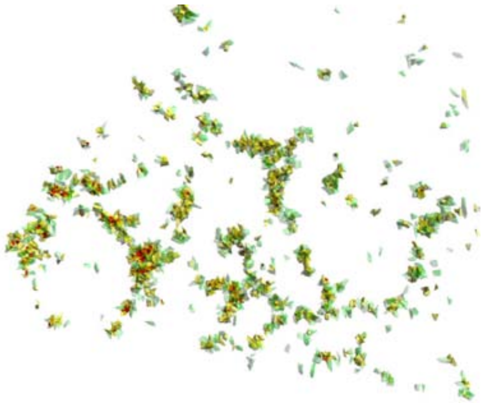


Figure 1: 3D Bayesian Block representation of a section of data from the Sloan Digital Sky Survey. A relatively high density threshold has been set, revealing the skeleton of the distribution.

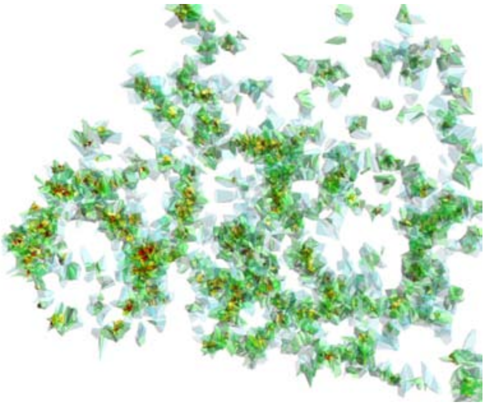


Figure 2: As in Figure 1, but with a lower density threshold, revealing the degree to which these large scale structures are interconnected.

An early version of Bayesian Blocks, based on the greedy algorithm, is in the Astrophysics Source Code Library at `http://ascl.net/block.html` Michael Nowak has developed S code implementing Bayesian blocks in 1D for the S-lang/ISIS Timing Analysis Routines (SITAR) home page `http://space.mit.edu/CXC/analysis/SITAR/` for the Chandra Science Center at MIT. A number of observers have used this approach to study time series data [1, 6, 9, 17, 18].

[4] A set A is *simply connected* if for any partition into two subsets, A_1 and A_2 ($A_1 \cup A_2 = A; A_1 \cap A_2 = \emptyset$), at least one cell in A_1 is adjacent to at least one cell in A_2. For Voronoi cells either of two notions of adjacency can be used: sharing at least one vertex, or sharing at least one face.

Acknowledgments

I am especially grateful to Tom Loredo for many extremely helpful comments on content and presentation. The Applied Information Sciences Research and Intelligent Systems Programs of NASA have supported this work.

References

[1] Bauer, F., and Brandt, W. (2003), "Chandra and HST Confirmation of the Luminous and Variable X-ray Source IC 10 X-1 as a Possible Wolf-Rayet, Black-Hole Binary," Ap. J. Lett., in press, `http://arxiv.org/abs/astro-ph/0310039`

[2] Coram, M. A., *Nonparametric Bayesian Classification*, Ph.D. thesis, Department of Statistics, Stanford University, 2002.

[3] Park, C., Gott, J., and Choi, Y., "Topology of the Galaxy Distribution in the Hubble Deep Fields," (2001), **Ap. J.**, 553, 33

[4] Hikage, C. et al, "Minkowski Functionals of SDSS Galaxies I: Analysis of Excursion Sets," (2003), **P.A.S.J.**, 55, 911

[5] Ebeling, H., and Wiedenmann, G. (1993), "Detecting structure in two dimensions combining Voronoi tessellation and percolation," *Physical Review* E, Volume 47, pp.704-710.

[6] Hambaryan, V., R. Neuhaeuser, R., and Stelzer, B. (1999), "Bayesian flare event detection: ROSAT X-ray observations of the UV Cetus type star G 131-026," Astron. Astrophys., **345**, pp. 121-126

[7] Jackson, B., Scargle, J., Barnes, D., Arabhi, S., Alt, A., Gioumousis, P., Gwin, E., Sangtrakulcharoen, P., Tan, L., and Tao Tsai, T. (2003) "An Algorithm for Optimal Partitioning of Data on an Interval," submitted `http://front.math.ucdavis.edu/math.NA/0309285`

[8] Kim, R., Kepner, J., Postman, M., Strauss, M., Bahcall, N., Gunn, J., Lupton, R., Annis, J., Nichol, R., Castander, F., Brinkmann, J., Brunner, R., Connolly, A., Csabai, I., Hindsley, R., Ivezic, Z., Vogeley, M., and York, D. (2002), "Detecting Clusters of Galaxies in the Sloan Digital Sky Survey. I. Monte Carlo Comparison of Cluster Detection Algorithms," Astronomical Journal, Vol. 123, pp. 20-36.

[9] Kim, D.-W., Cameron, R. A., Drake, J. J., Evans, N. R., Freeman, P., Gaetz, T. J., Ghosh, H., Green, P. J., Harnden, F. R., Jr., Karovska, M., Kashyap, V., Maksym, P. W., Ratzlaff, P. W., Schlegel, E. M., Silverman, J. D., Tananbaum, H. D., Vikhlinin, A. A., and Wilkes, B. J. (2003) Chandra Multi-wavelength Project (ChaMP). I. First X-ray Source Catalog, in preparation, `http://arxiv.org/abs/astro-ph/0308492`.

[10] Okabe, A., Boots, B., Sugihara, K., and Chiu, S. N. (2000), *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams*, John Wiley and Sons, Ltd., New York, Second Edition

[11] Ramella, M., Boschin, W., Fadda, D., and Nonino, M. (2001), "Finding galaxy clusters using Voronoi tessellations," Astronomy and Astrophysics, v.368, p.776-786.

[12] Rissanen, J., 1989, *Stochastic Complexity and Statistical Inquiry*, Singapore: World Scientific.

[13] Scargle, J., 1998, "Studies in Astronomical Time Series Analysis. V. Bayesian Blocks, A New Method to Analyze Structure in Photon Counting Data", *Astrophysical Journal*, **504**, p. 405-418, Paper V. `http://xxx.lanl.gov/abs/astro-ph/9711233`

[14] Scargle, J. D., (2001), Bayesian Blocks: Divide and Conquer, MCMC, and Cell Coalescence Approaches, in *Bayesian Inference and Maximum Entropy Methods in Science and Engineering,* 19th International Workshop, Boise, Idaho, 2-5 August, 1999. Eds. Josh Rychert, Gary Erickson and Ray Smith, AIP Conference Proceedings, Vol. 567, p. 245-256.

[15] Scargle, J. D., (2001a), "Bayesian Estimation of Time Series Lags and Structure," Contribution to **Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering (MAXENT 2001)**, held at Johns Hopkins University, Baltimore, MD USA on August 4-9, 2001.

[16] Scargle, J. D., (2001), "Bayesian Blocks in Two or More Dimensions: Image Segmentation and Cluster Analysis," Contribution to **Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering (MAXENT 2001)**, held at Johns Hopkins University, Baltimore, MD USA on August 4-9, 2001.

[17] Wheatland, M., Sturrock, P., and McTiernan, J. (1998), Ap. J., **509**, p. 448-455.

[18] Wheatland, M. (2000), Ap. J., 536 , L 109-112, 2000, "The Origin of the Solar Flare Waiting-Time Distribution."

Bayesian Adaptive Exploration in a Nutshell

T. J. Loredo
Dept. of Astronomy, Cornell University, Ithaca, NY, 14850, USA

I describe a framework for adaptive scientific exploration based on iterating an *Observation–Inference–Design* cycle that allows adjustment of hypotheses and observing protocols in response to the results of observation on-the-fly, as data are gathered. The framework uses a unified Bayesian methodology for the inference and design stages: Bayesian inference to quantify what we have learned from the available data, and Bayesian decision theory to identify which new observations would teach us the most. When the goal of the experiment is simply to make inferences, the framework identifies a computationally efficient iterative "maximum entropy sampling" strategy as the optimal strategy in settings where the noise statistics are independent of signal properties. Results of applying the method to two "toy" problems with simulated data—measuring the orbit of an extrasolar planet, and locating a hidden one-dimensional object—show the approach can significantly improve observational efficiency in settings that have well-defined, reliable models.

1. INTRODUCTION

The classical paradigm for the scientific method follows a rigid sequence of hypothesis formation, followed by experiment and then analysis. It bears little resemblance to the adaptive, self-adjusting behavior of the human brain, which learns from experience incrementally, making decisions and adjusting questions on-the-fly. The classical paradigm has served science well, but there are many circumstances where what has been learned from past data could be profitably used to alter the collection of future data to more efficiently address the questions of interest.

The idea that use of partial knowledge can improve the design of experiments has long been recognized in statistics; there are well-developed theories of experimental design using both the frequentist and Bayesian approaches to statistics. Unfortunately, practice has lagged theory, largely due to the complicated calculations required for rigorous experimental design with realistic models, particularly in adaptive settings where many designs must be calculated. Until recently most work focused on classes of problems that are analytically tractable (e.g., linear models with normal errors, and, in Bayesian design, with flat or conjugate priors). Treatment of nonlinear models was typically handled only approximately, by linearizing about a best-fit model. This focus has discouraged application to problems of interest to astronomers and physicists, which often have substantial nonlinearities and other complications. In addition, the gains offered by optimal designs in analytically tractable settings are often only modest. Finally, in these settings frequentist and Bayesian designs are the same or very similar, suggesting (erroneously) that the two approaches have little distinguishing themselves in this arena.

In recent years computational and theoretical developments finally enable one to undertake rigorous nonlinear Bayesian design in complicated settings. Here I describe the basic principles behind Bayesian design in an adaptive setting and report results of proof-of-concept calculations showing that Bayesian adaptive exploration (BAE) may improve observational efficiency in a variety of problems in astronomy and physics. A more lengthy treatment of BAE is available in a companion paper (Loredo [2004]).

2. BAYESIAN ADAPTIVE EXPLORATION

BAE iterates an *Observation–Inference–Design* cycle depicted in Figure 1. In the observation stage, new data are obtained based on an observing strategy produced by the previous cycle of exploration. The inference stage synthesizes the information provided by previous and new observations to assess hypotheses of interest. This synthesis produces interim results such as signal detections, parameter estimates, or object classifications. Finally, in the design stage the results of inference are used to predict future data for a variety of possible observing strategies; the strategy that offers the greatest predicted improvement in inferences (subject to any resource constraints) is passed on to the next Observation–Inference–Design cycle.

Bayesian statistics is used for both the inference and design stages. The inference stage uses the tools of Bayesian inference. In particular, Bayes's theorem, which combines prior information and data to produce posterior probabilities for hypotheses of interest, provides a formal description of learning perfectly suited for the tasks of the inference stage. The design stage uses Bayesian decision theory to find an optimal experimental or observational design by first specifying the purpose for a study, and then comparing how well candidate designs achieve that purpose by using the information in existing data to predict future data, and then determining how that data might improve inferences. Bayesian design can rigorously and straightforwardly account for uncertainties in assessing a design even in challenging settings (strongly nonlinear models, non-Gaussian noise). It interfaces naturally with Bayesian inference, so the tools of the inference and design stages work synergistically together.

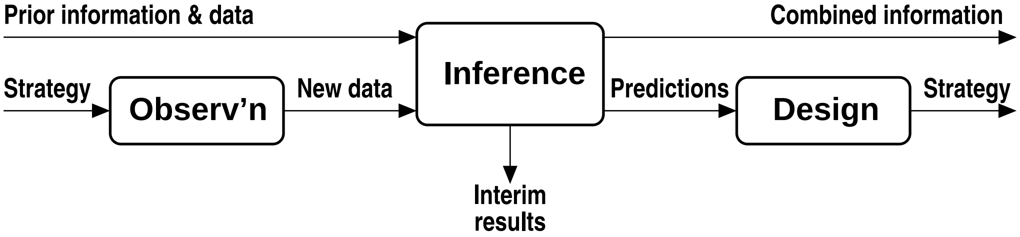


Figure 1: Information flow through one cycle of the adaptive exploration process.

The main ideas of Bayesian inference will be familiar to many readers of this volume, but Bayesian experimental design is not a common tool in the physical sciences, so this brief report will focus on describing the key elements of the design stage (see Chaloner and Verdinelli [1995] for a review of Bayesian experimental design and references). As already noted, Bayesian design is an application of Bayesian decision theory. In decision theory an optimal decision is made in the face of uncertainty by enumerating the possible actions, a (e.g., whether to bet on heads or tails in a coin flip), and the possible outcomes, o, of which we are uncertain (e.g., which side of the coin will come up), and assigning a utility $U(o,a)$ to action a if the outcome turns out to be o (e.g., the amount won or lost). If the available information, I, gives probability $p(o|I)$ to outcome o, then the *expected utility* associated with action a is $EU(a) = \sum_o p(o|I)U(o,a)$. Decision theory specifies that the optimal action is the one that maximizes the expected utility.

In Bayesian design, the space of actions is the space of possible experiments or observations (e.g., the location in time or space for the next sample). The uncertain outcome is the data one would get performing a candidate experiment. An optimal design is found by specifying a utility and maximizing the expected utility. In some settings (e.g., financial or medical experiments), there may be a natural utility function associated with actual costs associated with outcomes. This is seldom true in physics experiments where the goal is simply to gain knowledge about a phenomenon. In 1956, Lindley described how one could use tools from information theory and Bayesian statistics to find optimal designs for achieving such a goal. He later incorporated these ideas into a more general theory of Bayesian experimental design, described in his influential 1972 review of Bayesian statistics (Lindley [1972]). This theory unifies and generalizes non-Bayesian methods for optimal design that predated Lindley's work (see Chaloner and Verdinelli [1995] for discussion of the relationships between Bayesian and non-Bayesian design).

Lindley's idea is to use *information as utility*, with information quantified via information theory. Thus the utility for experiment e (described in the background information, I_e) producing data d is the negative entropy of the posterior distribution for the quantities (hypotheses) of interest, H_i:

$$\mathcal{I}(d,e) = \sum_i p(H_i|d,I_e)\, \log\left[p(H_i|d,I_e)\right]. \qquad (1)$$

The optimal experiment maximizes the expected information,

$$E\mathcal{I}(e) = \sum_d p(d|I_e) \sum_i p(H_i|d,I_e)\, \log\left[p(H_i|d,I_e)\right], \qquad (2)$$

where $p(d|I_e)$ is the predictive distribution for the data which can be calculated using

$$p(d|I_e) = \sum_i p(H_i|I_e)p(d|H_i,I_e), \qquad (3)$$

where $p(H_i|I_e)$ is the prior distribution for H_i given the current information, I_e (when some data is already available in I_e, this is the posterior distribution incorporating that data), and $p(d|H_i,I_e)$ is the sampling distribution for the future data.

Finding optimal designs using $E\mathcal{I}(e)$ requires calculating a triply-nested set of sums or integrals for each candidate design, a computationally challenging task unless the integrals are analytic. Two recent developments finally allow application of this approach to settings of realistic complexity in the physical sciences, where the needed integrals are not analytic. First is the development of sampling-based methods for calculating probability integrals like those appearing here. Physicists are familiar with such methods for calculating integrals over sample space via Monte Carlo generation of simulated data (e.g., for the d sum in (2)). The new spin on this has been the creation of good Monte Carlo algorithms for *hypothesis space* integrals, where the distributions one must sample from are multidimensional and with complicated structure. These are the Markov chain Monte Carlo (MCMC) methods now widely used for Bayesian inference. Muller and his colleagues have pioneered application of MCMC methods to Bayesian design (Muller [1999]).

The second development is the recognition that significant analytical simplification is possible in a restricted but common and very useful setting. It is often the case that the information in the sampling distribution, $p(d|H_i,I_e)$, is independent of H_i. That is, roughly speaking, the width of the noise distribution does not depend on the properties of the underlying signal. This is the case when noise is additive and is dominated by detector or background sources. Sebastiani and Wynn [2000] showed that when this is true, the expected information simplifies,

$$E\mathcal{I}(e) = C - \int dd\, p(d|I_e) \log[p(d|I_e)], \qquad (4)$$

where C is a constant (measuring the e-independent information in the prior and the sampling distribution). Thus the experiment that maximizes the expected information is the one for which the predictive distribution has minimum information, or maximum entropy. The strategy of sampling in this optimal way is called *maximum entropy sampling.* Colloquially, this strategy says you will learn the most by sampling where you know the least, an eminently sensible criterion.

To flesh out these ideas, consider the problem of optimally scheduling observations of a star in order to characterize the orbit of a planet detected via radial velocity measurements of the Keplerian reflex motion of the star. The data are modeled by

$$d_i = V(t_i; \tau, e, K) + e_i, \qquad (5)$$

where $V(t_i; \tau, e, K)$ gives the Keplerian velocity along the line of site as a function of time t_i and of the orbital parameters τ (period), e (eccentricity), and K (velocity amplitude); for simplicity three purely geometric parameters are suppressed. This function is strongly nonlinear in all variables except K. Our goal is to learn about the parameters τ, e and K.

Figure 2 shows results from a typical simulation iterating the BAE observation-inference-design cycle a few times. Figure 2a shows simulated data from a hypothetical "setup" observation stage. Observations were made at 10 equispaced times; the curve shows the true orbit with typical exoplanet parameters (τ = 800 d, e = 0.5, K = 50 ms^{-1}), and the noise distribution is Gaussian with zero mean and σ = 8 m s^{-1}. Figure 2b shows some results from the inference stage using these data. Shown are 100 samples from the marginal posterior density for τ and e (one could smooth this distribution and present contours; this display illustrates the sampling approach behind the algorithm). There is significant uncertainty that would not be well approximated by a Gaussian (even correlated). Figure 2c illustrates the design stage. The thin curves display the uncertainty in the predictive distribution as a function of sample time; they show the $V(t)$ curves associated with 15 of the parameter samples from the inference stage. The spread among these curves at a particular time displays the uncertainty in the predictive distribution at that time. A Monte Carlo calculation of the expected information vs. t (using all 100 samples) is plotted as the thick curve (right axis, in bits, offset so the minimum is at 0 bits). The curve peaks at t = 1925 d, the time used for observing in the next cycle.

Figure 2d shows interim results from the inference stage of the next cycle after making a single simulated observation at the optimal time. The period uncertainty has decreased by more than a factor of two, and the product of the posterior standard deviations of all three parameters (the "posterior volume") has decreased by a factor $\approx$ 5.8; this was accomplished by incorporating the information *from a single well-chosen datum.* Figures 2e,f show similar results from the next two cycles. The posterior volume continues to decrease much more rapidly than one would expect from the random-sampling "$\sqrt{N}$ rule" (by factors of $\approx$ 3.9 and 1.8).

Figure 3 provides a further example motivated by the problem of detecting buried land mines using a mix of technologies—inexpensive but noisy ferromagnetic scans, and more costly but more sensitive acoustic scans using laser doppler vibrometry. Figure 3a shows a hidden 1-d Gaussian (dashed curve; peak at x_0 = 5.2, amplitude A = 7, FWHM = 0.6) barely detected in an initial scan with 11 crude (σ = 1) observations spaced well over a full-width apart. Figure 3b shows samples from the marginal posterior density for A and x_0 from the first inference stage, with very substantial uncertainty. BAE proceeds, designing for subsequent more sensitive observations (σ = 1/3). The design stage produces the predictive distribution (thin curves) and entropy curve (thick curve; right axis) in Figure 3c, suggesting observing near the best guess for the peak. A simulated observation produces the more concentrated but complicated Cycle 2 inference of Figure 3d. Two subsequent cycles identify optimal sample locations that flip-flop to either side of the peak, producing the Cycle 3 and Cycle 4 inferences of Figures 3e and 3f. The posterior volume decreases by factors of $\approx$ 8.2, 6.6, and 5.6 between cycles, far more dramatically than expected from random sampling (even adjusting for the fact that only two of the original samples lie in the signal region). The final posterior distribution is nearly uncorrelated and simple in shape. If for the last step one samples just a few tenths of a unit from the optimal point, the posterior volume is 40% larger and remains strongly correlated.

These examples demonstrate the potential of BAE (and Bayesian design more generally) to greatly improve the return from planned experiments. But many issues must be addressed before the approach can be used efficiently and with confidence, including: adapting the algorithm to changing goals (e.g., from signal detection to signal characterization once the signal is detected); assessing robustness to model uncertainty (a possible "Achilles heel" in many settings); generalizing the utility to incorporate factors such as the cost of observations (financial or temporal); and finding good algorithms for higher dimensional models.

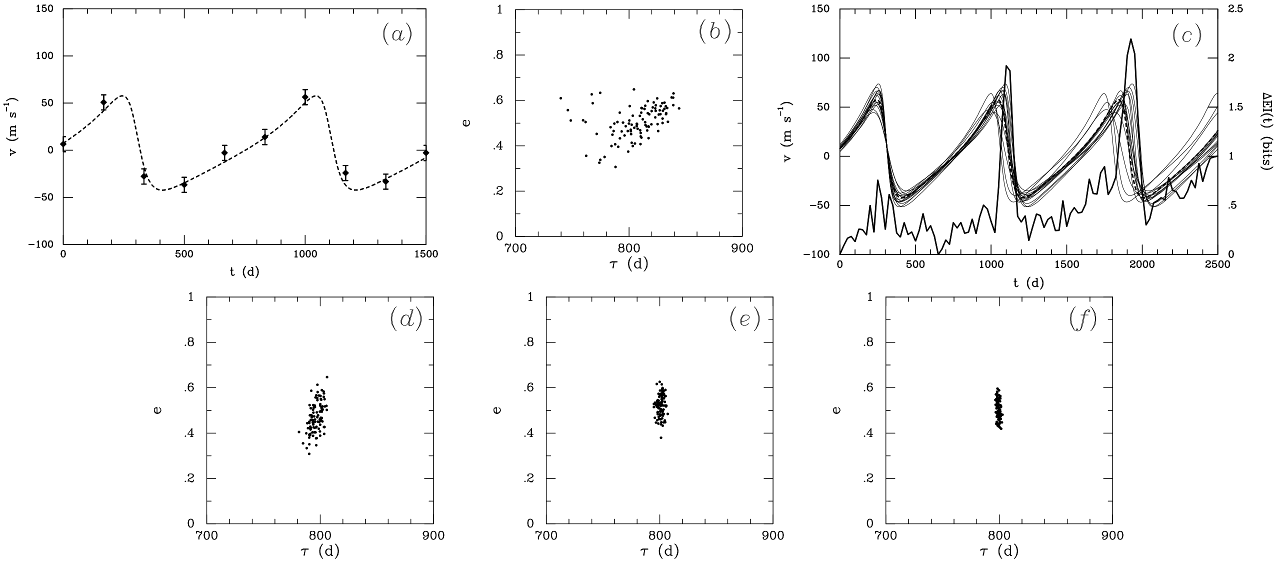


Figure 2: Results from various stages along four observation-inference-design cycles characterizing the orbit of an extrasolar planet with simulated radial velocity observations.

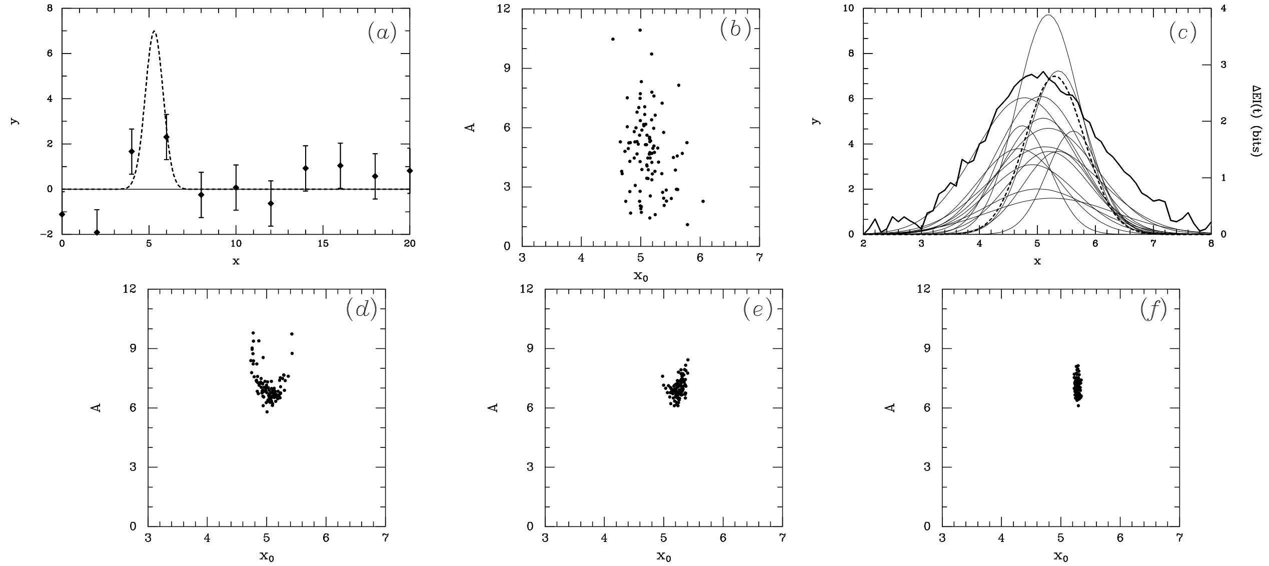


Figure 3: Results from various stages along four observation-inference-design cycles characterizing a hidden 1-d Gaussian object with simulated noisy observations.

Acknowledgments

I thank David Chernoff and Paul Goggins for key input to this work. This work was supported by NASA through grant NAG5-12082 and the SIM EPIcS Key Project.

References

T. J. Loredo, "Bayesian Adaptive Exploration," in *Maximum Entropy and Bayesian Methods, Jackson Hole, Wyoming, 2003*, edited by G. Erickson and Y. Zhai (Kluwer Academic Publishers, Dordrecht, 2004), in press.

K. Chaloner and I. Verdinelli, Stat. Sci. **10**, 273 (1995).

D. V. Lindley, *Bayesian statistics–a review* (SIAM, Philadelphia, 1972).

P. Muller, in *Bayesian Statistics 6*, edited by J. O. Berger, J. M. Bernardo, A. P. Dawid, and A. F. M. Smith (Oxford U. Press, 1999), pp. 459–474.

P. Sebastiani and H. P. Wynn, J. Roy. Stat. Soc. B **62**, 145 (2000).

Blind Analysis in Particle Physics

Aaron Roodman
Stanford Linear Accelerator Center, Stanford, CA 94025, USA

A review of the blind analysis technique, as used in particle physics measurements, is presented. The history of blind analyses in physics is briefly discussed. Next the dangers of *experimenter's bias* and the advantages of a blind analysis are described. Three distinct kinds of blind analysis in particle physics are presented in detail. Finally, the *BABAR* collaboration's experience with the blind analysis technique is discussed.

1. INTRODUCTION

A *blind analysis* is a measurement which is performed without looking at the answer. Blind analyses are the optimal way to reduce or eliminate *experimenter's bias*, the unintended biasing of a result in a particular direction.

In bio-medical research the double-blind randomized clinical trial is the standard way to avoid bias. In such experiments both patients and clinicians are blind to the individual assignments of treatments to patients, and that assignment is made randomly. A double-blind randomized trial was first used in 1948 by Hill in a study of antibiotic treatments for tuberculosis[1]. Amazingly, the concept of a double-blind trial dates back to at least 1662, when John Baptista van Helmont made the following challenge[1]:

> Let us take out of the hospitals, . . . 200, or 500 poor People, that have Fevers, Pleurisies, etc. Let us divide them into half, let us cast lots, that one half of them may fall to my share, and the other to yours; I will cure them without bloodletting and sensible evacuation... We shall see how many funerals both of us shall have. But let the reward of the contention or wager, be 300 florens, deposited on both sides . . .

A notable early use of a blind analysis in physics was in a measurement of the e/m of the electron, by Dunnington [2]. In this measurement, the e/m was proportional to the angle between the electron source and the detector. Dunnington asked his machinist to arbitrarily choose an angle around 340^o. Only when the analysis was complete, and Dunnington was ready to publish a result, did he accurately measure the *hidden* angle.

2. EXPERIMENTER'S BIAS

Experimenter's bias is defined as the unintended influence on a measurement towards prior results or theoretical expectations. Next, we consider some of the ways in which an unintended bias could be present in a measurement.

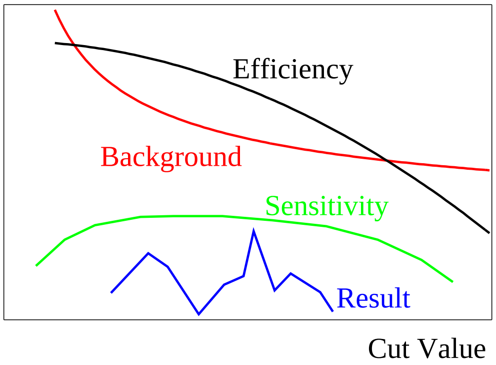


Figure 1: This cartoon illustrates how a result may vary, statistically, for different arbitrary choices of a cut.

One scenario involves the choice of experimental selection requirements, or cuts. Often, a measurement may be equally well done, in terms of sensitivity or uncertainties, with a range of values for a particular selection cut, and the exact cut value used may be chosen arbitrarily. This is illustrated in the cartoon in Fig 1, where there is a plateau in the sensitivity, and the cut value could be chosen anywhere in that plateau. However, the value of the result may vary, typically within the statistical uncertainty, depending on the exact value of cut chosen. If the cuts are set with knowledge of how that choice affects the results, experimenter's bias could occur. In this case, the size of the bias could be on the order of the statistical uncertainty.

Another, less subtle, scenario involves measurements of small signals, such as the search for rare processes or decays. Here experimenter's bias could occur if the event selection is determined with prior knowledge of the effect of that selection on the data. One danger is that the selection cuts can be tuned to remove a few extra background-like events, yielding a result biased to lower limits. Another danger is that the cuts can be tuned to improve the statistical significance of a small signal.

In general, experimenter's bias may occur if obtaining the *correct* result is the standard used to evalu-

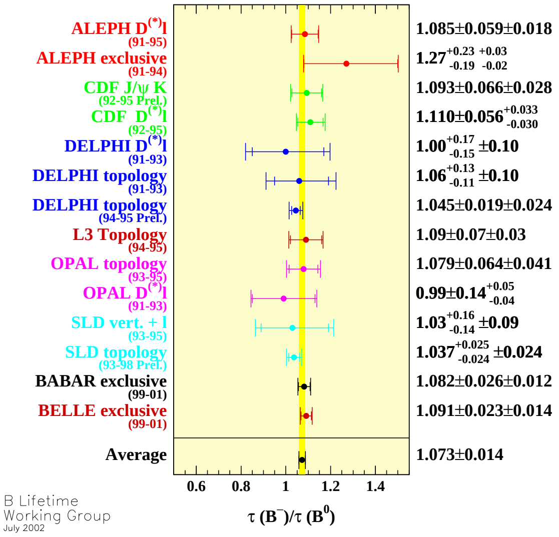


Figure 2: Summary of B meson lifetime ratio measurements. The average has a $\chi^2 = 4.5$ for 13 degrees of freedom.

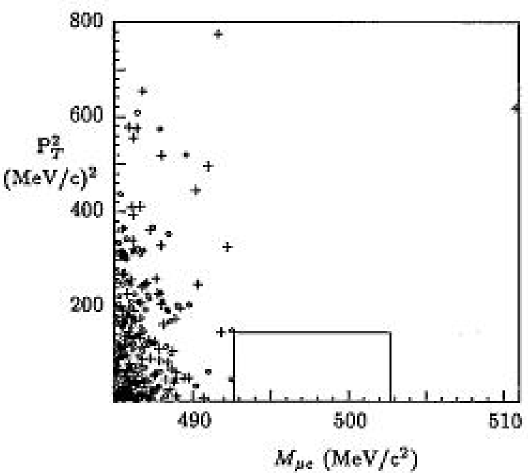


Figure 3: Hidden signal box from a search for the decay $K^0_L \rightarrow \mu^{\pm} e^{\mp}$ from Ref. [4].

ate the quality of the measurement. The validity of a measurement may be checked in a number of ways, such as internal consistency, stability under variations of cuts, data samples or procedures, and comparisons between data and simulation. The numerical result, and how well it agrees with prior measurements or the Standard Model, contains no real information about the internal correctness of the measurement. If such agreement is used to justify the completion of the measurement, then possible remaining problems could go unnoticed, and an experimenter's bias occur.

Does experimenter's bias occur in particle physics measurements? Consider the results on the ratio B meson lifetimes shown in Figure 2. The average has a $\chi^2 = 4.5$ for 13 degrees of freedom; a χ^2 this small or smaller occurs only 1.5% of the time. At this level, the good agreement between measurements is suspicious, but for each individual result no negative conclusion should be made. Nonetheless, it can be argued that even the possibility of a bias represents a problem. The PDG[3] has compiled a number of measurements that have curious time-histories. Likewise, while it is difficult to draw negative conclusions about a single measurement, the overall impression is that experimenter's bias does occur. Finally, there are numerous examples in particle physics of small signals, on the edge of statistical significance, that turned out to be artifacts. Here too, experimenter's bias may have been present.

In all of these cases, the possibility of experimenter's bias is akin to a systematic error. Unlike more typical systematic effects, an experimenter's bias cannot be numerically estimated. Therefore, a technique to reduce or eliminate this bias is needed.

3. BLIND ANALYSIS

A Blind Analysis is a measurement performed without looking at the answer, and is the optimal way to avoid experimenter's bias. A number of different blind analysis techniques have been used in particle physics in recent years. Here, several of these techniques are reviewed. In each case, the type of blind analysis is well matched to the measurement.

3.1. Hidden Signal Box

The *hidden signal box* technique explicitly hides the signal region until the analysis is completed. This method is well suited to searches for rare processes, when the signal region is known in advance. Any events in the signal region, often in two variables, are kept hidden until the analysis method, selection cuts, and background estimates are fixed. Only when the analysis is essentially complete is the box opened, and an upper limit or observation made.

The *hidden signal box* technique was used[1] in a search for the rare decay $K^0_L \rightarrow \mu^{\pm} e^{\mp}$. This decay was not expected to occur in the Standard Model, and the single event sensitivity of the experiment was

[1] This is the first use known to the author.

one event in 10^{11} K_L^0 decays. Any signal was expected inside the box in $M_{\mu e}$ and P_T^2 shown in Figure 3; the possible contents of this box were kept hidden until the analysis was completed[4].

The use of this method is now a standard method for rare decay searches, when the signal region is known in advance. One additional subtlety lies in the size of the hidden box. Generally, the box is initially chosen to be somewhat larger than the signal region, so that the final signal cuts may be chosen without bias as well. Otherwise, this technique is straightforward to apply.

3.2. Hidden Answer

For precision measurements of parameters, a different technique for avoiding bias must be used. In this case, *hiding the answer* is often the appropriate method. The KTeV experiment used this technique in its measurement of ϵ'/ϵ. The value of ϵ'/ϵ was found in a fit to the data, and a small value of order $10^{-4} - 10^{-3}$ was expected. In this case, KTeV inserted an unknown offset into its fitting program, so that the result of the fit was the hidden value:

$$\epsilon'/\epsilon\,(\mathrm{Hidden}) = \left\{ \begin{array}{c} 1 \\ -1 \end{array} \right\} \times \epsilon'/\epsilon + C \qquad (1)$$

where C was a hidden random constant, and the choice of 1 or -1 was also hidden and random. The value of the hidden constant, C, was made by a pseudo-random number generator with a reasonable distribution and mean. KTeV could determine its data samples, analysis cuts, Monte-Carlo corrections, and fitting technique while the result remained hidden, by looking only at $\epsilon'/\epsilon\,(\mathrm{Hidden})$. The use of the 1 or -1 factor prevented KTeV from knowing which direction the result moved as changes were made. In practice, the result[5] was unblinded only one week before the value was announced.

The hidden answer technique is well-suited to precise measurements of a single quantity. The complete analysis, as well as the error analysis, may proceed while blind to the result. An additional consideration is whether there are any distributions which will give away the blinded result. Often the exact value of the measurement is not readily apparent from the relevant plots; in this case those plots can be used without issue.

3.3. Hidden Answer and Asymmetry

For certain measurements hiding the answer is not sufficient; it may also be necessary to hide the visual aspect of the measurement. One example is an asymmetry measurement, such as the recent *CP*-violation measurement by *BABAR*. In this case, the rough size and sign of the asymmetry can be seen by looking at the Δt distributions for B^0 and $\overline{B}^0$ decays into *CP* eigenstates, as shown in Figure 4a. Before *CP* violation had been established, and to avoid any chance of bias, a blind analysis was developed to hide both the answer and the visual asymmetry.

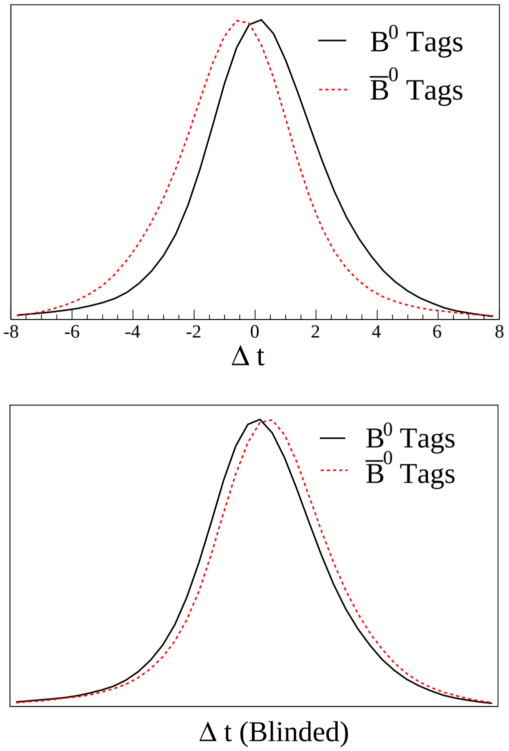


Figure 4: The Δt distributions for B decays into *CP* eigenstates, for sin$2\beta = 0.75$ with the B^0 flavor tagging and vertex resolution which are typical for the *BABAR* experiment. a) The number of B^0 (solid line) and $\overline{B}^0$ (dashed line) decays into *CP* eigenstates as a function of Δt. b) The $\Delta t_{\mathrm{Blind}}$ distributions for B^0 (solid) and $\overline{B}^0$ (dashed).

In *BABAR*'s *CP*-violation measurement the result, found from a fit to the data, was hidden as in Equation 1. In addition, the asymmetry itself was hidden by altering the Δt distribution used to display the data.[7] To hide the asymmetry the variable:

$$\Delta t\,(\mathrm{Blind}) = \left\{ \begin{array}{c} 1 \\ -1 \end{array} \right\} \times s_{\mathrm{Tag}} \times \Delta t + \mathrm{Offset} \qquad (2)$$

was used to display the data. The variable s_{Tag} is equal to 1 or -1 for B^0 or $\overline{B}^0$ flavor tags. Since the asymmetry is nearly equal and opposite for the different B flavors, we hid the asymmetry by flipping one of the distributions. In addition, the *CP*-violation can be visualized by the asymmetry of the individual B^0 and $\overline{B}^0$ distributions. In turn, this was hidden by adding the hidden offset which has the effect of hiding the $\Delta t = 0$ point. The result is shown in Figure 4b, where the amount of *CP*-violation is no longer visible (the remaining difference is due to charm lifetime effects). Also it is worth noting that for a given data sample, due to statistical fluctuations, the maximum of the distribution will not exactly correspond to the $\Delta t = 0$ point, as in the smooth curves shown.

This blind analysis technique allowed *BABAR* to use the $\Delta t_{\mathrm{Blind}}$ distribution to validate the analysis and explore possible problems, while remaining blind to the presence of any asymmetry. There was one additional restriction, that the result of the fit could not be superimposed on the data, since the smooth fit curve would effectively show the asymmetry. Instead to assess the agreement of the fit curve and the data, a distribution of just the residuals was used. In practice, this added only a small complication to the measurement. However, after the second iteration of the measurement, it became clear that the asymmetry would also remain blind if the only Δt distribution used was of the sum of B^0 and $\overline{B}^0$ events, and that no additional checks were needed using the individual Δt distributions.

3.4. Other Blind Methods

The kinds of measurements already discussed, such as rare searches and precision measurements of physical parameters, are well suited to the blind analysis technique. Other kinds of analyses are difficult to adapt to the methods described. For instance, branching fraction measurements typically require the careful study of the signal sample in both data and simulation, so it is not possible to avoid knowing the number of signal events or the efficiency. In this case, other techniques may be considered. One method is to fix the analysis on a sub-sample of the data, and then used the identical method on the full data sample. One may argue about the correct amount of data to use in the first stage, too little and backgrounds or other complications may not be visible, too much and the technique loses its motivating purpose. Another method is to mix an unknown amount of simulated data into the data sample, removing it only when the analysis is complete.

Another difficult example is the search for new particles, or bump-hunting. In this case, since the signal region is not known a-priori, there is no one place to put a hidden signal box. However, such measurements may be the most vulnerable to the effects of experimenter's bias. Certainly, there is some history of statistically significant bumps that are later found to be artifacts. The possibility of using a blind analysis technique may depend on the understanding of the relevant background. If the background can be estimated independently of the bump-hunting region, than the analysis and selection cuts may be set independently of the search for bumps. Here again is a case in which the exact method used must be well matched to the measurement in question.

4. CONCLUSION

The experience of the *BABAR* collaboration in using blind analyses is instructive. While the collaboration had initial reservations about the blind analysis technique, it has now become a standard method for *BABAR* [8]. Often the blind analysis is a part of the internal review of *BABAR* results. Results are presented and reviewed, before they are unblinded, and changes are made while the analysis is still blind. Then when either a wider analysis group or an internal review committee is satisfied with the measurement the result is unblinded, ultimately to be published. With several years of data taking, and many results, *BABAR* has successfully used blind analyses.

Acknowledgments

Work supported by the U.S. Department of Energy under contract number DE-AC03-76SF00515.

References

[1] R. Doll, *Controlled trials: the 1948 watershed*, *British Medical Journal* **318**, 1217, (1998).

[2] F.G. Dunnington, *Phys. Rev.* **43**, 404, (1933). See also L. Alvarez, *Adventures of a Physicist*, (1987).

[3] Review of Particle Properties, *Phys. Rev.* **D66**, 010001-14.

[4] K. Ariska *et al.* [E791 Collaboration], *Phys. Rev. Lett.* **70**, 1049, (1993).

[5] A. Alavi-Harati *et al.* [KTeV Collaboration], *Phys. Rev. Lett.* **83**, 22 (1999).

[6] B. Aubert *et al.* [BABAR Collaboration], *Phys. Rev. Lett.* **86**, 2515 (2001).

[7] A. Roodman, *Blind Analysis of* $\sin 2\beta$, Babar Analysis Document # 41, (2000).

[8] Blind Analysis Task Force [Babar Collaboration], Babar Analysis Document # 91, (2000).

Constraints on Neutrino Mixing Parameters with the SNO data

A. Bellerive
Ottawa-Carleton Institute for Physics, Department of Physics,
Carleton University, 1125 Colonel By Drive, Ottawa, K1S 5B6, Canada

This paper reviews the constraints imposed on the solar neutrino mixing parameters by data collected by the Sudbury Neutrino Observatory (SNO). The SNO multivariate analysis is reviewed. The global solar neutrino analysis is emphasized in terms of matter-enhanced oscillation of two active flavors. An outline of how SNO uses the data to produce oscillation contour plots and how to include the relevant correlations for the new salt data in similar oscillation analyses is summarized.

1. INTRODUCTION

The deficit of detected neutrinos coming from the Sun compared with our expectations based on laboratory measurements, known as the Solar Neutrino Problem, was one of the outstanding problems in basic physics for over thirty years. It appeared inescapable that either our understanding of the energy producing processes in the Sun was seriously defective, or neutrinos, one of the fundamental particles in the Standard Model, had important properties which had not been measured. It was indeed argued by some that we needed to change our ideas on how energy was produced in fusion reactions inside the Sun. Others suggested that the problem arose due to peculiar characteristics of neutrinos such as vacuum or matter oscillations. It is useful to review the evolution of our understanding from the data collected by various solar neutrino experiments. The new analysis of the salt data collected by the Sudbury Neutrino Observatory (SNO) [1] will be described, together with the technique used to combine the results of many solar neutrino experiments.

2. SOLAR NEUTRINOS

The energy in the Sun is produced by nuclear reactions that transform hydrogen into helium. Through the fusion reactions, four protons combine to form a helium nucleus containing two protons and two neutrons. The only reactions that allow this to happen are caused by weak interactions like nuclear beta decay. Each time a neutron is formed, there must be an associated positron and electron neutrino produced. Neutrinos can travel directly from the core of the Sun to the Earth in a about eight minutes and hence provide a direct way to study thermonuclear processes in the Sun. The detailed predictions of the solar electron neutrino flux have been produced by John Bahcall and his collaborators from the 1960's until now. Their calculations are refereed to as the Standard Solar Model (SSM). In this proceeding, the Bahcall-Pinsonneault calculations [2] are compared to experimental results.

It is known that neutrinos exist in different flavors corresponding to the three charged leptons: the electron, muon, and tau particles. If neutrinos have masses, flavor can mix and a neutrino emitted in a weak interaction is represented as a superposition of mass eigenstates. In the case of three flavors of neutrino, the mixing matrix U is called the Maki-Nakagawa-Sakata-Pontecorvo (MNSP) matrix [3] and $\nu_\ell = \sum_i U_{\ell i}|\nu_i\rangle$. Here the neutrino mass eigenstates are denoted by ν_i with $i = 1, 2, 3$, while the flavor eigenstates are labelled (e, μ, τ). The most general form of mixing for three families of neutrinos can be simplified so that only two neutrinos participate in the oscillations. Hence, the survival probability for solar neutrinos propagating in time takes the approximate form

$$P_{e\beta} = \delta_{e\beta} - (2\delta_{e\beta} - 1)\sin^2 2\theta \sin^2(1.27\frac{\Delta m^2 L}{E}) . \quad (1)$$

The mixing angle is represented by θ, L is the distance between the production point of ν_e and the point of detection of ν_β, E is the energy of the neutrino, and $\Delta m^2 \equiv m_j^2 - m_i^2$ is the difference in the squares of the masses of the two states ν_j and ν_i which are mixing. The function $\delta_{e\beta}$ is the usual Kronecker delta. The numerical constant 1.27 is valid for L in meters, E in MeV, and Δm^2 in eV^2. The energy of a neutrino depends on the type of nuclear reaction which produced it. By studying the evolution of the solar neutrinos as a function of L, all the physics is embedded in one angle θ, one mass difference Δm^2, and the sign of Δm^2. This corresponds to the extraction of the three MNSP elements: U_{e1}, U_{e2}, and U_{e3}.

3. SUDBURY NEUTRINO OBSERVATORY

The Sudbury Neutrino Observatory (SNO) is a 1,000 ton heavy-water Čerenkov detector[4] situated 2 km underground in INCO's Creighton mine in Canada. Another 7,000 tons of ultra-pure light water is used for support and shielding. The heavy water is in an acrylic vessel (12 m diameter and 5 cm thick) viewed by 9,456 PMT mounted on a geodesic structure

18 m in diameter; all contained within a polyurethane-coated barrel-shaped cavity (22 m diameter by 34 m high). The solar-neutrino detectors in operation prior to SNO were mainly sensitive to the electron neutrino type; while the use of heavy water by SNO allows neutrinos to interact through charged-current (CC), elastic-scattering (ES), or neutral-current (NC) interactions. The determination of these reaction rates is a critical measurement in determining if neutrinos oscillate in transit between the core of the Sun and their observation on Earth.

During the pure D_2O phase of the experiment, the signal was determined with a statistical analysis based on the direction, $\cos\theta_{\rm sun}$, the position, R, and the kinetic energy, T, of the reconstructed events assuming the SSM energy spectrum shape [5]. The final selection criteria were $T \geq 5$ MeV and $R \leq 550$ cm. The result of the extended maximum-likelihood fit yields [6]

$$\begin{aligned}\Phi_{\rm CC} &= 1.76\,^{+0.06}_{-0.05}\,^{+0.09}_{-0.09}\,\times 10^6\ {\rm cm^{-2}s^{-1}}\,,\\ \Phi_{\rm ES} &= 2.39\,^{+0.24}_{-0.23}\,^{+0.12}_{-0.12}\,\times 10^6\ {\rm cm^{-2}s^{-1}}\,,\\ \Phi_{\rm NC} &= 5.09\,^{+0.44}_{-0.43}\,^{+0.46}_{-0.43}\,\times 10^6\ {\rm cm^{-2}s^{-1}}\,.\end{aligned} \tag{2}$$

The excess of the NC flux over the CC and ES fluxes implies neutrino flavor transformations. There is also a good agreement between the SNO NC flux and the total ^8B flux of $5.05^{+1.01}_{-0.81}$ $\times 10^6$ cm^{-2}s^{-1} predicted by the SSM. A simple change of variables that resolves the data directly into electron and non-electron components [6] indicates clear evidence of solar neutrino flavor transformation at 5.3 standard deviations

$$\phi_e = 1.76\,^{+0.06}_{-0.05}\,^{+0.09}_{-0.09}\,\times 10^6\ {\rm cm^{-2}s^{-1}}\,, \tag{3}$$

$$\phi_{\mu\tau} = 3.41\,^{+0.45}_{-0.45}\,^{+0.48}_{-0.45}\,\times 10^6\ {\rm cm^{-2}s^{-1}}\,. \tag{4}$$

Allowing a time variation of the total flux of solar neutrinos leads to day/night measurements by SNO, which are sensitive to the neutrino type [7]

$$A_{\rm DN}({\rm total}) = (-24.2 \pm 16.1\,^{+2.4}_{-2.5})\,\%\,, \tag{5}$$

$$A_{\rm DN}(e) = (12.8 \pm 6.2\,^{+1.5}_{-1.4})\,\%\,. \tag{6}$$

By forcing no asymmetry in the $\phi_e + \phi_{\mu\tau}$ rate, i.e. $A_{\rm DN}({\rm total}) = 0$, the day/night asymmetry for the electron neutrino is [7] $A_{\rm DN}(e) = (7.0 \pm 4.9\,^{+1.3}_{-1.2})$.

SNO published its first results of the salt phase [1] in coincidence with the PHYSTAT2003 conference. The measurements were made with dissolved $NaCl$ in the heavy water to enhance the sensitivity and signature for neutral-current interactions. Neutron capture on ^{35}Cl typically produces multiple γ rays while the CC and ES reactions produce single electrons. The greater isotropy of the Čerenkov light from neutron capture events relative to CC and ES events allows good statistical separation of the event types. The degree of the Čerenkov light isotropy is determined by the pattern of PMT hits. This separation allows a precise measurement of the NC flux to be made independent of assumptions about the CC and ES energy spectra. To minimize the possibility of introducing biases, SNO performed a blind analysis for the model independent determination of the total active ^8B solar neutrino. In this analysis, events are statistically separated into CC, NC, ES, and external-source neutrons using an extended maximum-likelihood technique based on the distributions of isotropy, $\cos\theta_{\rm sun}$, and radius, R, within the detector. To take into account correlations between isotropy and energy, a 2D joint probability density function (PDF) is constructed. This analysis differs from the analyses of the pure D_2O data [6, 7] since (1) correlations are explicitly incorporated in the signal extraction and (2) the spectral distributions of the ES and CC events are not constrained to the ^8B shape, but are extracted from the data. Čerenkov event backgrounds from $\beta - \gamma$ decays are reduced with an effective electron kinetic energy threshold $T \geq 5.5$ MeV and a fiducial volume with radius $R \leq 550$ cm.

The extended maximum-likelihood analysis gives the following ^8B fluxes [1]

$$\begin{aligned}\Phi_{\rm CC} &= 1.59\,^{+0.08}_{-0.07}\,^{+0.06}_{-0.08}\,\times 10^6\ {\rm cm^{-2}s^{-1}}\,,\\ \Phi_{\rm ES} &= 2.21\,^{+0.31}_{-0.26} \pm 0.10\,\times 10^6\ {\rm cm^{-2}s^{-1}}\,,\\ \Phi_{\rm NC} &= 5.21 \pm 0.27 \pm 0.38\,\times 10^6\ {\rm cm^{-2}s^{-1}}\,.\end{aligned} \tag{7}$$

The systematic uncertainties on the derived fluxes are shown in Table I. These fluxes are in agreement with previous SNO measurements and the SSM. The ratio of the ^8B flux measured with the CC and NC reactions then provides confirmation of solar neutrino oscillations

$$\frac{\Phi_{\rm CC}}{\Phi_{\rm NC}} = 0.306 \pm 0.026 \pm 0.024\,. \tag{8}$$

4. HOW TO USE THE SNO DATA

The SNO CC, ES and NC fluxes are statistically correlated, since they are derived from a fit to a single data set. The statistical correlation coefficients between the fluxes in the salt phase are

$$\begin{aligned}\rho_{\rm CC,NC} &= -0.521\,,\\ \rho_{\rm CC,ES} &= -0.156\,,\\ \rho_{\rm ES,NC} &= -0.064\,.\end{aligned} \tag{9}$$

These can be used with the statistical uncertainties quoted by SNO [1] to write down the statistical covariance matrix for the salt fluxes. Systematic uncertainties between fluxes can be correlated as well. Some sources of systematic error, such as neutron capture efficiency, affect only one of the three fluxes, and so can be considered to be uncorrelated with the

Source	NC	CC	ES
Energy scale	-3.7,+3.6	-1.0,+1.1	±1.8
Energy resolution	±1.2	±0.1	±0.3
Energy non-linearity	±0.0	-0.0,+0.1	±0.0
Radial accuracy	-3.0,+3.5	-2.6,+2.5	-2.6,+2.9
Vertex resolution	±0.2	±0.0	±0.2
Angular resolution	±0.2	±0.2	±2.4
Isotropy mean	-3.4,+3.1	-3.4,+2.6	-0.9,+1.1
Isotropy resolution	±0.6	±0.4	±0.2
Radial energy bias	-2.4,+1.9	±0.7	-1.3,+1.2
Vertex Z accuracy	-0.2,+0.3	±0.1	±0.1
Internal neutrons	-1.9,+1.8	±0.0	±0.0
Internal background	±0.1	±0.1	±0.0
Neutron capture	-2.5,+2.7	±0.0	±0.0
Čerenkov backgrounds	-1.1,+0.0	-1.1,+0.0	±0.0
AV events	-0.4,+0.0	-0.4,+0.0	±0.0
Total uncertainty	-7.3,+7.2	-4.6,+3.8	-4.3,+4.5

Table I Systematic uncertainties (in %) on fluxes for the spectral shape unconstrained analysis of the salt data set.

Source	NC	CC	ES
Energy scale	+1	+1	+1
Energy resolution	+1	+1	+1
Energy non-linearity	+1	+1	+1
Radial accuracy	+1	+1	+1
Vertex resolution	+1	+1	+1
Angular resolution	+1	+1	−1
Isotropy mean	+1	−1	−1
Isotropy resolution	+1	+1	+1
Radial energy bias	+1	+1	+1
Vertex X accuracy	+1	+1	+1
Vertex Y accuracy	+1	+1	+1
Vertex Z accuracy	+1	−1	−1
Internal neutrons	+1	0	0
Internal background	+1	+1	+1
Neutron capture	+1	0	0
Čerenkov backgrounds	+1	+1	+1
AV events	+1	+1	+1

Table II Signs of systematic correlations, relative to its effect on the NC flux. An entry of +1 indicates a 100% positive correlation, −1 a 100% negative correlation, and 0 means no correlation.

other fluxes. Other systematics can be either 100% correlated (e.g. radial accuracy) or 100% anticorrelated (e.g. isotropy mean). The most important anticorrelated systematic is the isotropy mean. Isotropy is important for separating CC and ES events from NC events, so CC and ES will have a negative correlation with the NC flux (and a positive correlation with each other) for the isotropy uncertainty. Table II shows the sign of the correlation for each systematic of Table I. Using the table of systematics and the signs for the correlations, one can assemble an individual covariance matrix for each systematic. Then, to get the total covariance matrix for the CC, ES and NC fluxes, one simply adds all of the covariance matrices together.

Even when fluxes are being analyzed as opposed to energy spectra, it is best to determine the effect of energy-related systematics at each grid point in the $\Delta m^2 - \tan^2\theta$ plane. For the salt analysis, these include energy scale and energy resolution; the uncertainty due to energy non-linearity is tiny so that it can reasonably be ignored. The energy scale uncertainty is implemented as a 1.1% uncertainty in the total energy; while the energy resolution has an uncertainty which is energy dependent for $T > 4.975$ MeV

$$\frac{\Delta\sigma_T}{\sigma_T} = 0.035 + 0.00471 \times (T - 4.975)\,, \qquad (10)$$

and $\frac{\Delta\sigma_T}{\sigma_T} = 0.034$ for $T < 4.975$ MeV. Here T is the reconstructed kinetic energy. For all other systematics, it is assumed that the effect on the fluxes is the same for all oscillation parameters.

When SNO quotes $\Phi_{\rm CC} = 1.59 \times 10^6$ cm^{-2}s^{-1}, it refers to the integral flux from zero to the endpoint assuming an undistorted ^8B spectrum. It implies that the number of events attributed to CC interactions above $T = 5.5$ MeV is equal to the number of events that would be observed if the ν_e flux follows the ^8B spectral shape. The ^8B spectral shape aspect of this definition is only for normalization; there is no assumption of any spectral shape when extracting the number of events during the salt phase. Similar definitions apply for the NC and ES fluxes.

For the comparison of the SNO CC rate with the theoretical rates for a set of oscillation parameters, the $\Phi_{\rm CC}$ flux is

$$f_B \int_0^\infty \phi_{\rm SSM}(E_\nu)\, dE_\nu\, S(T, T_e, E_\nu)\,, \qquad (11)$$

with the scale $S(T, T_e, E_\nu)$ is equal to

$$\frac{\int_0^\infty \int_0^\infty \int_{5.5}^\infty F(T, T_e, E_\nu) P_{ee}(E_\nu) dT dT_e dE_\nu}{\int_0^\infty \int_0^\infty \int_{5.5}^\infty F(T, T_e, E_\nu) dT dT_e dE_\nu}\,, \qquad (12)$$

where

$$F(T, T_e, E_\nu) = \phi_{\rm SSM}(E_\nu)\, \frac{d\sigma(E_\nu, T_e)}{dT_e}\, N(T_e, \sigma_T^2)\,. \qquad (13)$$

The factor f_B allows the total ^8B solar neutrino flux to float from the SSM value, E_ν is the neutrino energy, P_{ee} is the survival probability, T_e is the true recoil electron kinetic energy, and T is the observed electron kinetic energy; while $N(T_e, \sigma_T^2)$ is a Gaussian energy response function for T with $\sigma_T(T) =$

$-0.145 + 0.392\sqrt{T} + 0.0353T$. It is a similar definition for the SNO ES flux, remembering to include the contribution from $\nu_{\mu\tau}$ using the appropriate cross section and $(1 - P_{ee})$. There is no ambiguity in interpreting NC flux since it is equal to the total SSM flux.

5. GLOBAL FITS

This section summarizes the constraints from solar neutrino data in a global analysis. The allowed region in the oscillation $\Delta m^2 - \tan^2\theta$ plane is obtained by comparing the measured rates to the calculated SSM solar neutrino rate. We consider a set of N observables R_n for $n = 1, 2, \cdots, N$ with the associated set of experimental observations $R_n^{\rm exp}$ and theoretical predictions $R_n^{\rm th}$. In general, one wants to build a χ^2 function which measures the differences $R_n^{\rm exp} - R_n^{\rm th}$ in units of the total experimental and theoretical uncertainties. This task is completely determined from the estimated uncorrelated errors u_n and a set of correlated systematic errors c_n^k caused by K independent sources. The correlation coefficients between the different observables are $\rho(u_n, u_m) = \pm\delta_{nm}$ and $\rho(c_n^k, c_m^h) = \pm\delta_{kh}$. The covariance matrix takes the form $\sigma_{nm}^2 = \delta_{nm} u_n u_m + \sum_{k=1}^{K} c_n^k c_m^h$ and all the experimental information is combined together in a global χ^2

$$\chi^2_{\rm cov} = \sum_{n,m=1}^{N} (R_n^{\rm exp} - R_n^{\rm th})[\sigma_{nm}^2]^{-1}(R_m^{\rm exp} - R_m^{\rm th})\,. \quad (14)$$

The salt shape-unconstrained fluxes presented here, combined with shape-constrained fluxes and day/night energy spectra from the pure D_2O phase [6, 7], place impressive constraints on the allowed neutrino flavor mixing parameters. In the fit, the ratio f_B of the total ^8B flux to the SSM value is a free parameter together with the mixing parameters. A combined χ^2 fit to SNO D_2O and salt data alone yields the allowed regions in Δm^2 and $\tan^2\theta$ shown in Fig. 1. There are certainly correlations between the salt and the D_2O phase, since it's the same detector. However, these correlations are estimated to be negligibly small.

The $\chi^2_{\rm cov}$ calculated above from the SNO NC, CC and ES fluxes is added to a global analysis which includes data from all the other solar neutrino experiments. Systematic errors that are correlated between different experiments, such as cross section uncertainties or uncertainties on the ^8B, are accounted for by including the covariance terms between different experimental results. The effect of the ^8B spectral shape uncertainty is determined at each grid point in the oscillation plane.

The global analysis includes the Homestake results [8], the updated Gallium flux measurements [9, 10], the SK zenith spectra [11], and the D_2O and salt results from SNO [1, 6, 7]. At each grid point in the $\Delta m^2 - \tan^2\theta$ plane, the expected rate for each energy bin is calculated and compared to the measured rate. The free parameters in the global fit are the total ^8B flux, the difference of the squared masses Δm^2, and the mixing angle θ. The higher energy *hep* ν_e flux is fixed at 9.3×10^3 cm^{-2} s^{-1}. Contours are generated in Δm^2 and $\tan^2\theta$ for $\Delta\chi^2_{\rm cov} = 4.61$ (90% CL), 5.99 (95% CL), 9.21 (99% CL), and 11.83 (99.73% CL). We assume a Gaussian distribution of $R_n^{\rm exp}$ for a given value of the true parameters δm^2 and $\tan^2\theta$ when we map the survival probability into the MSW plane [12]. As presented in Fig 2(a), the combined results of all solar neutrino experiments can be used to determine a unique region of the oscillation parameters; the allowed region in this parameter space shrinks considerably to a portion of the Large Mixing Angle (LMA) region.

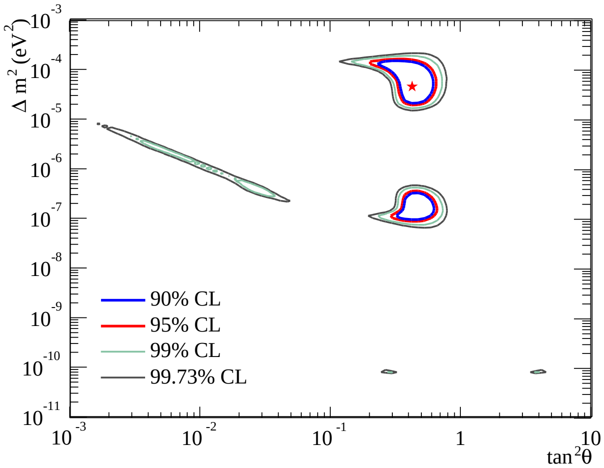


Figure 1: SNO-only neutrino oscillation contours, including pure D_2O day/night spectra, salt CC, NC, ES fluxes, with ^8B flux free and *hep* flux fixed. The best-fit point is $\Delta m^2 = 4.7 \times 10^{-5}$, $\tan^2\theta = 0.43$, $f_B = 1.03$, with χ^2/d.o.f.=26.2/34. The inside of the covariance regions is allowed.

A global analysis including the KamLAND reactor anti-neutrino results [13] shrinks the allowed region further, with a best-fit point of $\Delta m^2 = 7.1^{+1.2}_{-0.6} \times 10^{-5}$ eV^2 and $\theta = 32.5^{+2.4}_{-2.3}$ degrees, where the errors reflect 1σ constraints on the 2-dimensional region. This is summarized in Fig. 2(b). With the new SNO measurements, the allowed region is constrained to only the lower band of LMA at $> 99\%$ CL. The best-fit point with a one dimensional projection of the uncertainties in the individual parameters (marginalized uncertainties) is $\Delta m^2 = 7.1^{+1.0}_{-0.3} \times 10^{-5}$ eV^2 and $\theta = 32.5^{+1.7}_{-1.6}$ degrees. This disfavors maximal mixing at a confidence level equivalent to 5.4 standard deviations and indicates $\tan^2\theta < 1$. In our interpretation, the $\chi^2_{\rm cov}$ for $\theta = 45.0$ is 5.4^2 higher than the

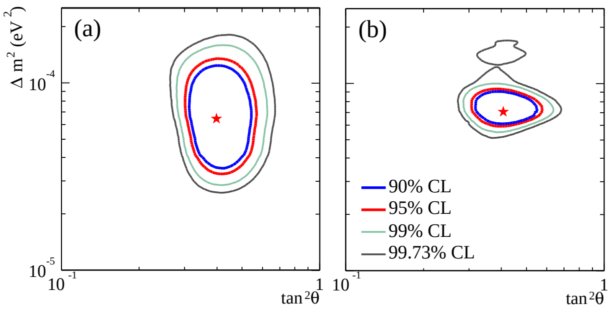


Figure 2: Allowed region of the $\Delta m^2 - \tan^2\theta$ plane determined by a χ^2 fit to (a) the Chlorine, Gallium, SK, and SNO experiments. The best-fit point is $\Delta m^2 = 6.5 \times 10^{-5}$, $\tan^2\theta = 0.40$, $f_B = 1.04$, with χ^2/d.o.f.=70.2/81. (b) Solar global + KamLAND. The best-fit point is $\Delta m^2 = 7.1 \times 10^{-5}$, $\tan^2\theta = 0.41$, $f_B = 1.02$. The inside of the covariance contours is the allowed region.

best LMA fit. The solution $\tan^2\theta < 1$ corresponds to the neutrino mass hierarchy $m_2 > m_1$.

6. PULL ANALYSIS

The pull method allows a split of the residuals from the observables and the systematic uncertainties [14]. This alternative approach embeds the effect of each independent $k^{\rm th}$ source of systematics through a shift of the difference $(R_n^{\rm exp} - R_n^{\rm th})$ by an amount $\epsilon_k c_n^k$. The normalization condition for the K independent sources of systematic uncertainty is implemented through quadratic penalties in the global χ^2, which is minimized with respect to all ϵ_k's

$$\chi^2_{\rm pull} = \sum_{n=1}^{N}\left(\frac{R_n^{\rm exp} - R_n^{\rm th} - \sum_{k=1}^{K}\epsilon_k c_n^k}{u_n}\right)^2 + \sum_{k=1}^{K}\epsilon_k^2 \,. \tag{15}$$

In an experimental context, the pull approach is not blind since it uses the data to constrain the systematic uncertainties. Systematic shifts calculated with the pull method should not be used as iterative corrections to experimental systematic uncertainties since it might lead to biases in the estimation of the mixing parameters. Nevertheless, the pull approach provides a nice framework to study each component of a global fit after a detailed study of the systematic uncertainty of each observables. See details in Ref. [14].

7. SUMMARY

A summary of how to use the new salt data published by SNO is described in the context of solar neutrino analyses of matter-enhanced oscillation of two active flavors. Solar neutrino oscillation is clearly established by SNO. Matter effects [15] explain the energy dependence of solar oscillations with Large Mixing Angle (LMA) solutions favored. The global analysis of the solar and reactor neutrino results yields $\Delta m^2 = 7.1^{+1.0}_{-0.3} \times 10^{-5}$ eV^2 and $\theta = 32.5^{+1.7}_{-1.6}$ degrees.

SNO is presently analyzing its full salt data set with a detailed treatment of the day/night and spectral information. In the future SNO will perform a global oscillation fit with a maximum-likelihood method.

Acknowledgments

This article builds upon the careful and detailed work of many people. Special thanks for the contributions of M. Boulay, M. Chen, S. Oser, Y. Takeuchi, G. Tešić, and D. Waller. This research has been financially supported in Canada by the Natural Sciences and Engineering Research Council (NSERC), the Canada Research Chair (CRC) Program, and the Canadian Foundation for Innovation (CFI).

References

[1] Q.R. Ahmad *et al.*, Submitted to *Phys. Rev. Lett.*, Sept. 2003, nucl-ex/0309004.

[2] J.N. Bahcall, H.M. Pinsonneault, and S. Basu, *Astrophys. J.* **555**, 990 (2001).

[3] Z. Maki, M. Nakagawa, S. Sakata, *Prog. Theor. Phys.* **28**, 870 (1962); B. Pontecorvo, *Sov. Phys. JETP* **26**, 984 (1968).

[4] J. Boger *et al.*, *Nucl. Inst. Meth.* A **449**, 172 (2000).

[5] C.E. Ortiz *et al.*, *Phys. Rev. Lett.* **85**, 2909 (2000).

[6] Q.R. Ahmad *et al.*, *Phys. Rev. Lett.* **89**, 011301 (2002).

[7] Q.R. Ahmad *et al.*, *Phys. Rev. Lett.* **89**, 011302 (2002).

[8] B.T. Cleveland *et al.*, *Ap. J.* **496**, 505 (1998).

[9] V. Gavrin, 4th International Workshop on Low Energy and Solar Neutrinos, Paris, May 2003.

[10] T. Kirsten, XXth Int. Conf. on Neutrino Phy Astrophysics, Munich, May 2002; *Nucl. Phys.* B (Proc. Suppl.) **118** (2003).

[11] S. Fukuda *et al.*, *Phys. Rev. Lett.* **86**, 5651 (2001); S. Fukuda *et al.*, *Phys. Lett.* B **539**, 179 (2002).

[12] P. Creminelli, G. Signorelli, and A. Strumia, *JHEP* **05**, 052 (2001).

[13] K. Eguchi *et al.*, *Phys. Rev. Lett.* **90**, 021802 (2003).

[14] G.L. Fogli, E. Lisi, A. Marrone, D. Montanino, and A. Palazz, *Phys. Rev.* D **66**, 053010 (2002).

[15] S.P. Mikheyev and A.Yu. Smirnov, *Sov. J. Nucl. Phys.* **42**, 913 (1985); L. Wolfenstein, *Phys. Rev.* D **17**, 2369 (1978).

A Feldman-Cousins Likelihood Analysis of Soudan 2 Data for Atmospheric Neutrino Oscillations

Peter J. Litchfield
Minnesota University, Minneapolis, MN 55455, USA

A bin-free Feldman-Cousins style likelihood analysis has been carried out on Soudan 2 atmospheric neutrino data. Reference is given to a full description of the statistical methods used.

A likelihood analysis scheme has been developed for the determination of neutrino oscillation parameters from atmospheric neutrino data using the Feldman-Cousins prescription [1]. The features that distinguish this analysis are;

1. the data are used without binning,
2. background events are fully integrated into the formalism
3. pdfs for the likelihood function are calculated by gaussian smearing of Monte Carlo and background data,
4. nuisance parameters (background fractions and data normalization) are determined in an integrated manner,
5. systematic errors in calibration, fluxes and cross-sections are incorporated in the Feldman-Cousins analysis and
6. the Feldman-Cousins scheme gives proper coverage for the 90% confidence region determination and for the discrimination against the no oscillation hypothesis.

The resultant 90% confidence levels on the oscillation parameters were shown and it was demonstrated that the data agreed with the expected sensitivity of the data and that not using the full Feldman-Cousins procedure would have seriously underestimated the allowed region.

Full details of the analysis are given in reference [2], including the full mathematical formalism.

References

[1] G.J.Feldman and R.D. Cousins, Phys. Rev. D**57**,3873 (1998).

[2] M. Sanchez *et al*, Phys ReV D**68**, to be published, and hep-ex/0307069

Statistical Issues in High-Energy Gamma-Ray Astronomy for GLAST

S. W. Digel
Hansen Experimental Physics Laboratory, Stanford University, Stanford, CA 94305, USA

This paper describes the statistical issues involved in analyzing data from high-energy gamma-ray telescopes, at levels from event reconstruction to correlations of populations of astrophysical sources. Some motivation for attempting to do astronomy with high-energy gamma rays is also given, along with some of the constraints implied by operating the instrument in orbit. Specific attention is given to the Large Area Telescope (LAT) under development for launch in late 2006 on the Gamma-ray Large Area Space Telescope (GLAST) mission.

1. INTRODUCTION

Gamma-ray astronomy has developed only relatively recently owing to many technical challenges in detecting gamma rays. In the energy range below ∼50 GeV, gamma-ray detectors must be in space. (At higher energies, air showers from interactions of gamma rays with the upper atmosphere can be detected from the ground.) The missions that have flown to date with sensitivity in the >20 MeV range (Table 1), in particular the Energetic Gamma-Ray Experiment Telescope (EGRET) on the Compton Gamma-Ray Observatory, have revealed a remarkable variety of astrophysical sources of high-energy gamma rays, and plausible prospective source classes remain to be discovered with future missions, in particular the Large Area Telescope (LAT) on the Gamma-ray Large Area Space Telescope (GLAST) [1, 2], which promises a great increase in sensitivity.

1.1. Sources of Celestial Gamma Rays

Astrophysical sources of gamma rays are nonthermal, accelerating particles in shocks, e.g., in jets and supernova remnants, and in the intense fields of pulsars (rotating magnetized neutron stars). Gamma rays are produced in these sources by high-energy electrons via bremsstrahlung scattering on nucleons, or inverse Compton scattering low-energy photons, or in the case of sufficiently strong magnetic fields via synchrotron or curvature radiation. They are also produced in pion decay from high-energy proton-nucleon interactions. The universe is essentially transparent to gamma rays, and their observation provides a unique, direct probe of these processes in nature. (At ∼TeV energies and cosmological distances, attenuation does occur by γ-γ interactions on the cosmic microwave background.) The known or prospective classes of celestial gamma-ray sources for the next generation of instruments are described briefly below.

1.1.1. Diffuse gamma-ray emission

Interactions of Galactic cosmic rays with interstellar gas and low-energy photons make the Milky Way itself a diffuse source of high-energy gamma rays. The intensity is greatest at low Galactic latitudes, owing to the concentration of the interstellar gas and sources of cosmic rays in the spiral arms of the Milky Way.

1.1.2. Active galaxies

With EGRET, a class of active galaxies called blazars was discovered to be a powerful source of gamma rays [3, 4]. Approximately 70 were identified in EGRET data. The generally-accepted interpretation is that blazars have jet emission associated with a massive black holes in their nuclei, and that the jets are closely aligned with the line of sight to the earth. Other active galaxies, with less favorable alignment of their jets, are also known gamma-ray sources, although much less intense; the only example from EGRET is Centaurus A [5].

1.1.3. Gamma-Ray Bursts

These are extremely bright, short-lived sources, most of which have been identified with some kind of cataclysmic explosions in star-forming galaxies at cosmological distances.

1.1.4. Pulsars

In the Milky Way, a subset of rotation-powered pulsars comprise a well-established class of gamma-ray sources, with approximately 9 identified in the EGRET data. The usual method of discovery is to phase fold the gamma rays according to timing information derived from radio observations (see Sec. 5.2). At least one gamma-ray pulsar, Geminga, is not a radio pulsar; searching for periodicity without a timing ephemeris is discussed in Section 5.2.

1.1.5. Other classes

In general, Galactic sources are associated with tracers or remnants of massive star formation—pulsars (possibly radio quiet), binary pulsars, millisecond pulsars, supernova remnants, plerions (filled center supernova remnants powered by a pulsar), OB/WR associations, microquasars, microblazars, and isolated black holes have all been proposed as sources of gamma rays in the Galaxy.

Table I High-Energy Gamma-Ray Astronomy Missions

Instr.	**Years**	$\theta_{0.1}$[a]	θ_{10}[b]	**Energies (GeV)**	$A_{eff}\Omega$ **cm^2 sr**	**No. γ-Rays**
OSO-3	1967–68	18°	-	>0.05	1.9	621
SAS-2	1972–73	7	-	0.03–10	40	$\sim1\times10^4$
COS-B	1975–82	7	-	0.03–10	40	$\sim2\times10^5$
EGRET	1991–00	5.8	0.5°	0.03–10	750	1.4×10^6
AGILE	2005–	4.7	0.2	0.03–50	1500	4×10^6/yr
AMS	2005[c]–	-	0.1	1–300	500	$\sim2\times10^5$/yr
LAT	2007–	3.5	0.1	0.02–300	25,000	1×10^8/yr

[a] Angular resolution at 0.1 GeV
[b] Angular resolution at 10 GeV
[c] Scheduled for the 16th shuttle mission once launches resume.

Outside the Milky Way, Galaxy clusters and starburst galaxies are prospective new classes of gamma-ray sources. EGRET detected the Large Magellanic Cloud in the light of its diffuse emission.

EGRET detected impulsive GeV emission from intense, X-class flares that occurred near solar maximum in 1991 [6].

Figure 1 shows the gamma-ray sky seen by EGRET and Figure 2 shows a simulated all-sky image from the planned one-year sky survey with the LAT. Projections are that the LAT will detect several thousand gamma-ray sources. Owing to the scanning coverage of the sky survey, the LAT will also provide extremely well sampled light curves.

1.2. History of high-energy gamma-ray astronomy in space

Celestial gamma rays were first detected by OSO-3, which saw the diffuse emission of the Milky Way in 1968 [7], and there have been three missions since of increasing size and resolution. Table I summarizes the past and upcoming missions; all have used pair conversion trackers (see Sec. 2). OSO-3 used plastic scintillators as the tracking material. SAS-2, COS-B, and EGRET had wire grid spark chambers for tracking. Upcoming missions will use silicon strip detectors, which have the advantages of finer pitch (for better angular resolution), orders of magnitude faster readout (for limiting the dead time), and no reliance on expendables (like spark chamber gas).

Three missions in Table I are under development. The LAT, under development for launch by NASA in early 2007 on the GLAST mission, will provide a great increase in sensitivity . The development of the LAT is being supported by NASA, DOE; CEA and IN2P3 in France; ASI, CNR, and INFN in Italy; and institutions in Japan and Sweden. Astro-rivelatore Gamma a Immagini LEggero (AGILE), under development by ASI [8], is a smaller instrument of generally similar design that is planned for launch approximately one year before GLAST. The Alpha Magnetic Spectrometer (AMS) is a cosmic-ray experiment to be launched on the space shuttle for installation on the ISS [9]. It will have sensitivity to gamma rays in the range >1 GeV.

AGILE and the LAT have very large fields of view relative to preceding instruments because they will not rely on time-of-flight (TOF) scintillators below the tracking section to discriminate upward-moving, i.e., background, events. The field of view is limited because events must cross both the tracker and the TOF system in order to be accepted. The trade off is a greater event rate, and the need to rely on post processing to reject upward-moving events.

2. DETECTION OF HIGH-ENERGY GAMMA RAYS

At X-ray energies, photons can be focused with grazing incidence mirrors, but gamma rays cannot be focused similarly. The collecting area of a gamma-ray telescope is therefore directly related to the physical size of the detector, which is not the case in X-ray astronomy.

X-ray detectors more or less can count individual X-rays. Gamma-ray detectors convert the gamma rays to positron-electron pairs, then track their trajectories through the instrument (see Fig. 3) and measure their energies to infer the directions and energies of the gamma rays. Most conversions of the gamma rays happen in heavy metal (W in the case of the LAT) foils interleaved with the tracking layers (silicon strip detectors for the LAT). The trade off for including conversion foils to increase the probability of conversion is that the electron and positron tend to scatter on passage through the foils in subsequent layers, decreasing the accuracy with which their directions and energies can be determined. As a result, gamma-ray telescopes have much poorer angular resolution than X-ray instruments, typically measured in degrees rather than arcseconds (Table I).

The LAT has a modular design, arranged as a 4×4 grid of independent towers, each with a tracker (TKR) and calorimeter (CAL) section (Fig. 3). The TKR section of each tower has 18 tracking planes, each with two layers of silicon strip detectors, one for measuring x coordinates and the other for y, with W foils interleaved between the planes. The CAL section of each tower has CsI(Tl) crystals arranged as a hodoscope: 8 layers of 12 crystals each, with the orientations of the layers alternating between the x and y directions. Each end of each log is instrumented with PIN photodiodes to detect the scintillations. The anticoincidence detector (ACD) surrounds the top and sides of

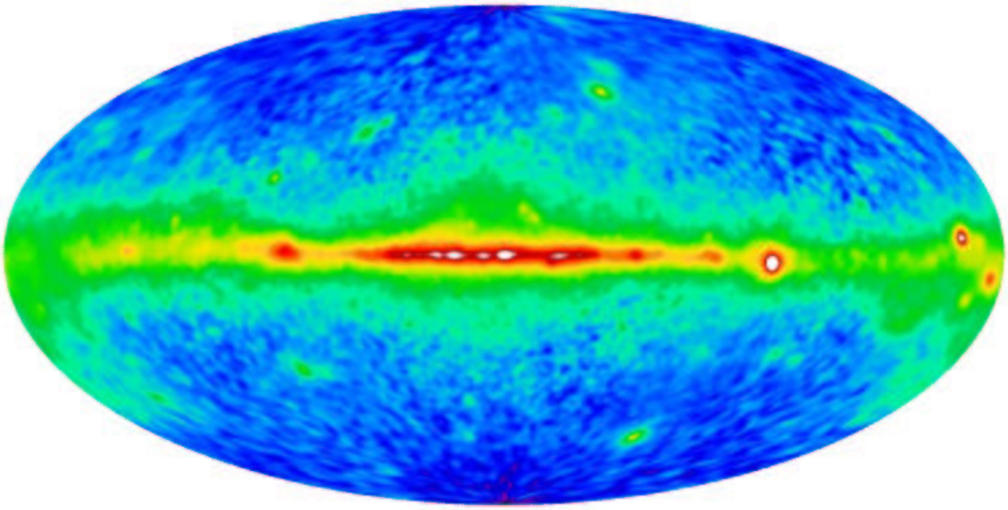


Figure 1: Intensity of gamma-ray emission >100 MeV observed by EGRET, displayed in false color. The Aitoff projection is in Galactic coordinates, and the bright band across the center of the image is the diffuse emission from the Galactic plane. The bright point sources at low latitude are rotation-powered pulsars. Many of the bright sources removed from the plane are blazars.

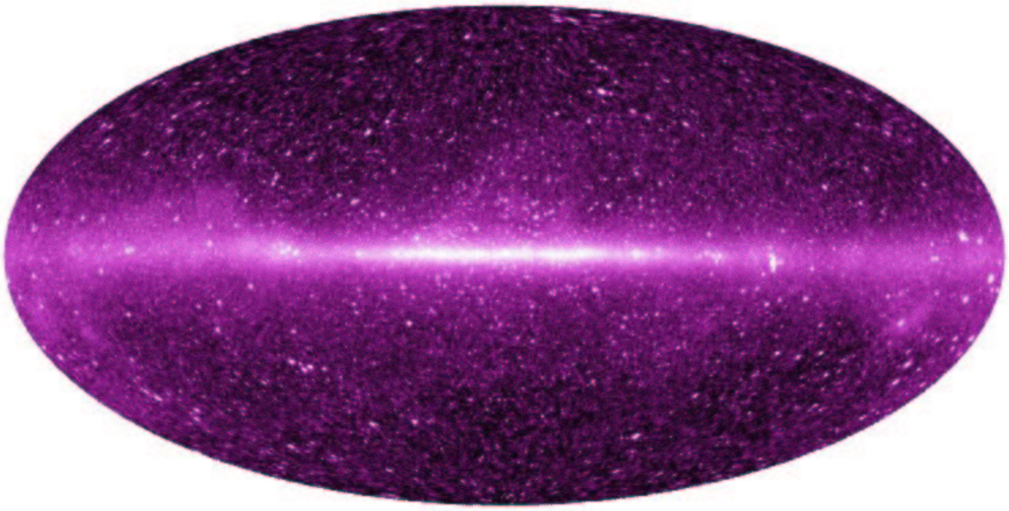


Figure 2: Simulated gamma-ray intensity observed by the LAT during the planned one-year sky survey. The energy range shown here is >1 GeV, where the angular resolution and effective collecting area of the LAT are much greater than for EGRET. The model of the sky includes the cataloged EGRET sources as well as populations of fainter sources and the diffuse emission of the Milky Way [10]

the array; it registers the passage of charged particles and therefore is used in anticoincidence with the TKR and CAL in forming the trigger for gamma rays.

One advantage of a pair production tracking detector is that the field of view is enormous, ∼2.2 sr for the LAT. The LAT can observe many targets simultaneously and does not need to be pointed at a particular target. In fact, to increase the overall observing efficiency, the standard operating mode will have the LAT continuously scan the sky. This avoids the loss of observing time due to earth occultations and also limits the need to detect (and reject on board) the bright background of albedo gamma rays from interactions of cosmic rays in the upper atmosphere.

2.1. Design Issues for LAT Data Handling

A number of design compromises must be made for a gamma-ray detector to be operated in space. Most importantly, the collecting area is limited by the size of the rocket fairing, the mass is limited by the lift capacity, the power by the feasible solar cell and radiator capacities, and the data rate to the ground by allocations of telemetry bandwidth. The charged par-

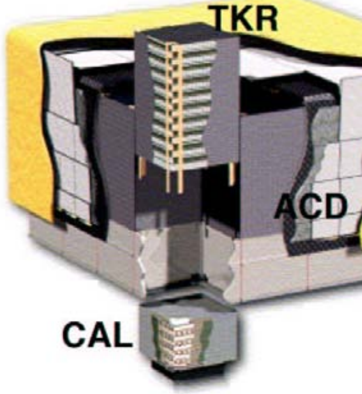


Figure 3: Cutaway view of the LAT. One of the sixteen towers is shown with its TKR module on top of the CAL module. The ACD is an array of plastic scintillator tiles that cover the towers. Surrounding the ACD is a thermal blanket and micrometeoroid shield. The overall dimensions are 1.8 × 1.8 × 0.75 m.

ticle background in orbit is intense; the orbit-averaged trigger rate will be approximately 3 kHz for the LAT. The actual rate of triggers from celestial sources will be ∼2 Hz. The telemetry bandwidth is sufficient to send event data (∼10 kbits per event) at an average rate of ∼30 Hz, so efficient filtering of the data in flight is essential.

The combination of a signal:background event rate ratio of $< 10^{-3}$, the need to reconstruct gamma-ray information from tracks and energy depositions in the LAT, the resulting limited angular resolution, the bright and structured diffuse gamma-ray emission from the Milky Way, and low fluxes of celestial point sources provide ample motivation for careful treatment of the data at every step of the analysis of LAT data.

3. LOW-LEVEL ANALYSIS

3.1. Nature of the Data

Readout of the LAT is triggered by the occurrence of hits in 3 successive X-Y TKR planes in tower, or a large energy deposition in the CAL. Simple, robust algorithms are used to filter the data. As described above, on board filtering of the data is required by the available average telemetry rate. The data are lists of the s that were hit, measurements of light output from the ends of the CsI(Tl) logs in the CAL, and a list of the tiles of the ACD that were hit.

3.2. Event Reconstruction

In ground processing, reconstruction of events (interactions of a cosmic ray or gamma ray in the LAT) starts with grouping the hits (silicon strips that registered a charged particle that trigger a readout) in the TKR into clusters, because adjacent strips can register the passage of the same particle. A pattern recognition algorithm is applied to associate the clusters into 'tracks', with preference for finding the longest, straightest tracks. The current algorithm is combinatorial (i.e., brute force).

The identified track or tracks are then fit via Kalman filtering (e.g., [11]). This defines the best estimate of the initial direction of the charged particle. This process is iterative with analysis of the energy depositions in the CAL, which is used for estimating the overall energy of the event. The energy information is used to evaluate the scattering angles expected in each tracker plane. Multiple scattering in the W conversion foils is quite non-Gaussian. In principle, this is a problem for the method. However, the uncertainty in the energy determination is great enough that it (rather than the non-Gaussian tails of multiple scattering) dominates the uncertainty of estimated scattering angles. The assumption implicit in the current analysis is that in these circumstances, the Kalman filtering method is applicable; more study of the validity probably would be prudent. An approach for track reconstruction that uses concepts from particle filtering is being investigated as a potential alternative [12]. It is more challenging computationally but should be able to explicitly take into account processes like multiple scattering.

Reconstructed tracks are analyzed to define the conversion point of the gamma ray and its initial direction. An example reconstruction is shown in Figure 4. At higher energies, the positron and electron tracks may not separate at the resolution of the tracker, so the vertex and the estimated initial direction come from analysis of a single track.

The energy deposition in the CAL in general must be corrected to account for partial containment of the showers. The CAL is only 8.5 radiation lengths deep (owing to the constraint on the mass that can be placed in orbit), and even at moderately large inclination angles, significant corrections are required. Two approaches are being evaluated, shower profiling and last-layer correlation. The development of the showers in the CAL can be reconstructed with coarse resolution, owing to the hodoscopic design discussed in Sec. 2. Intrinsic fluctuations in energy deposition as showers develop limit the resolution achievable by these techniques. At energies >100 GeV, typical energy containment is less than 40%; showers are still developing at the point that they leave the CAL and any correction scheme necessarily involves a large extrapolation.

If multiple tracks are found, the best (straightest, highest energy) tracks are checked for intersection. The estimated energies and initial directions of the two tracks are used to calculate the energy and direc-

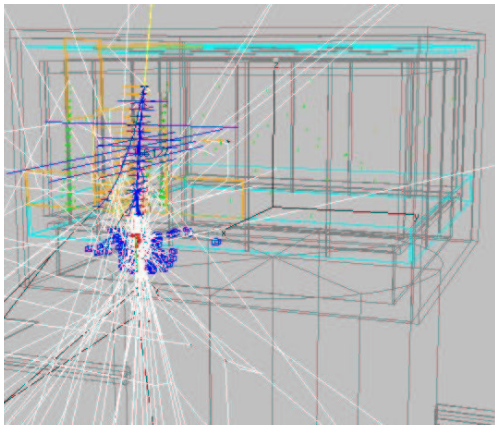


Figure 4: Simulated interaction of a 1 GeV gamma ray in the LAT. The LAT is indicated by a wire frame outline that also includes a schematic, cylindrical spacecraft. The reconstructed tracks are indicated in blue. The white lines are soft (X-ray) photons and the ACD tiles that are hit are outlined in orange. The CsI logs in the CAL with significant energy depositions are also indicated. Courtesy T. Usher.

tion of the incident gamma-ray.

3.3. Event Classification

Once an event is reconstructed, the final classification needs to be made. Fundamentally, the classification discriminates between charged particles and gamma rays, although sub-classifications also will be made. For example, heavy cosmic rays that do not undergo nuclear interactions in the CAL will be flagged for use in calibration of the CAL in flight. Also, gamma rays with especially well-measured energies or directions will be flagged.

Through extensive Monte Carlo simulation of the instrument, informed by beam tests of prototypes, useful diagnostics for discriminating cosmic rays from gamma rays have been identified. By far the most powerful is the intersection of the reconstructed event direction with a tile of the ACD that recorded a hit (passage of a charged particle). Other cuts are not as obvious, and not completely orthogonal. The production of a ‘clean’ gamma-ray data set is vitally important, owing to the orders-of-magnitude greater intensity of cosmic rays than celestial gamma rays. Currently, the classification of events is implemented using a classification tree trained with simulated data. Each node in the tree applies a single test (e.g., one based on projected distance to the nearest ACD tile that was hit). The result from traversing the tree is the probability that the event is a gamma ray; the probabilities are defined from the results of passing simulated events through the tree.

Decision trees are also used to identify the events that probably have well measured energies and directions. As mentioned above, multiple scattering in the tracker unavoidably causes long ‘tails’ in the point-spread function (PSF). The tails can confound the analysis of sky regions of high source density or intense diffuse emission, and to the extent that the events in the tails can be identified (and ignored) the cost to the effective collecting area may be worth the trade off for these circumstances.

One issue with this approach is the stability of the classification trees. For a classification tree analysis, a small change in input quantities can dramatically affect the path through the tree, and the resulting classification. More recently developed methods, such as boosted decision trees (e.g., Friedman, this volume), do not suffer this shortcoming and may be adopted for the classification of LAT events.

4. HIGH-LEVEL ANALYSIS

High-level analysis of LAT data is gamma-ray astronomy, the detection and characterization of celestial sources of gamma rays. Generally, for the reasons described above, the characterization will be via model fitting, where the parameters of a model quantify what we are trying to learn from the data. This approach has a long history in gamma-ray astronomy. Pollock et al. [13] introduced the maximum likelihood method for model fitting for analysis of data from COS-B, and the same approach was used for EGRET [14]. Technically, the method is extended maximum likelihood, because the number of gamma rays is itself random variable.

For likelihood analysis, the detector is represented by its response functions, high-level descriptions of how the point-spread function, energy resolution, and effective collecting area depend on energy, direction (relative to the instrument coordinate system), and other measurable quantities, like plane of conversion of the gamma ray, and the results of the event classification trees. This high-level description of the LAT, derived from Monte Carlo studies and accelerator beam tests, abstracts the instrument, the event data, reconstruction, and particle background rejection into what is needed for modelling the sky. The likelihood function is the probability of the data given the model. The response functions relate a model defined on the sky (sources of given spectra, positions, etc.) to the data space of measured energies, directions, etc., taking into account the pointing and live time history of the LAT for the period of interest.

For analysis of LAT data we may encounter practical limitations to the evaluation of the likelihood function. In maximizing the likelihood, changes in ln L of $\sim$1 are significant. Owing to the breadth of the PSFs at low energies, the region of the sky in a typical analysis will be of $\sim 15^\circ$ across and may contain hundreds of thousands of gamma rays (in an analysis of a one-year time frame). Numerical accuracy will have to be carefully maintained in the evaluation of the likelihood function.

The source models may also contain dozens of parameters (source positions, spectral indicies, scaling factors for diffuse emission). Only a fraction of these may be adjustable in any given analysis—e.g., coordinates of known sources may be held fixed—but even so maximization of the likelihood function will be a multidimensional optimization. In principle, this is manageable, but optimizations will be most reliable with good initial guesses for parameters. An implementation of the Expectation Maximization (EM) algorithm [15] is being explored for possible use in speeding up likelihood optimizations of models for LAT data. In this approach, gamma rays are provisionally assigned to specific sources in the model, based on the current values of the parameters of the model. Next, values of the parameters are optimized source by source, requiring likelihood evaluations for only a fraction of the gamma rays at a time. Source assignments are then updated, and the whole process is iterated. For models with a large number of sources, this approach potentially offers a tremendous advantage in computation time.

4.1. Nonparametric source detection

The fundamental limitation of likelihood analysis is that it does not answer a question that you are not asking. Also, as mentioned above, the method is computationally intensive and subject to limitation of numerical accuracy. A practical (fast and accurate) nonparametric method for detecting sources would have a great deal of appeal, of course. Even if a method provides just a useful starting point for detailed likelihood analysis to derive parameter estimations, it could be very useful. Several methods are under consideration for analysis of LAT data, including wavelet transformations (continuous and discrete), independent component analysis, general multiresolution image deconvolution, and a multidimensional extension of the Bayesian blocks algorithm (see Sec. 5.3). To the extent that they depend on knowledge of the instrument response functions for filtering, these methods may have difficulties due to the scanning observing mode of the LAT, which effectively mixes response functions for each source.

A further complication to the analysis of celestial gamma-ray sources in the LAT data is the brightness of the Earth's limb in gamma rays. These 'albedo' gamma rays produced in cosmic-ray interactions in the upper atmosphere have been characterized with data from SAS-2 [16]. The emission is quite intense relative to the gamma-ray sky, although fairly soft. The intensity exhibits a strong east-west variation, and also depends on solar activity. Even for the routine scanning sky coverage of the LAT, the horizon is never far from the field of view. The current plan is to exclude from the high-level analysis regions of the sky at large zenith angles ($>\sim 110^\circ$). The cuts necessarily will be made based on measured zenith angle, and owing to the relatively strong dependence of the PSF on angle, must be more conservative at lower energies. The cuts on zenith angle complicate both the data selection and the calculation of exposure.

4.2. Characterization of sources

By whatever means a gamma-ray source is detected, characterization of the source means determining the confidence region for its location on the sky, and measuring its spectrum and variability. If a searched-for source is not detected, a meaningful upper limit for its flux should also be determined. For EGRET, these were evaluated by applying the likelihood ratio test and appealing to Wilks' theorem for interpretation of the results [14]. The likelihood ratio test was used to compare source models, e.g., one with a given point source with its maximum likelihood position and one with that source shifted somewhat in position, and Wilks' theorem was used to relate the likelihood ratios to significance levels. The interpretations of significances in the EGRET data was backed up by Monte Carlo simulations.

Recently, Protassov et al. [17] have pointed out that often in astronomy the likelihood ratio test is misapplied to circumstances where one of the parameters (e.g., source flux) is on the border of the range on which it is defined (like 0 in the case of source flux). For questions like this, Protassov et al. propose evaluating Bayesian posterior predictive P-values. The procedure to be implemented for determination of confidence ranges and upper limits in routine analysis of LAT data is still being evaluated.

4.3. Identification of gamma-ray sources

By standards of astronomy at other wavelengths, the positions of gamma-ray sources are measured very poorly. The majority of the EGRET sources are unidentified, $\sim$170 out of 271 in the Third EGRET Catalog [4], largely for this reason (see Sec. 2 and Fig. 5). For EGRET, 95% confidence contours for source locations were typically 1-2° across. The number of potential counterparts is so large that no compelling case for any particular counterpart can be made on the basis of positional coincidence.

For the LAT we plan to adopt an objective procedure for identifying potential counterparts, taking advantage of all of the information that we can derive, such as variability (especially correlated variability). For the LAT, source location regions will be much

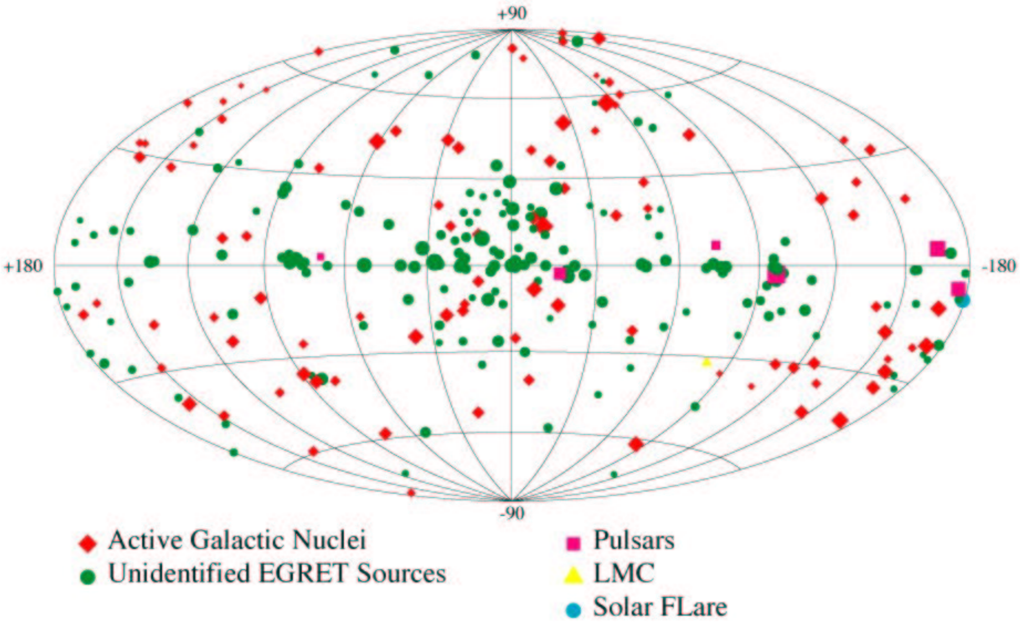


Figure 5: Sky locations of the sources in the Third EGRET Catalog [4]. Larger symbols indicate greater fluxes, scaled logarithmically. A concentration toward the Galactic plane is evident. The majority of the sources are unidentified.

smaller, on the order of several arcminutes for typical sources. This is still relatively large for counterpart searches, but will certainly make the problem easier. Mattox et al. [18, 19] made a Bayesian analysis of potential blazar counterparts to unidentified EGRET sources that used positional correlations of EGRET sources with radio continuum sources, as well as the flux and spectral index distributions of radio sources already known to be blazars. Sowards-Emmerd et al. [20] introduced a 'figure of merit' for assessing source counterparts that also includes X-ray spectral information. The figure of merit essentially includes weighting factors based on the X-ray and radio characteristics of known blazars. Establishing a new class of sources is more difficult, as statistical results for other members of the population are not available.

Variability is a common characteristic of high-energy gamma-ray sources. Blazars undergo episodic flares, during which fluxes can increase by factors of several on time scales of hours or less. Gamma-ray pulsars are periodic sources, typically with periods of hundreds of milliseconds; integrated over many periods their gamma-ray fluxes are quite steady. Indeed the brighter EGRET pulsars were used as calibration sources in flight. So, for unidentified sources, measures of variability can be used to distinguish between blazars and pulsars.

For candidate pulsars for which ephemeris data are available, generally from monitoring observations in the radio, epoch folding the gamma rays is a well-established way to search for gamma-ray pulsations. Well-defined statistical techniques have been applied to such pulsation searches [21]. The sensitivities of the tests are limited by the lack of a 'template' for pulsations. Some tests are most powerful for detecting sinusoidal variations, for example.

For suspected pulsars of unknown timing parameters, period searching is in principle possible, but hampered by many complications; see [22]. As is apparent from Fig. 6, the gamma-ray pulsations have no standard template. Before timing searches, the arrival times of the gamma rays need to be corrected for the arrival time variations due to the changes in the position of the spacecraft. If the direction of the source on the sky is not known very accurately (and it will likely not be), then uncertainties in the arrival time correction accumulate quickly. Phase drifts owing to the unknown spin-down rate (as large as 10^{-13} Hz s^{-1} for a young pulsar) can also become significant over the days or weeks required to accumulate enough gamma rays from a given source. So a period search is effectively multidimensional, including the coordinates of the prospective pulsar and its spin-down rate.

4.4. Source Identification

Positional coincidence alone is in general not adequate to establish the identification of a gamma-ray source with a counterpart detected at other wavelengths. The accuracy of position determinations with the LAT will typically be at the few arcminute level (depending on source spectrum and diffuse intensity). This is inadequate, owing to the high density of potential counterparts (e.g., the NRAO VLA Sky Survey

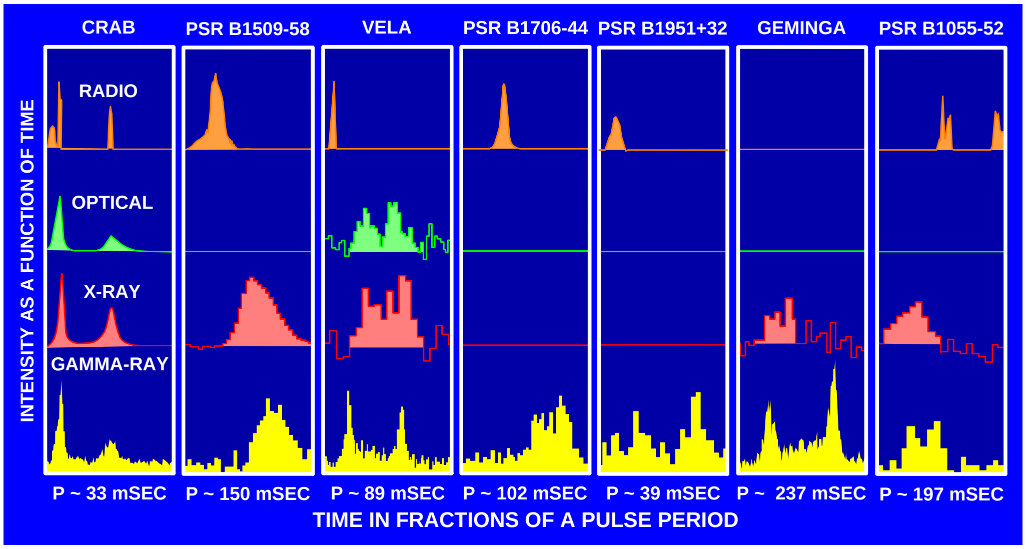


Figure 6: Composite of light curves at different wavelengths for many of the pulsars detected by EGRET. (Source: D. J. Thompson)

has >50 sources deg^{-2} [23]), without additional information that supports the identification.

Useful information can be applied based on the characteristics of either the γ-ray source or of the potential counterparts. For example, blazars have been established to be associated with flat-spectrum radio sources [14, 18].

Correlated variability is a powerful technique for identification, the prototypical example being γ-ray pulsars. For suspected counterparts with known ephemerides, statistical tests have been developed to evaluate whether the γ-ray source is pulsing with the same period. Gregory & Loredo [24] presented a Bayesian approach for periodicity searching that relies only on the assumption that the pulse profile can be assumed to be stepwise continuous; in their analysis, more complicated profiles, i.e., those with greater numbers of steps are naturally discouraged in favor of simpler profiles.

With any method of searching for periodicity, pulsations must be detected against the background of non-pulsed emission, e.g., from an associated nebula and diffuse interstellar emission, and the 'signal-to-noise' may be optimized by making PSF-dependent cuts the γ-rays included in the searches, taking advantage of the narrowing of the PSF at high energies to reduce the fraction of γ-rays of diffuse origin. An alternative approach that has been proposed instead of cutting events is to weight them according to the widths of the corresponding PSFs.

Gamma-ray bursts (GRBs) are well-known as brief, intense, and intensely variable gamma-ray sources, in recent years firmly established to be at cosmological distances and associated with galaxies with extensive massive star formation. Typical GRBs are brightest in the ~few MeV range, and from EGRET relatively little is known about their higher-energy behavior. This is primarily due to the large dead time per trigger (~0.2 s) of the EGRET spark chamber tracker; the dead time per event for the LAT likely will be significantly less than 0.1 ms.

That there are two populations of GRBs, short duration with hard spectra and long duration with relatively soft spectra, has long been established [25]. Typical durations are ~0.3 s for the short bursts vs. ~30 s for the long bursts, and the short bursts have spectral indexes harder by an increment of ~0.5. To date, only the long-soft population have been able to have counterpart identifications via rapid follow up observations, because the localization of the short-hard bursts with (typically) hard X-ray detectors is very poor. One goal for science with the LAT is to obtain excellent positions for the hard-spectrum bursts and GRB 'trigger' algorithms are being explored. By the nature of the LAT (see Sec. 2.1), some on board processing is required to make provisional reconstruction and classification of events on board. The trigger algorithms look for clusters of events in direction and time, using a moving time window. The on board algorithm will be tuned carefully via Monte Carlo simulations, and the LAT is designed to send GRB notifications to the ground using the demand access Tracking and Data Relay Satellite (TDRS) system, which promises latencies measured only in seconds.

The time profiles of GRBs typically consist of a train of pulses of different profiles (the details of which also depend on energy). Objective decomposition of GRB profiles into intervals with constant event rates (i.e., within which the variations of the event rate are not statistically significant), with proper attention to the background noise level, can be achieved using the Bayesian Block algorithm developed by Scargle [26]. The algorithm cannot sort out overlapping pulses, an unsolved problem in the general case, but still can provide useful characterizations of time profiles of GRBs. It does not require the events to be binned in time, so no minimum time scale for detecting variability is imposed by the method.

4.5. Population studies

Several related approaches have been used to demonstrate that the low-latitude EGRET gamma-ray sources are correlated with tracers of massive star formation, without needing to claim identifications for any particular unidentified source. Kaaret & Cottam [27] fitted Gaussian profiles to the latitude and longitude distributions of low-latitude unidentified EGRET sources and generated random sets of point sources consistent with these distributions. For each set, the number of sources lying within 1° of an OB association was counted. The probability of a chance association of as many as 16 sources out of 25, as observed with the actual EGRET sources, was evaluated from the distribution as 6.1 $\times 10^{-5}$.

Romero et al. [28] applied a somewhat different technique to evaluate the significance of positional correlations between the distributions of low-latitude EGRET gamma-ray sources and Wolf-Rayet stars, OB associations, and supernova remnants. The observed distributions of positional offsets was compared with the offsets obtained for the sources scrambled in longitude and latitude in such a way that their latitude distribution was exactly maintained. The conclusion that correlations were statistically significant was especially strong for SNR. Of course, Wolf-Rayet stars, OB associations and supernova remnants necessarily have fairly similar distributions, and may in fact interact to produce gamma-ray sources [29]

Grenier [30] investigated the distribution of unidentified EGRET sources by evaluating the log N-log S (flux distribution) for the sources and comparing the distribution of expected detections on the sky for various assumed intrinsic spatial distributions of the sources. This likelihood approach naturally compensates for sensitivity variations owing to different depths of exposure and intensities of diffuse emission across the sky. The results indicated a significant correlation with tracers of dense interstellar gas and star formation. Correlation with the population of radio pulsars was notably weaker, although most radio pulsars have lifetimes as gamma-ray sources much shorter than as radio sources and pulsars are known to have large proper motions.

A stacked source analysis can be used to study whether a population of putative gamma-ray sources can be detected collectively, even if individual sources are not bright enough for detection. Stacking the data for a population of sources means coadding counts and exposure centered on each source position. (Generally, unrelated nearby point sources must be excised.) This technique was recently applied by Reimer et al. [31] to investigate whether nearby, X-ray bright clusters of galaxies are also gamma-ray sources; see Fig. 7. The coadded exposure for 58 clusters resulted in an upper limit of ~6 $\times 10^{-9}$ cm^{-2} s^{-1} for the average flux above 100 MeV, approximately 8 times more sensitive than the upper limit for any of the galaxy clusters individually.

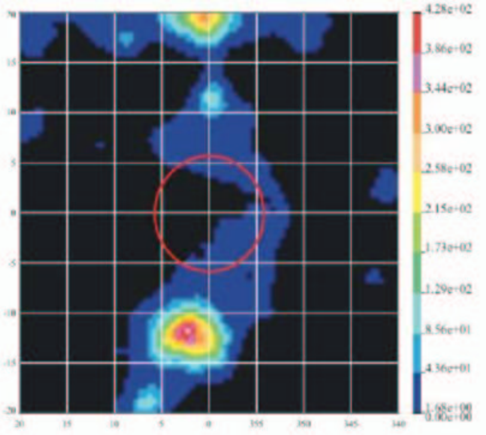


Figure 7: Source location (likelihood test statistic) map for the stacked EGRET counts and exposure for 58 X-ray bright galaxy clusters. No significant emission is seen at the composite source location at the center of the field [31].

5. CONCLUSIONS

The LAT instrument on GLAST will have revolutionary sensitivity and should revolutionize gamma-ray astronomy. Fulfilling the promise of the instrument will require careful statistical treatment of the data at all levels, from event reconstruction and classification, to source detection, source identification, and population studies. This relates to the detection method for gamma rays, their low intrinsic fluxes, and the scanning observing mode that will be routine for the LAT. Some classes of sources, in particular blazars and gamma-ray bursts, will require triggers for near-real time alerts. Searches for periodic emission from pulsars will also be needed.

References

[1] http://glast.gsfc.nasa.gov
[2] http://glast.stanford.edu/

[3] C. von Montigny, et al., "High-Energy Gamma-Ray Emission from Active Galaxies: EGRET Observations and Their Implications", ApJ, 440, 525, 1995

[4] R. C. Hartman, et al., "The Third EGRET Catalog of High-Energy Gamma-Ray Sources", ApJS, 123, 79, 1999

[5] P. Sreekumar, et al., "GeV emission from the nearby radio galaxy Centaurus A", APh, 11, 221, 1999

[6] G. Kanbach, et al., "Detection of a long-duration solar gamma-ray flare on June 11, 1991 with EGRET on COMPTON-GRO", A&AS, 97, 349, 1993

[7] W. L. Kraushaar, et al., "High-Energy Cosmic Gamma-Ray Observations from the OSO-3 Satellite", ApJ, 177, 341, 1972

[8] http://agile.mi.iasf.cnr.it/

[9] http://ams.pg.infn.it/ams-italy/ams.htm

[10] S. D. Hunter, et al., "EGRET Observations of the Diffuse Gamma-Ray Emission from the Galactic Plane", ApJ, 481, 205, 1997

[11] B. B. Jones, "A Search for Gamma-Ray Bursts and Pulsars, and the Application of Kalman Filters to Gamma-Ray Reconstruction", Ph. D. thesis, Stanford Univ. 1998, http://arxiv.org/abs/astro-ph/0202088

[12] R. D. Morris and J. Cohen-Tanugi, "An Analysis Methodology for the Gamma-ray Large Area Space Telescope", Proc. MaxEnt 23, "Bayesian and Maximum-Entropy Methods", Jackson Hole, WY, in press, 2004

[13] A. M. T. Pollock, et al., "Search for gamma-radiation from extragalactic objects using a likelihood method", A&A, 94, 116, 1981

[14] J. R. Mattox, et al., "The Likelihood Analysis of EGRET Data", ApJ, 461, 396, 1996

[15] A. P. Dempster, et al., "Maximum likleihood from imcomplete data via the EM algorithm", J. Royal Stat. Soc., Ser B., 39(1), 1, 1977

[16] D. J. Thompson, G. A. Simpson, and M. E. Ozel, "SAS 2 observations of the earth albedo gamma radiation above 35 MeV", JGR,, 86, 1265, 1981

[17] R. Protassov, et al., "Statistics, Handle with Care: Detecting Multiple Model Components with the Likelihood Ratio Test", ApJ, 571, 545, 2002

[18] J. R. Mattox, et al., "The Identification of EGRET Sources with Flat-Spectrum Radio Sources", ApJ, 481, 95, 1997

[19] J. R. Mattox, et al., "A Quantitative Evaluation of Potential Radio Identifications for 3EG EGRET Sources", ApJS, 135, 155, 2001

[20] D. Sowards-Emmerd, R. W. Romani, and P. F. Michelson, "The Gamma-Ray Blazar Content of the Northern Sky", ApJ, 590, 109, 2003

[21] O. C. de Jager, "On periodicity tests and flux limit calculations for gamma-ray pulsars", ApJ, 436, 239, 1994

[22] A. M. Chandler, et al., "A Search for Radio-Quiet Gamma-Ray Pulsars", ApJ, 556, 59, 2001

[23] J. J. Condon, et al., "The NRAO VLA Sky Survey", AJ, 115, 1693, 1998

[24] P. C. Gregory and T. J. Loredo, "A new method for the detection of a periodic signal of unknown shape and period", ApJ, 398, 146, 1992

[25] C. Kouveliotou, et al., "Identification of Two Classes of Gamma-Ray Bursts", ApJ, 413, L101, 1993

[26] J. D. Scargle, "Studies in Astronomical Time Series Analysis. V. Bayesian Blocks, a New Method to Analyze Structure in Photon Counting Data", ApJ, 504, 405, 1998

[27] P. Kaaret and J. Cottam, "Do the Unidentified EGRET Sources Lie in Star-forming Regions?", ApJ, 462, L35, 1996

[28] G. E. Romero, P. Benaglia, and D. F. Torres, "Unidentified 3EG gamma-ray sources at low galactic latitudes", A&A, 348, 868, 1999

[29] T. Montmerle, "On gamma-ray sources, supernova remnants, OB associations, and the origin of cosmic rays", ApJ, 231, 95, 1979

[30] I. A. Grenier, "Spatial distribution of the unidentified EGRET sources off the galactic plane", AdSpR, 15, 573, 1995

[31] O. Reimer, "EGRET Upper Limits on the High-Energy Gamma-Ray Emission of Galaxy Clusters", ApJ, 588, 155, 2003

Statistical Issues in Particle Physics – A View from BaBar

F. C. Porter
(For the BaBar collaboration)
Lauritsen Laboratory of Physics, Caltech, Pasadena, CA 91125, USA

The statistical methods used in deriving physics results in the BaBar collaboration are reviewed, with especial emphasis on areas where practice is not uniform in particle physics.

1. INTRODUCTION

The purpose of the BaBar experiment at the PEP-II accelerator at SLAC is to study e^+e^- collisions in the 10 GeV center-of-mass region, namely the region around $B\bar{B}$ threshold. In particular the program is to investigate extensively CP violation and rare decays of B mesons, as well as topics in charm and tau physics.

Here, BaBar's approach to statistical issues is summarized. Emphasis is given to areas which are often controversial.

2. BABAR ANALYSIS ORGANIZATION

BaBar is a collaboration of approximately 600 physicists, from $\sim$ 80 institutions in a dozen countries. Thus, managing the production of physics results, from initial analysis to final publication, while maintaining collaboration involvement is a daunting task. An organizational structure has been established to facilitate this process, as illustrated in Fig. 1.

The "Statistics Working Group" was appointed by the Publications Board in order to provide guidelines and advice on statistical matters [1]. This group is advisory; I'll note how well the guidelines are actually adopted in some cases.

3. PHILOSOPHY

The approach to choosing a statistical procedure is to start by considering the goal. We adopt the view that there are two broad domains in terms of goal:

- The first goal is that of summarizing the relevant information in a measurement. This is "descriptive" statistics. It is considered obligatory to report such a description of the result of the experiment. Inherent in this is the view that it is actually useful to do so, a notion that is not uniformly accepted. The use of frequency statistics is recommended for this purpose. The choice within the domain of possible frequency statistics is driven by an emphasis on clarity and the facility to compare and combine with other measurements.

- The second goal is that of interpreting the relevant information in the context of making a

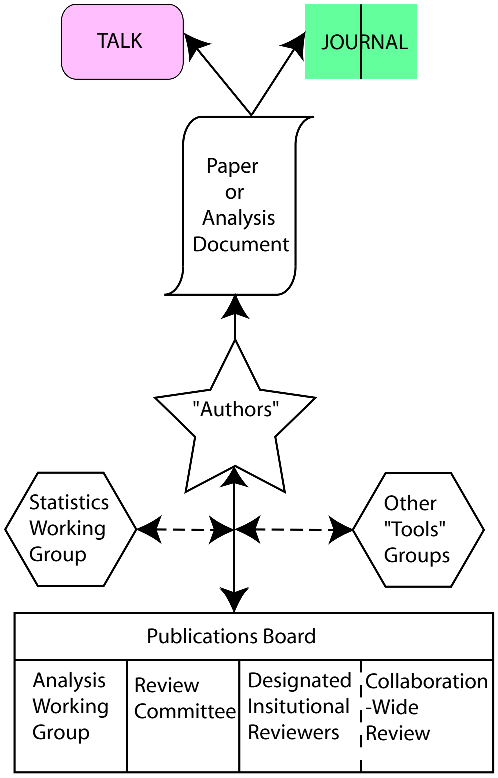


Figure 1: BaBar analysis organization. A detailed analysis for some physics result is typically performed by a subset of the collaboration, labelled "authors" here. There are several layers of review that occur as an analysis moves towards publication: an Analysis Working Group interacts with the authors from the earliest stages; once a document is produced, a Review Committee of typically three people is assigned by the Publications Board to critically examine the analysis; upon approval from the Review Committee, the paper is circulated for collaboration-wide review, including several institutions specially designated to look closely at it. Oversight of the process and final review is carried out by the Publications Board.

statement about "physics". This is regarded as optional, since once the relevant information is available people are in principle able to do this step for themselves. Because a statement about physical reality may depend on other information, and on theoretical input, Bayesian statistics are recommended.

It may be remarked that there may be other goals, such as making a decision concerning how to spend money for the next experiment. This would involve, beyond the above interpretive aspects, a consideration of the risks and benefits. We take the point of view that this is outside the scope of the analysis and reporting of results, and hence do not discuss it further.

4. STATISTICAL PRACTICE IN BABAR

We turn now to a review of the specific statistical practices recommended or adopted in BaBar analyses. Not included are the methods and tools used for optimizing analyses, and pattern recognition, data reduction, and simulation procedures. These matters are crucial, but here we emphasize instead areas which are traditionally more controversial. It should be mentioned that the typical products of a BaBar physics analysis are:

1. "Best" estimates for physical parameters.

2. Interval estimates for physical parameters.

3. Significance levels of observations (e.g., of a possible discovery).

4. Goodness-of-fit of models to the data.

4.1. Blind Analysis

Many BaBar results are obtained in "blind analyses". The purpose of a blind analysis is to avoid the introduction of bias, which could occur if the analyst is looking at the results as the analysis is designed. There is more than one approach to "blindness", see the talk by Aaron Roodman [2] for a summary of BaBar practice. We'll give one example here.

For example, consider the measurement of the rare B decay $B^{\pm} \to K^{\pm}e^{+}e^{-}$ [3], of interest because of its sensitivity to possible physics beyond the standard model. The basic idea of the analysis is to look for a signal which peaks in the distribution of two kinematic variables, known as "ΔE" and "m_{ES}" (Fig. 2). A fit is performed to this two-dimensional distribution in order to extract the strength of any signal present. However, before performing the fit, an event selection is made in order to suppress backgrounds. In order to avoid biasing the result by looking at the data while tuning the selection, a blind analysis is performed.

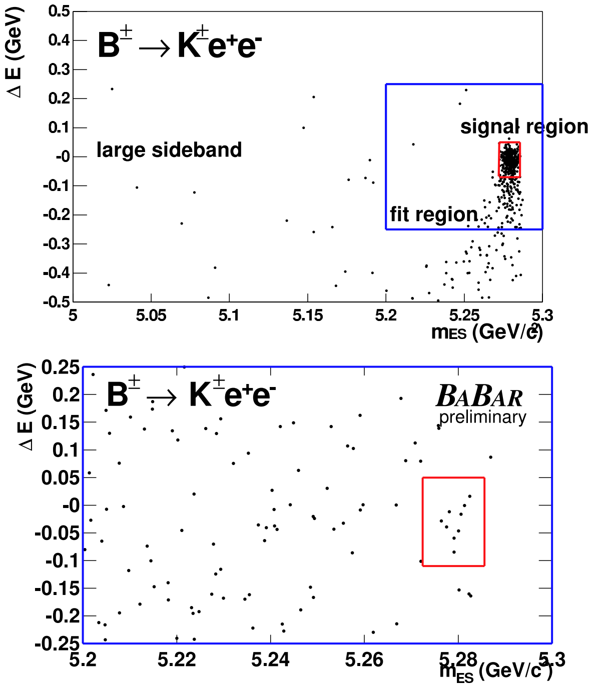


Figure 2: Example of a blind analysis in BaBar. The upper plot shows a Monte Carlo simulation of the signal $B^{\pm} \to K^{\pm}e^{+}e^{-}$ process. The outside boundaries delimit the "large sideband region"; the intermediate box is the "fit region", and the inner box is the region in which the signal is concentrated (referred to as the "signal region", but in fact playing no special role in the analysis). The lower plot shows the BaBar data after unblinding. Here, the outside boundaries demarcate the fit region, and the smaller box is the "signal region".

The $\Delta E - m_{\text{ES}}$ plane is divided into two regions: a region where the fit will be performed, which includes the region where a signal might appear; and a larger ("large sideband") region which excludes the fit region. During the tuning of the analysis, the data may not be looked at in the fit region, only in the large sideband region. Monte Carlo and control sample data (including a type of data resembling signal) are used to tune the analysis. Once the selection criteria have been established, the fit region of the data is revealed, and the fit performed to extract the result.

As BaBar is continuing to accumulate data, an issue arises when it is desired to update a blind analysis to include new data. In principle, one could simply add the new data, without changing the analysis. However, this may be impractical, or undesirable. For example, the entire dataset may be re-reconstructed with improved constants or pattern recognition code. Or, there may have been improvements in tools such as particle identification. One would like to incorporate the benefit from such improvements. Additionally, it might be desirable to work harder to optimize the analysis, or to optimize on different criteria, such as precision instead of sensitivity. BaBar often takes a practical compromise approach to incoporate new data, and such improvements. We have the notion of "re-blinding" the data, and re-optimizing. It is considered safe in this re-optimization to use variables which have not been inspected too carefully in the blind region in the first dataset. Nonetheless, once we have done this, we do not refer to the new result as having been done with a blind analysis.

BaBar is perhaps the first large HEP collaboration to have embraced the blind methodology so enthusiastically. However, not every BaBar analysis is blind. In particular, analyses which may be called exploratory are generally not blinded. A recent example from BaBar is the discovery of the $D^*_{sJ}(2317)^\pm$ [4], which was not the result of a blind analysis. There are many examples of people being led astray by such non-blind exploratory analyses, so extreme caution is warranted. The exploratory nature of such analyses makes it difficult to apply rigorous methodologies with well-defined statistical properties. It may not be impossible to do better though [5].

4.2. Confidence Intervals

The recommendation in BaBar is to use frequency statistics for summarizing information (Sect. 3). The goal is to describe what is observed, stressing simplicity and coherence of interpretation, as well as facility in combining with other results. With these criteria, we think it can be counter-productive to impose "physical" constraints. There is no reason to obscure the observation of an "unlikely" result. Imposing constraints may also complicate combination of results. Generally, the recommendation is to quote two-sided 68% confidence intervals as the primary result. Where there may be doubt, a check for frequency validity (coverage) should be performed.

4.2.1. Example in Two Dimensions

As an example of the construction of a confidence region in a BaBar analysis, consider the measurement of D mixing and doubly Cabibbo suppressed D decays [6]. In this analysis, two parameters of interest are to be determined, which may be expressed as x' and y' according to the relations:

$$x' \equiv \frac{\Delta m}{\Gamma}\cos\delta + \frac{\Delta\Gamma}{2\Gamma}\sin\delta, \tag{1}$$

$$y' \equiv \frac{\Delta\Gamma}{2\Gamma}\cos\delta - \frac{\Delta m}{\Gamma}\sin\delta, \tag{2}$$

where m and Γ are the D mass and width, Δm and $\Delta\Gamma$ are the (small) differences in masses and widths between the two D mass eigenstates, and δ is an unknown strong phase (between Cabibbo-favored and doubly Cabibbo suppressed amplitudes). The measurement is only sensitive to x'^2 and y, and it is possible that the maximum of the likelihood will occur at $x'^2 < 0$ ("unphysical" region). At the current level of sensitivity, we should find a result consistent with $x'^2 = y' = 0$, if the standard model is correct.

The construction of a confidence region in the two-dimensional (x'^2, y') plane, corresponding to 95% confidence level with the frequency interpretation, is performed as follows (Fig. 3):

1. Pick a point $(x_0'^2, y_0')$ in the plane.
2. Form the "data" likelihood ratio comparing the observed maximum likelihood with the likelihood at $(x_0'^2, y_0')$:
$$\lambda_{\rm Data} = \frac{\mathcal{L}_{\max}(\rm Data)}{\mathcal{L}_{(x_0'^2,y_0')}(\rm Data)}. \tag{3}$$
3. Simulate many experiments with $(x_0'^2, y_0')$ taken as the true values of the parameters.
4. For each Monte Carlo simulation form the "MC" likelihood ratio:
$$\lambda_{\rm MC} = \frac{\mathcal{L}_{\max}(\rm MC)}{\mathcal{L}_{(x_0'^2,y_0')}(\rm MC)}. \tag{4}$$
5. From the ensemble of simulations, determine the probability $P(\lambda_{\rm MC} > \lambda_{\rm Data})$. If this probability is greater than 0.95, then the point $(x_0'^2, y_0')$ is inside the contour; if less than 0.95, then the point is outside the contour.
6. This procedure is repeated for many choices of $(x_0'^2, y_0')$ in order to map out the contour.

Fig. 3 shows the result of this algorithm. The choice was made to stop computng the contour at the border of the "physical" region. The computation could in principle have been carried into the "unphysical" region (up to technical difficulties of the sort we shall discuss anon). It of course makes no difference to the frequency interpretation whether it is extended into the "unphysical" region or not.

4.2.2. Low Statistics Issues

Issues arise in applying the recommendation of always quoting a two-sided interval for a parameter when the sampling is not from an approximate normal distribution. Most often this involves the low-statistics regime of a counting process.

The first issue is a technical one: it can happen that a search in parameter space wants to go into a region where the probability distribution is undefined. This is distinct from going into an "unphysical" region as in the example above: we'll call it crossing a "math boundary". As a simple example, consider the case of

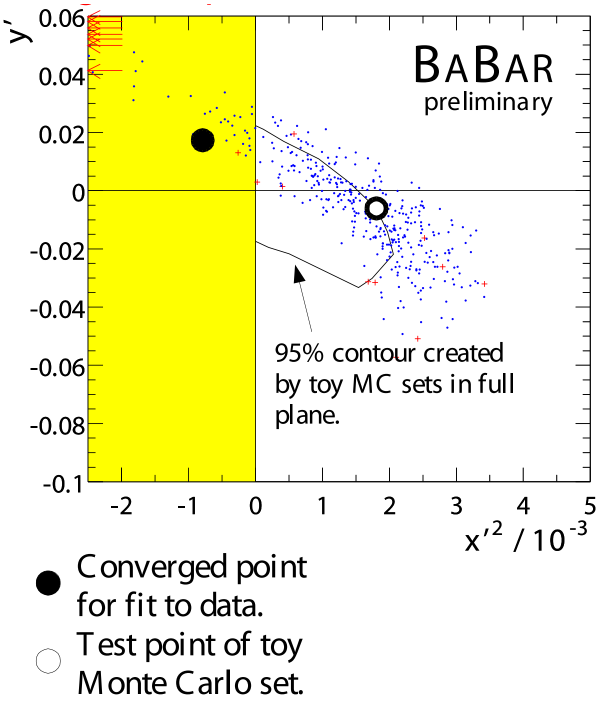


Figure 3: Finding a confidence contour in two dimensions [7]. The large filled dot shows the location of the maximum likelihood for the BaBar data. The open dot shows the value of $(x_0'^2, y_0')$ chosen for a simulation. The small dots show simulated experiments for which $\lambda_{\rm MC} > \lambda_{\rm Data}$. The pluses, as well as the arrows pointing offscale, show simulated experiments for which $\lambda_{\rm MC} < \lambda_{\rm Data}$. The 95% contour resulting from the algorithm described in the text is shown. The shaded region is the "unphysical" region. Note that the evaluation of the maximum likelihood is not restricted to the "physical" region.

a normal "signal" on a flat "background", with PDF (Fig. 4):

$$p(x;\theta) = \frac{\theta}{2} + \frac{1-\theta}{A\sqrt{2\pi}\sigma}e^{-\frac{x^2}{2\sigma^2}},\ x \in (-1,1). \qquad (5)$$

The parameter of interest is the strength of the signal, here expressed as $1-\theta$, the probability of sampling a signal event. An experiment samples N events from this distribution, with likelihood function:

$$\mathcal{L}(\theta;\{x_i, i=1\ldots,N\}) = \prod_{i=1}^{N} p(x_i;\theta). \qquad (6)$$

It is quite possible that the likelihood will be maximal for a value of θ for which the PDF is not defined. The function $p(x;\theta)$ may become negative in some region of x. If there are no events in this region, the likelihood is still "well-behaved". However, the resulting fit, as a description of the data, will typically look poor even where the PDF is positive. This is considered unacceptable.

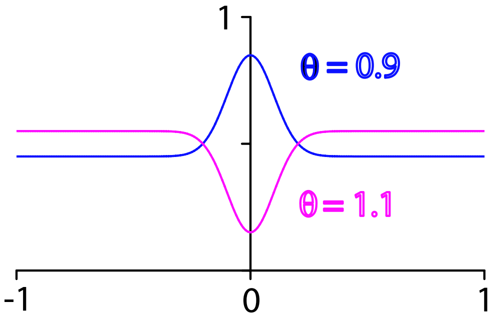


Figure 4: Graph of the example sampling PDF for two values of parameter θ: $\theta = 0.9$, and "unphysical" (negative signal) value $\theta = 1.1$. Note that both values are mathematically permissible.

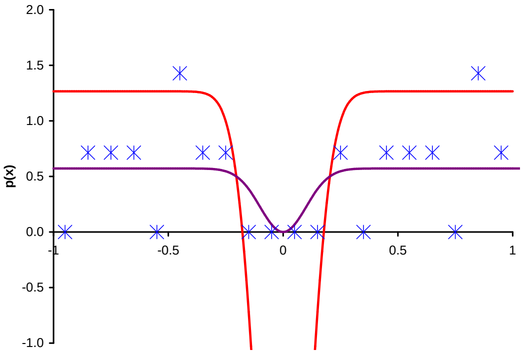


Figure 5: Example of a possible dataset generated according to the flat background plus normal signal PDF. The data are displayed in histogram form by the points. The curve that goes negative (and is cut off at the plot boundary) is the result of the (unbinned) maximum likelihood fit. The other curve is the result of the same fit, except with the constraint that it cannot become negative.

An illustration of a possible sampled dataset from this distribution is shown in Fig. 5, displayed as a histogram. An (unbinned) maximum likelihood fit to this data gives an estimate for θ in a region outside the math boundary. The graph of the "PDF" curve for this estimate does not give a good representation of the data. On the other hand, if the fit is constrained to the math region, the graph of the PDF curve looks like a reasonable representation of the data.

Thus, we suggest as a practical resolution to this problem to constrain the fit to remain within bounds such that the PDF is everywhere legitimate (n.b., parameters may still be "unphysical"). Experience is that this gives fits which "look" like the data, as in the present example, Fig. 5. This same practical recommendation applies in interval evaluation (but coverage should be checked, as always).

Another issue that arises frequently in low statistics (Poisson) sampling may be expressed in the form of the following example: A "cut and count" analysis for a branching fraction B finds n events. The mean expected background contribution is estimated as $\hat{b} \pm \sigma_b$ events. The efficiency and parent sample are estimated to give a scale factor (relating observed signal events to B) of $\hat{f} \pm \sigma_f$. The problem is to determine a confidence interval (at 68% confidence, say), in the frequency sense, for B.

We'll assume that n is sampled from a Poisson distribution with mean $\mu = \langle n \rangle = fB + b$, that $\hat{b}$ is sampled from a normal distribution, $N(b, \sigma_b)$, and that $\hat{f}$ is sampled from a normal distribution, $N(f, \sigma_f)$. Thus the likelihood function is:

$$\mathcal{L}(n, \hat{b}, \hat{f}; B, b, f) = \frac{\mu^n e^{-\mu}}{n!} \frac{1}{2\pi\sigma_b\sigma_f} e^{-\frac{1}{2}\left(\frac{\hat{b}-b}{\sigma_b}\right)^2 - \frac{1}{2}\left(\frac{\hat{f}-f}{\sigma_f}\right)^2}. \tag{7}$$

It should be noted that this example is realistic, arising in practice (to a good approximation). A variant is to assume a normal distribution in $\widehat{(1/f)}$

Several methods have been proposed, and used, for dealing with this problem (see Ref. [8] for further discussion of these):

1. Just give n, $\hat{b} \pm \sigma_b$, $\hat{f} \pm \sigma_f$. This provides a complete summary of the relevant information, and should be done anyway. But it isn't a confidence interval for B.

2. Integrate out the nuisance parameters according to

$$\mathcal{L}(n, \hat{b}, \hat{f}; B) = \tag{8}$$
$$\int df \int db \frac{\mu^n e^{-\mu}}{n!} \frac{1}{2\pi\sigma_b\sigma_f} e^{-\frac{1}{2}\left(\frac{\hat{b}-b}{\sigma_b}\right)^2 - \frac{1}{2}\left(\frac{\hat{f}-f}{\sigma_f}\right)^2}.$$

   This is easy, and often done. It may be interpreted as a partially Bayesian approach, where a uniform prior has been assumed for f and b. The frequency properties could be investigated, but usually aren't.

3. A very common approach when quoting upper limits is to do the appropriate Possion statistical analysis for n, but with the scale and background parameters fixed at the estimated values shifted by one standard deviation (in the direction to make the limit higher than with the central values). This has the benefit of being very easy to do, but it is clearly ad hoc, and the coverage is usually not investigated.

Here, I would like to comment on the possibility of evaluating these confidence intervals in another way.

The method I consider is actually a very common method that seems to have been rather neglected as an approach to the present problem. The algorithm is as follows: First, find the global maximum of the likelihood function with respect to B, f, b. Then search in the B parameter for the point where $-\ln\mathcal{L}$ increases from the minimum by a specified amount (perhaps by $\Delta = 1/2$ for a 68% confidence interval), making sure that the likelihood is re-maximized with respect to f and b during this search. The resulting points B_ℓ, B_u then give an estimated interval for parameter B which we would like to be a confidence interval.

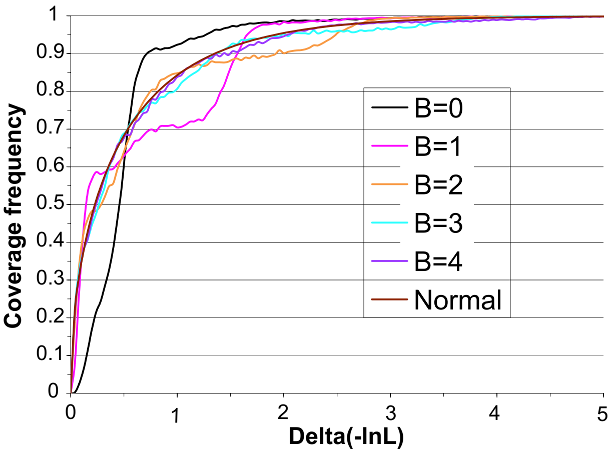


Figure 6: Coverage frequency as a function of Δ for $f = 1$, $\sigma_f = 0.1$, $b = 0.5$, $\sigma_b = 0.1$. There are several curves corresponding to different numbers of expected signal events, B. The smoothest curve is the coverage in the high statistics (normal) limit.

The question, of course, is: Does it work? To answer this, we need to investigate the frequency property of the algorithm. For large statistics (i.e., the normal limit) we know it works — for $\Delta = 1/2$ this method produces a 68% confidence interval for B. We expect that it will fail in the extreme small statistics limit, and the question becomes a quantitative one of how far it can be pushed into the low statistics regime. We answer this with Figs. 6–10.

Figure 6 shows the dependence of the coverage of this algorithm on the value of Δ, for several values of B and an expected background of 1/2 event. The branching fraction scale is adjusted so that B may be interpreted as the mean number of signal events. It may be seen that $\Delta = 1/2$ gives coverage reasonably close to 68% for $B \geq 2$. Figure 7 shows the coverage for $B = 0$, for several backgrounds. Even at zero branching fraction, the $\Delta = 1/2$ coverage is fairly close to 68% for expected backgrounds $b \geq 2$. Note that extending this to intervals with higher confidence may result in different conclusions.

It may be remarked that uncertainties in the background and/or scale factor help to obtain the desired coverage (Figs. 8 and 9). This is because they smooth out the effect of the discreteness of the Poisson sampling space.

One issue is when the coverage is deemed to be

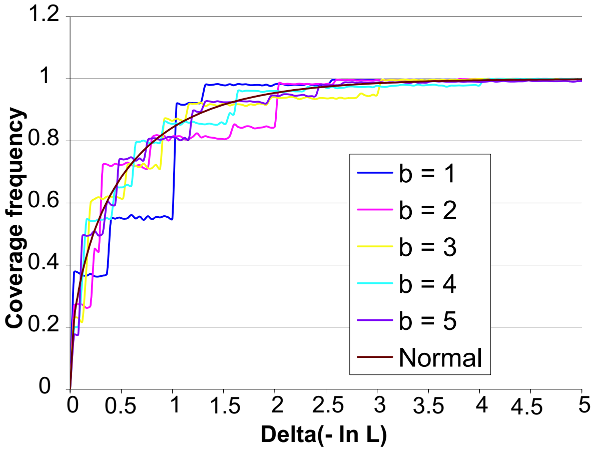


Figure 7: Coverage frequency as a function of Δ for $B = 0$, $f = 1$, $\sigma_f = 0$, $\sigma_b = 0.1$. There are several curves corresponding to different numbers of expected background events, b. The smoothest curve is the coverage in the high statistics (normal) limit.

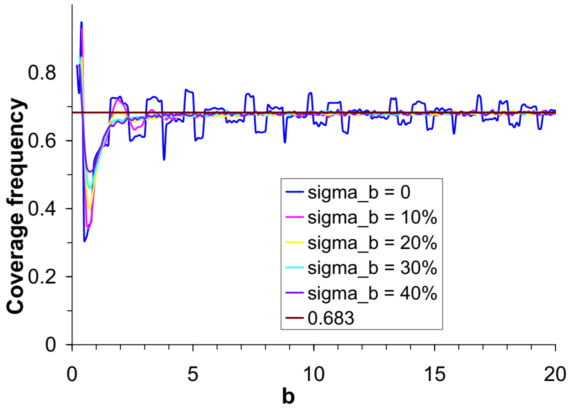


Figure 8: Coverage frequency as a function of mean background b for $B = 0$, $f = 1$, $\sigma_f = 0$, $\Delta = 1/2$. There are several curves corresponding to different values of σ_b, becoming smoother as σ_b increases. The horizontal line is at 68%.

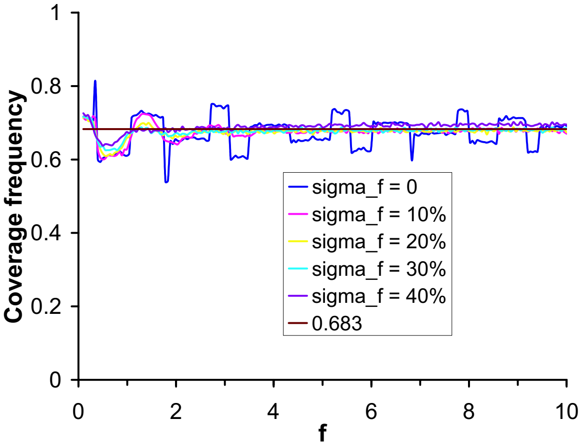


Figure 9: Dependence of coverage on scale factor f and σ_f for $B = 1$, $b = 2$, $\sigma_b = 0$, $\Delta = 1/2$. There are several curves corresponding to different values of σ_f, becoming smoother as σ_f increases. The horizontal line is at 68%.

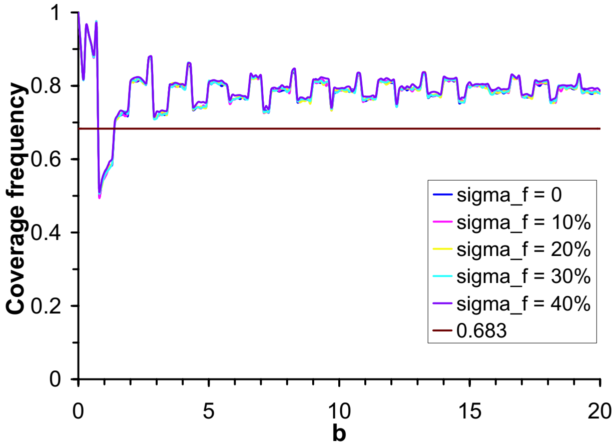


Figure 10: Coverage as a function of expected background for $\Delta = 0.8$, $B = 0$, $f = 1$, $\sigma_b = 0$. There are several curves corresponding to different values of σ_b, becoming smoother as σ_b increases. The horizontal line is at 68%.

"good enough". It might be suggested that if the coverage is known to be within some amount, say 5% of 68%, that this is good enough for anything we are going to use those numbers for. However, one could also decide to take a "conservative" approach, and insist that the coverage be at least at the quoted level. One way to accomplish this is to shift the value of Δ. Fig. 10 shows the coverage as a function of expected background (in the worst-case of zero signal branching fraction and $\sigma_b = 0$) for a value of $\Delta = 0.8$. We see that at least 68% coverage is guaranteed as long as the mean background is greater than 1.4.

We'll conclude this discussion with a few summary remarks: First, it is a good idea to always quote $n, \hat{b} \pm \sigma_b$, and $\hat{f} \pm \sigma_f$. Second, any approach used should be justified with a computation of the coverage. The likelihood analysis studied here works pretty well even down to rather low statistics for 68% confidence intervals. It should be kept in mind however that "good enough" for 68% intervals does not imply good enough for other purposes, such as tests of significance. Finally, if $\sigma_b \approx b$ or $\sigma_f \approx f$ this is outside the regime studied here; the normal assumption is likely invalid in this case.

4.2.3. Interpretation Intervals

In the interpretation stage, Bayesian intervals may be given, as deemed useful to the consumer. In BaBar practice, this is typically done when someone wants to give an upper limit, and is usually implemented with the assumption of a uniform prior in the parameter of interest. BaBar recognizes the issues surrounding the choice of prior. The recommendation is to consider it carefully, and to make checks on how sensitive the result is to the choice. Even this recommendation is not routinely adopted however.

4.3. Significance

The "significance" of an observation (e.g., of the presence of a signal for some process) is defined as the probability of the observed deviation (or larger) from the null (no signal) model, under the null hypothesis. The recommended procedure in BaBar is to compute this probability according to the frequentist methodology. It may be noted that knowing the 68% confidence interval does not always provide much insight into the significance. The tails of the null sampling distribution may be non-normal. A separate analysis is generally required, in which the tails are appropriately modelled.

No recommendation is tendered for when to label a result as "significant". We struggled with possible algorithms, but eventually gave up, because such a label implies an interpretation. No uniform prescription seems to make sense; judgement is involved. For example, deciding that the observation of a bizarre new particle is significant may involve a different standard than the claim that an expected decay mode of an established particle is significant. It isn't really our primary role as experimenters; it is up to the reader ultimately to decide what they wish to believe. This is perhaps the least-accepted of the Statistics Working Group's points in BaBar: people insist on making qualitative statements, e.g., "observation of", "evidence for", "discovery of", "not significant", "consistent with". A code exists in which "observation of" becomes quantified as $> 4\sigma$ significance, and "evidence for" means $> 3\sigma$.

This preoccupation with qualitative interpretive terminology is pervasive beyond BaBar. For example, the following excerpt appeared in Physics Today [9], (italics mine, references deleted):

> "In March, back-to-back papers in Physical Review Letters reported the measurement of CP symmetry violation in the decay of neutral B mesons by groups in Japan and California. Now the word "*measurement*" has been replaced by "*observation*" in the titles of two new back-to-back reports by these same groups in the 27 August Physical Review Letters. That is to say, with a lot more data and improved event reconstruction, the BaBar collaboration at SLAC and the Belle collaboration at KEK in Japan have at last produced the *first compelling evidence* of CP violation in any system other than the neutral K mesons."

For another example, some people think a measurement should not be called a "measurement" unless the result is significantly different from zero. An editor at a prominent journal has suggested that "*bounds on*" might be more appropriate than "*measurement*" in reference to a CP asymmetry angle which was observed as consistent with zero. This can lead to amusing ironies: Finding $\sin 2\beta = 0.00 \pm 0.01$ would be an exciting contradiction with the standard model. But it isn't a "measurement"?

A further issue that arises is that many people mix the question of significance with the choice of interval (i.e., one-sided vs two-sided). This has a drawback, because basing how one quotes the interval based on the result of the measurement can introduce a bias. The algorithm of Feldman and Cousins [10] is designed to address this. However, this methodology is not adopted in BaBar because of the constraint on the physical region, as discussed earlier. Instead, our recommendation is to always give a two-sided interval (if otherwise appropriate), independent of the significance. The significance is quoted separately. Quoting a one-sided interval may optionally also be done, and is usually regarded as part of the interpretation (hence a Bayesian approach is suggested). This recommendation is typically followed in BaBar, but there have been exceptions.

Another issue that arises in the quoting of significance has to do with the tradition of quoting significance as $n\sigma$. Unfortunately, this is used to mean different things: Sometimes it actually means n standard deviations. But sometimes it means the probability content of an $n\sigma$ fluctuation for a normal distribution. We recommend to quote directly the probability if the sampling distirbution is not normal. However, this has met with very limited implementation.

4.4. Systematic Uncertainties

BaBar makes many checks in a typical analysis. For the purpose of defining systematic uncertainties, we divide these into two broad categories:

1. "Blind checks": This is a test for mistakes. No correction to the data is anticipated. If the test passes, then there is no contribution to the systematic error. An example of such a check is dividing the data into two chronological subsets and comparing the results.

2. "Educated checks": This is a measurement of biases or corrections, and may affect the quoted result. It involves a contribution to the systematic error. An example is the model dependence of the efficiency calculation.

It is recommended that the systematic uncertainty be quoted separately from the statistical uncertainty. The sources of systematic uncertainty should be described, and may contain statistical components, for

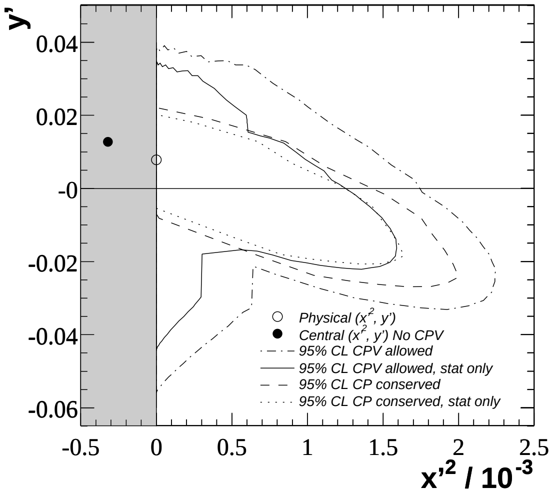


Figure 11: Incorporating systematic uncertainties in the confidence contour for D mixing. The filled dot is the location of the best fit point, $(\hat{x}'^2, \hat{y}')$, the open circle the best fit point in the physical region. The solid contour (dotted if restricting to CP conserving models) shows the 95% confidence contour according to statistical errors only. The dot-dash contour (dash for CP conserving models) shows how this contour becomes scaled on incorporating systematic uncertainties.

example due to limited Monte Carlo statistics in the efficiency evaluation.

We return to our earlier example (Sec. 4.2.1) of D mixing for an example of the treatment of systematic uncertainties. The goal here is to produce a two-dimensional confidence contour in the parameter space which incorporates the systematic uncertainites. In this case, the statistical uncertainties are large, and we are willing to accept an approximation in order to keep the procedure simple. Thus, it is decided to use a method which takes the statistics-only contour and scales it uniformly along rays from the best fit value. The scaling factor is $\sqrt{1+\sum m_i^2}$, where m_i is an estimate of systematic uncertainty i in units of the statistical uncertainty. This estimate is obtained by determining the effect of the systematic uncertainty on $\hat{x}'^2, \hat{y}'$ (the position of the best fit). Figure 11 shows the result of this procedure. This method is conservative (or lazy) in the sense that scaling for a given systematic in one (worst case) direction is applied uniformly in all directions. On the other hand, by evaluating the error at the best fit position, a linear approximation is being made.

4.5. Goodness of Fit

There appears to be no perfect general goodness-of-fit test. Given a dataset generated under the null hypothesis, one can usually find a test which rejects the null hypothesis (and this may be taken as a warning that choosing the test after you see the data is dangerous). Given a dataset generated under an alternative hypothesis, one can usually find a test for which the null passes. It seems advisable to think about what one wants to test for in choosing the test.

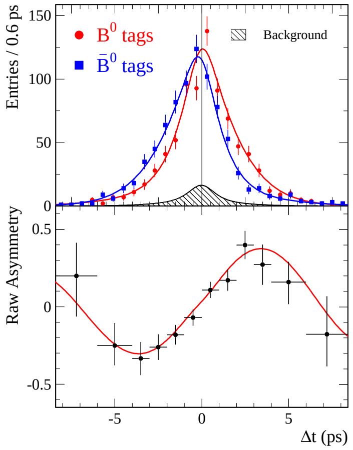


Figure 12: Measurement of CP violation (BaBar) [11]. The upper plot shows the measurement (points) of the time distributions for B^0 and $\bar{B}^0$ decays to selected CP eigenstates. The curves show the result of a maximum likelihood fit to the data. The lower plot shows the time-dependent asymmetry between the B^0 and $\bar{B}^0$ decays, again with the fitted curve overlaid. The asymmetry would be zero in the absence of CP violation.

For example, Fig. 12 shows data used in a measurement of CP violation by BaBar. A likelihood ratio (or a chi-square) test of the time distribution may be a good test for the lifetime fit to the data, but it may have little sensitivity to testing the goodness-of-fit of the CP asymmetry, which is a low-fequency question.

So far, BaBar generally uses likelihood ratio tests or chi-square tests if appropriate. The Kolmogorov-Smirnov test is also used. If a test statistic such as the likelihood ratio is used, then a Monte Carlo evaluation of the distribution of the statistic is recommended, rather than assuming an asymptotic property.

4.6. Consistency of Analyses

BaBar has encountered several times the question of whether a new analysis is consistent with an old analysis. Often, the new analysis is a combination of additional data plus changed (improved) analysis of original data. The stickiest issue is handling the correlation in testing for consistency in the overlapping data. People sometimes have difficulty understanding that statistical differences can arise even comparing results based on the same events, so we expound on this.

Given a sampling $\hat{\theta}_1, \hat{\theta}_2$ from a bivariate normal distribution $N(\theta, \sigma_1, \sigma_2, \rho)$, with $\langle\hat{\theta}_1\rangle = \langle\hat{\theta}_2\rangle = \theta$, the difference $\Delta\theta \equiv \hat{\theta}_2 - \hat{\theta}_1$ is $N(0, \sigma)$-distributed with

$\sigma^2 = \sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2$. If the correlation is unknown, all we can say is that the variance of the difference is in the range $(\sigma_1 - \sigma_2)^2 \dots (\sigma_1 + \sigma_2)^2$. If we at least believe $\rho \geq 0$ then the maximum variance of the difference is $\sigma_1^2 + \sigma_2^2$.

Suppose we measure a neutrino mass, m, in a sample of $n = 10$ independent events. The measurements are $x_i, i = 1, \dots, 10$. Assume the sampling distribution for x_i is $N(m, \sigma_i)$.

We may form *unbiased* estimator, $\hat{m}_1$, for m:

$$\hat{m}_1 = \tfrac{1}{n}\textstyle\sum_{i=1}^n x_i \pm \sqrt{\tfrac{1}{n^2}\sum_{i=1}^n \sigma_i^2}. \tag{9}$$

The result (from a Monte Carlo simulation) is $\hat{m}_1 = 0.058 \pm 0.039$.

Then we notice that we have some further information which might be useful: we know the experimental resolutions, σ_i for each measurement. We form another *unbiased* estimator, $\hat{m}_2$, for m:

$$\hat{m}_2 = \textstyle\sum_{i=1}^n \frac{x_i}{\sigma_i^2} \big/ \sum_{i=1}^n \frac{1}{\sigma_i^2} \pm 1 \big/ \sqrt{\sum_{i=1}^n \frac{1}{\sigma_i^2}}. \tag{10}$$

The result (from the same simulation, i.e., from the *same events*) is $\hat{m}_1 = 0.000 \pm 0.016$.

The results are certainly correlated, so the question of consistency arises (we know the error on the difference is between 0.023 and 0.055). In this example, the difference between the results is 0.058 ± 0.036, where the 0.036 error includes the correlation ($\rho = 0.41$).

Art Snyder has developed an approximate formula for evaluating the correlation in a comparison of maximum likelihood analyses. Suppose we perform two maximum likelihood analysis, with event likelihoods $\mathcal{L}_1$, $\mathcal{L}_2$, on the same set of N events [n.b., we may use different information in each analysis]. The results are estimators $\hat{\theta}_1$, $\hat{\theta}_2$ for parameter θ (restricting to the one-dimensional case for simplicity). The correlation coefficient ρ may be estimated according to:

$$\rho \approx \frac{\sum_{i=1}^N R_i \frac{d\ln\mathcal{L}_{1i}}{d\theta}|_{\theta=\hat{\theta}_1} \frac{d\ln\mathcal{L}_{2i}}{d\theta}|_{\theta=\hat{\theta}_2}}{\sqrt{\left(\sum_{i=1}^N \frac{d^2\ln\mathcal{L}_{1i}}{d\theta^2}|_{\theta=\theta_0}\right)\left(\sum_{i=1}^N \frac{d^2\ln\mathcal{L}_{2i}}{d\theta^2}|_{\theta=\theta_0}\right)}}, \tag{11}$$

where (θ_0 is an expansion reference point):

$$R_i = \left[1 - (\hat{\theta}_1 - \theta_0)\frac{d^2\ln\mathcal{L}_{1i}}{d\theta^2}|_{\theta=\theta_0} \bigg/ \frac{d\ln\mathcal{L}_{1i}}{d\theta}|_{\theta=\theta_0}\right] \left[1 - (\hat{\theta}_2 - \theta_0)\frac{d^2\ln\mathcal{L}_{2i}}{d\theta^2}|_{\theta=\theta_0} \bigg/ \frac{d\ln\mathcal{L}_{2i}}{d\theta}|_{\theta=\theta_0}\right].$$

If $\theta_0 \approx \hat{\theta}_1 \approx \hat{\theta}_2$, then

$$\rho \approx \tilde{\sigma}_{\theta_1}\tilde{\sigma}_{\theta_2} \sum_{i=1}^N \frac{d\ln\mathcal{L}_{1i}}{d\theta}|_{\theta=\theta_0} \frac{d\ln\mathcal{L}_{2i}}{d\theta}|_{\theta=\theta_0}, \tag{12}$$

where $\tilde{\sigma}_{\theta_k}^2 \equiv 1/\sum_{i=1}^N \left(\frac{d\mathcal{L}_{ki}}{d\theta}|_{\theta=\theta_0}\right)^2$.

Let us look at a real example of the consistency question in a BaBar analysis, the measurement of the CP-violation parameter $\sin 2\beta$. In August 2001, we published a result based on a dataset of 32×10^6 $B\bar{B}$ pairs [12]:

$$\sin 2\beta = 0.59 \pm 0.14(\text{stat}) \pm 0.05(\text{syst}) \tag{13}$$

An updated result was produced in March 2002, based on 62×10^6 $B\bar{B}$ pairs [13]:

$$\sin 2\beta = 0.75 \pm 0.09(\text{stat}) \pm 0.04(\text{syst}) \tag{14}$$

The second result includes the earlier data, re-reconstructed. The analysis is not simply counting events; it involves multivariate maximum likelihood fits, reprocessing changes, and relative likelihoods for an event to be signal or background, for example. The question is, are the two results statistically consistent?

If these were independent data sets, a difference of 0.16 ± 0.17 would not be a worry. The issue is the correlation. A specialized analysis deriving from Eqn. 11 is performed on the events in common between the two analyses. A correlation of $\rho = 0.87$ is deduced, yielding a difference of $\sim 2.2\sigma$. This corresponds to a probability of 3%, which is small enough that we noticed, and looked hard for possible systematic problems, but not so small to be alarming, especially in an experiment with many such tests being made.

There has been some impression that BaBar may be seeing more diffences between old vs updated results than people are used to, and the question arises whether BaBar is making mistakes. The answer to this seems to be, first of all, based on studies such as the above, there is no compelling statistical evidence to support the contention that mistakes are being made. There should be differences, purely due to statistical fluctuations, among results, and BaBar sees nothing clearly beyond what might be expected from statistics. The second part of the answer is a speculation to why the impression may exist. BaBar is different from most other experiments in that it makes extensive use of the blind methodology. There is little opportunity to react to observed differences with further changes in analysis. Without using the blind methodology, there is the potential for bias, tending towards making results agree with earlier results better than they should.

5. REFLECTIONS

It is my observation that statistical sophistication in particle physics (not specific to BaBar) has grown significantly, not so much in the choice of methods, which are often long-established, but in the understanding attached to them. People now understand that there is a choice of approach between Bayesian example, BaBar now relies heavily on blind methodology.

BaBar adopts frequency statistics for describing results, and much attention is devoted to Monte Carlo validation and verification of coverage. The use of the Bayesian approach in high energy physics, including BaBar, is still not mature: There is no established methodology for choosing the prior distribution, other than to default on a uniform prior. The justification for this is basically that it usually doesn't matter very much. There are, however, even issues still in frequency statistics. Controversies involve such notions as restricting to the "physical region", or that the presence of backgrounds should "always" lead to higher upper limits. Both of these notions are not a concern in the BaBar recommendations.

BaBar is attempting to provide a coherent, documented approach to its use of statistics in its results. This is very much a work in progress.

Acknowledgments

I would like to thank Louis Lyons for organizing an informative and stimulating conference. I am grateful to my collaborators on BaBar for many interesting discussions of statistical issues. Work supported in part by Department of Energy grant DE-FG03-92ER40701 and contract DE-AC03-76SF00515.

References

[1] BaBar Statistics Working Group, http://www.slac.stanford.edu/BFROOT/www/Statistics/index.html

[2] A. Roodman, http://www-conf.slac.stanford.edu/phystat2003/talks/roodman/roodman-blind-stat2003.pdf, Talk given at PHYSTAT2003 (2003).

[3] J. Walsh, First International Conference on Flavor Physics and CP Violation (FPCP 2002), http://www.hep.upenn.edu/FPCP/

[4] B. Aubert et al. (BaBar collaboration), Phys. Rev. Lett., **90**, 242001 (2003).

[5] B. Knuteson, http://www-conf.slac.stanford.edu/phystat2003/talks/knuteson/, Talk given at PHYSTAT2003 (2003).

[6] B. Aubert et al. (BaBar collaboration), Phys. Rev. Lett., **91**, 171801 (2003).

[7] U. Egede, International Workshop on Frontier Science, Frascati, October 6-11, 2002.

[8] R. Barlow, "A Calculator for Confidence Intervals", MAN/HEP/2001/04.

[9] B. Schwarzschild, Physics Today, http://www.physicstoday.org/pt/vol-54/iss-9/p19.html

[10] G. Feldman and R. Cousins, Phys. Rev. D, **57**, 3873 (1998).

[11] B. Aubert et al. (BaBar collaboration), Phys. Rev. Lett., **89**, 201802 (2002).

[12] B. Aubert et al. (BaBar collaboration), Phys. Rev. Lett., **87**, 091801 (2001).

[13] B. Aubert et al. (BaBar Collaboration), SLAC-PUB-9153 (2002).

Recent Advances in Predictive (Machine) Learning

Jerome H. Friedman
Deptartment of Statisitcs and Stanford Linear Accelerator Center,
Stanford University, Stanford, CA 94305

Prediction involves estimating the unknown value of an attribute of a system under study given the values of other measured attributes. In prediction (machine) learning the prediction rule is derived from data consisting of previously solved cases. Most methods for predictive learning were originated many years ago at the dawn of the computer age. Recently two new techniques have emerged that have revitalized the field. These are support vector machines and boosted decision trees. This paper provides an introduction to these two new methods tracing their respective ancestral roots to standard kernel methods and ordinary decision trees.

1. INTRODUCTION

The predictive or machine learning problem is easy to state if difficult to solve in general. Given a set of measured values of attributes/characteristics/properties on a object (observation) $\mathbf{x} = (x_1, x_2, \cdots, x_n)$ (often called "variables") the goal is to predict (estimate) the unknown value of another attribute y. The quantity y is called the "output" or "response" variable, and $\mathbf{x} = \{x_1, \cdots, x_n\}$ are referred to as the "input" or "predictor" variables. The prediction takes the form of function

$$\hat{y} = F(x_1, x_2, \cdot \cdot \cdot, x_n) = F(\mathbf{x})$$

that maps a point $\mathbf{x}$ in the space of all joint values of the predictor variables, to a point $\hat{y}$ in the space of response values. The goal is to produce a "good" predictive $F(\mathbf{x})$. This requires a definition for the quality, or lack of quality, of any particular $F(\mathbf{x})$. The most commonly used measure of lack of quality is prediction "risk". One defines a "loss" criterion that reflects the cost of mistakes: $L(y, \hat{y})$ is the loss or cost of predicting a value $\hat{y}$ for the response when its true value is y. The prediction risk is defined as the average loss over all predictions

$$R(F) = E_{y\mathbf{x}} L(y, F(\mathbf{x})) \qquad (1)$$

where the average (expected value) is over the joint (population) distribution of all of the variables $(y, \mathbf{x})$ which is represented by a probability density function $p(y, \mathbf{x})$. Thus, the goal is to find a mapping function $F(\mathbf{x})$ with low predictive risk.

Given a function f of elements w in some set, the choice of w that gives the smallest value of $f(w)$ is called $\arg\min_w f(w)$. This definition applies to all types of sets including numbers, vectors, colors, or functions. In terms of this notation the optimal predictor with lowest predictive risk (called the "target function") is given by

$$F^* = \arg\min_F R(F). \qquad (2)$$

Given joint values for the input variables $\mathbf{x}$, the optimal prediction for the output variable is $\hat{y} = F^*(\mathbf{x})$.

When the response takes on numeric values $y \in R^1$, the learning problem is called "regression" and commonly used loss functions include absolute error $L(y, F) = |y - F|$, and even more commonly squared–error $L(y, F) = (y - F)^2$ because algorithms for minimization of the corresponding risk tend to be much simpler. In the "classification" problem the response takes on a discrete set of K unorderable categorical values (names or class labels) $y, F \in \{c_1, \cdot \cdot \cdot, c_K\}$ and the loss criterion $L_{y,F}$ becomes a discrete $K \times K$ matrix.

There are a variety of ways one can go about trying to find a good predicting function $F(\mathbf{x})$. One might seek the opinions of domain experts, formally codified in the "expert systems" approach of artificial intelligence. In predictive or machine learning one uses data. A "training" data base

$$D = \{y_i, x_{i1}, x_{i2}, \cdot \cdot \cdot, x_{in}\}_1^N = \{y_i, \mathbf{x}_i\}_1^N \qquad (3)$$

of N previously solved cases is presumed to exist for which the values of all variables (response and predictors) have been jointly measured. A "learning" procedure is applied to these data in order to extract (estimate) a good predicting function $F(\mathbf{x})$. There are a great many commonly used learning procedures. These include linear/logistic regression, neural networks, kernel methods, decision trees, multivariate splines (MARS), etc. For descriptions of a large number of such learning procedures see Hastie, Tibshirani and Friedman 2001.

Most machine learning procedures have been around for a long time and most research in the field has concentrated on producing refinements to these long standing methods. However, in the past several years there has been a revolution in the field inspired by the introduction of two new approaches: the extension of kernel methods to support vector machines (Vapnik 1995), and the extension of decision trees by boosting (Freund and Schapire 1996, Friedman 2001). It is the purpose of this paper to provide an introduction to these new developments. First the classic kernel and decision tree methods are introduced. Then the extension of kernels to support vector machines is described, followed by a description of applying boosting to extend decision tree methods. Finally, similarities and differences between these two approaches will be discussed.

Although arguably the most influential recent developments, support vector machines and boosting are not the only important advances in machine learning in the past several years. Owing to space limitations these are the ones discussed here. There have been other important developments that have considerably advanced the field as well. These include (but are not limited to) the bagging and random forest techniques of Breiman 1996 and 2001 that are somewhat related to boosting, and the reproducing kernel Hilbert space methods of Wahba 1990 that share similarities with support vector machines. It is hoped that this article will inspire the reader to investigate these as well as other machine learning procedures.

2. KERNEL METHODS

Kernel methods for predictive learning were introduced by Nadaraya (1964) and Watson (1964). Given the training data (3), the response estimate $\hat{y}$ for a set of joint values $\mathbf{x}$ is taken to be a weighted average of the training responses $\{y_i\}_1^N$:

$$\hat{y} = F_N(\mathbf{x}) = \sum_{i=1}^{N} y_i \, K(\mathbf{x}, \mathbf{x}_i) \bigg/ \sum_{i=1}^{N} K(\mathbf{x}, \mathbf{x}_i). \tag{4}$$

The weight $K(\mathbf{x}, \mathbf{x}_i)$ assigned to each response value y_i depends on its location $\mathbf{x}_i$ in the predictor variable space and the location $\mathbf{x}$ where the prediction is to be made. The function $K(\mathbf{x}, \mathbf{x}')$ defining the respective weights is called the "kernel function", and it defines the kernel method. Often the form of the kernel function is taken to be

$$K(\mathbf{x}, \mathbf{x}') = g(d(\mathbf{x}, \mathbf{x}')/\sigma) \tag{5}$$

where $d(\mathbf{x}, \mathbf{x}')$ is a defined "distance" between $\mathbf{x}$ and $\mathbf{x}'$, σ is a scale ("smoothing") parameter, and $g(z)$ is a (usually monotone) decreasing function with increasing z; often $g(z) = \exp(-z^2/2)$. Using this kernel (5), the estimate $\hat{y}$ (4) is a weighted average of $\{y_i\}_1^N$, with more weight given to observations i for which $d(\mathbf{x}, \mathbf{x}_i)$ is small. The value of σ defines "small". The distance function $d(\mathbf{x}, \mathbf{x}')$ must be specified for each particular application.

Kernel methods have several advantages that make them potentially attractive. They represent a universal approximator; as the training sample size N becomes arbitrarily large, $N \to \infty$, the kernel estimate (4) (5) approaches the optimal predicting target function (2), $F_N(\mathbf{x}) \to F^*(\mathbf{x})$, provided the value chosen for the scale parameter σ as a function of N approaches zero, $\sigma(N) \to 0$, at a slower rate than $1/N$. This result holds for almost any distance function $d(\mathbf{x}, \mathbf{x}')$; only very mild restrictions (such as convexity) are required. Another advantage of kernel methods is that no training is required to build a model; the training data set *is* the model. Also, the procedure is conceptually quite simple and easily explained.

Kernel methods suffer from some disadvantages that have kept them from becoming highly used in practice, especially in data mining applications. Since there is no model, they provide no easily understood model summary. Thus, they cannot be easily interpreted. There is no way to discern how the function $F_N(\mathbf{x})$ (4) depends on the respective predictor variables $\mathbf{x}$. Kernel methods produce a "black–box" prediction machine. In order to make each prediction, the kernel method needs to examine the entire data base. This requires enough random access memory to store the entire data set, and the computation required to make each prediction is proportional to the training sample size N. For large data sets this is much slower than that for competing methods.

Perhaps the most serious limitation of kernel methods is statistical. For any *finite* N, performance (prediction accuracy) depends *critically* on the chosen distance function $d(\mathbf{x}, \mathbf{x}')$, especially for regression $y \in R^1$. When there are more than a few predictor variables, even the largest data sets produce a very sparse sampling in the corresponding n–dimensional predictor variable space. This is a consequence of the so called "curse–of–dimensionality" (Bellman 1962). In order for kernel methods to perform well, the distance function must be carefully matched to the (unknown) target function (2), and the procedure is not very robust to mismatches.

As an example, consider the often used Euclidean distance function

$$d(\mathbf{x}, \mathbf{x}') = \left[\sum_{j=1}^{n} (x_j - x_j')^2\right]^{1/2}. \tag{6}$$

If the target function $F^*(\mathbf{x})$ dominately depends on only a small subset of the predictor variables, then performance will be poor because the kernel function (5) (6) depends on all of the predictors with equal strength. If one happened to know *which* variables were the important ones, an appropriate kernel could be constructed. However, this knowledge is often not available. Such "kernel customizing" is a requirement with kernel methods, but it is difficult to do without considerable a priori knowledge concerning the problem at hand.

The performance of kernel methods tends to be fairly insensitive to the detailed choice of the function $g(z)$ (5), but somewhat more sensitive to the value chosen for the smoothing parameter σ. A good value depends on the (usually unknown) smoothness properties of the target function $F^*(\mathbf{x})$, as well as the sample size N and the signal/noise ratio.

3. DECISION TREES

Decision trees were developed largely in response to the limitations of kernel methods. Detailed descriptions are contained in monographs by Brieman, Friedman, Olshen and Stone 1983, and by Quinlan 1992. The minimal description provided here is intended as an introduction sufficient for understanding what follows.

A decision tree partitions the space of all joint predictor variable values $\mathbf{x}$ into J–disjoint regions $\{R_j\}_1^J$. A response value $\hat{y}_j$ is assigned to each corresponding region R_j. For a given set of joint predictor values $\mathbf{x}$, the tree prediction $\hat{y} = T_J(\mathbf{x})$ assigns as the response estimate, the value assigned to the region containing $\mathbf{x}$

$$\mathbf{x} \in R_j \Rightarrow T_J(\mathbf{x}) = \hat{y}_j \,. \tag{7}$$

Given a set of regions, the optimal response values associated with each one are easily obtained, namely the value that minimizes prediction risk in that region

$$\hat{y}_j = \arg\min_{y'} E_y[L(y, y') \,|\, \mathbf{x} \in R_j]. \tag{8}$$

The difficult problem is to find a good set of regions $\{R_j\}_1^J$. There are a huge number of ways to partition the predictor variable space, the vast majority of which would provide poor predictive performance. In the context of decision trees, choice of a particular partition directly corresponds to choice of a distance function $d(\mathbf{x}, \mathbf{x}')$ and scale parameter σ in kernel methods. Unlike with kernel methods where this choice is the responsibility of the user, decision trees attempt to use the data to estimate a good partition.

Unfortunately, finding the optimal partition requires computation that grows exponentially with the number of regions J, so that this is only possible for very small values of J. All tree based methods use a greedy top–down recursive partitioning strategy to induce a good set of regions given the training data set (3). One starts with a single region covering the entire space of all joint predictor variable values. This is partitioned into two regions by choosing an optimal splitting predictor variable x_j and a corresponding optimal split point s. Points $\mathbf{x}$ for which $x_j \leq s$ are defined to be in the left daughter region, and those for which $x_j > s$ comprise the right daughter region. Each of these two daughter regions is then itself optimally partitioned into two daughters of its own in the same manner, and so on. This recursive partitioning continues until the observations within each region all have the same response value y. At this point a recursive recombination strategy ("tree pruning") is employed in which sibling regions are in turn merged in a bottom–up manner until the number of regions J^* that minimizes an estimate of future prediction risk is reached (see Breiman *et al* 1983, Ch. 3).

3.1. Decision tree properties

Decision trees are the most popular predictive learning method used in data mining. There are a number of reasons for this. As with kernel methods, decision trees represent a universal method. As the training data set becomes arbitrarily large, $N \to \infty$, tree based predictions (7) (8) approach those of the target function (2), $T_J(\mathbf{x}) \to F^*(\mathbf{x})$, provided the number of regions grows arbitrarily large, $J(N) \to \infty$, but at rate slower than N.

In contrast to kernel methods, decision trees do produce a model summary. It takes the form of a binary tree graph. The root node of the tree represents the entire predictor variable space, and the (first) split into its daughter regions. Edges connect the root to two descendent nodes below it, representing these two daughter regions and their respective splits, and so on. Each internal node of the tree represents an intermediate region and its optimal split, defined by a one of the predictor variables x_j and a split point s. The terminal nodes represent the final region set $\{R_j\}_1^J$ used for prediction (7). It is this binary tree graphic that is most responsible for the popularity of decision trees. No matter how high the dimensionality of the predictor variable space, or how many variables are actually used for prediction (splits), the entire model can be represented by this two–dimensional graphic, which can be plotted and then examined for interpretation. For examples of interpreting binary tree representations see Breiman *et al* 1983 and Hastie, Tibshirani and Friedman 2001.

Tree based models have other advantages as well that account for their popularity. Training (tree building) is relatively fast, scaling as $nN \log N$ with the number of variables n and training observations N. Subsequent prediction is extremely fast, scaling as $\log J$ with the number of regions J. The predictor variables need not all be numeric valued. Trees can seamlessly accommodate binary and categorical variables. They also have a very elegant way of dealing with missing variable values in both the training data and future observations to be predicted (see Breiman *et al* 1983, Ch. 5.3).

One property that sets tree based models apart from all other techniques is their invariance to monotone transformations of the predictor variables. Replacing any subset of the predictor variables $\{x_j\}$ by (possibly different) arbitrary strictly monotone functions of them $\{\, x_j \leftarrow m_j(x_j)\}$, gives rise to the same tree model. Thus, there is no issue of having to experiment with different possible transformations $m_j(x_j)$ for each individual predictor x_j, to try to find the best ones. This invariance provides immunity to the presence of extreme values "outliers" in the predictor variable space. It also provides invariance to changing the measurement scales of the predictor variables, something to which kernel methods can be very sensitive.

Another advantage of trees over kernel methods is fairly high resistance to irrelevant predictor variables. As discussed in Section 2, the presence of many such irrelevant variables can highly degrade the performance of kernel methods based on generic kernels that involve all of the predictor variables such as (6). Since the recursive tree building algorithm estimates the optimal variable on which to split at each step, predictors unrelated to the response tend not to be chosen for splitting. This is a consequence of attempting to find a good partition based on the data. Also, trees have few tunable parameters so they can be used as an "off–the–shelf" procedure.

The principal limitation of tree based methods is that in situations not especially advantageous to them, their performance tends not to be competitive with other methods that might be used in those situations. One problem limiting accuracy is the piecewise–constant nature of the predicting model. The predictions $\hat{y}_j$ (8) are constant within each region R_j and sharply discontinuous across region boundaries. This is purely an artifact of the model, and target functions $F^*(\mathbf{x})$ (2) occurring in practice are not likely to share this property. Another problem with trees is instability. Changing the values of just a few observations can dramatically change the structure of the tree, and substantially change its predictions. This leads to high variance in potential predictions $T_J(\mathbf{x})$ at any particular prediction point $\mathbf{x}$ over different training samples (3) that might be drawn from the system under study. This is especially the case for large trees.

Finally, trees fragment the data. As the recursive splitting proceeds each daughter region contains fewer observations than its parent. At some point regions will contain too few observations and cannot be further split. Paths from the root to the terminal nodes tend to contain on average a relatively small fraction of all of the predictor variables that thereby define the region boundaries. Thus, each prediction involves only a relatively small number of predictor variables. If the target function is influenced by only a small number of (potentially different) variables in different local regions of the predictor variable space, then trees can produce accurate results. But, if the target function depends on a substantial fraction of the predictors everywhere in the space, trees will have problems.

4. RECENT ADVANCES

Both kernel methods and decision trees have been around for a long time. Trees have seen active use, especially in data mining applications. The classic kernel approach has seen somewhat less use. As discussed above, both methodologies have (different) advantages and disadvantages. Recently, these two technologies have been completely revitalized in different ways by addressing different aspects of their corresponding weaknesses; support vector machines (Vapnik 1995) address the computational problems of kernel methods, and boosting (Freund and Schapire 1996, Friedman 2001) improves the accuracy of decision trees.

4.1. Support vector machines (SVM)

A principal goal of the SVM approach is to fix the computational problem of predicting with kernels (4). As discussed in Section 2, in order to make a kernel prediction a pass over the entire training data base is required. For large data sets this can be too time consuming and it requires that the entire data base be stored in random access memory.

Support vector machines were introduced for the two–class classification problem. Here the response variable realizes only two values (class labels) which can be respectively encoded as

$$y = \begin{cases} +1 & \text{label} = \text{class } 1 \\ -1 & \text{label} = \text{class } 2 \end{cases} . \tag{9}$$

The average or expected value of y given a set of joint predictor variable values $\mathbf{x}$ is

$$E\,[y \,|\, \mathbf{x}] = 2 \cdot \Pr(y = +1 \,|\, \mathbf{x}) - 1. \tag{10}$$

Prediction error rate is minimized by predicting at $\mathbf{x}$ the class with the highest probability, so that the optimal prediction is given by

$$y^*(\mathbf{x}) = sign(E\,[y \,|\, \mathbf{x}]).$$

From (4) the kernel estimate of (10) based on the training data (3) is given by

$$\hat{E}\,[y \,|\, \mathbf{x}] = F_N(\mathbf{x}) = \sum_{i=1}^{N} y_i\, K(\mathbf{x}, \mathbf{x}_i) \Bigg/ \sum_{i=1}^{N} K(\mathbf{x}, \mathbf{x}_i) \tag{11}$$

and, assuming a strictly non negative kernel $K(\mathbf{x}, \mathbf{x}_i)$, the prediction estimate is

$$\hat{y}(\mathbf{x}) = sign(\hat{E}\,[y \,|\, \mathbf{x}]) = sign\left(\sum_{i=1}^{N} y_i\, K(\mathbf{x}, \mathbf{x}_i)\right). \tag{12}$$

Note that ignoring the denominator in (11) to obtain (12) removes information concerning the absolute value of $\Pr(y = +1 \,|\, \mathbf{x})$; only the estimated sign of (10) is retained for classification.

A support vector machine is a weighted kernel classifier

$$\hat{y}(\mathbf{x}) = sign\left(a_0 + \sum_{i=1}^{N} \alpha_i \, y_i \, K(\mathbf{x}, \mathbf{x}_i)\right). \qquad (13)$$

Each training observation $(y_i, \mathbf{x}_i)$ has an associated coefficient α_i additionally used with the kernel $K(\mathbf{x}, \mathbf{x}_i)$ to evaluate the weighted sum (13) comprising the kernel estimate $\hat{y}(\mathbf{x})$. The goal is to choose a set of coefficient values $\{\alpha_i\}_1^N$ so that many $\alpha_i = 0$ while still maintaining prediction accuracy. The observations associated with non zero valued coefficients $\{\mathbf{x}_i \,|\, \alpha_i \neq 0\}$ are called "support vectors". Clearly from (13) only the support vectors are required to do prediction. If the number of support vectors is a small fraction of the total number of observations computation required for prediction is thereby much reduced.

4.1.1. Kernel trick

In order to see how to accomplish this goal consider a different formulation. Suppose that instead of using the original measured variables $\mathbf{x} = (x_1, \cdot\cdot\cdot, x_n)$ as the basis for prediction, one constructs a very large number of (nonlinear) functions of them

$$\{z_k = h_k(\mathbf{x})\}_1^M \qquad (14)$$

for use in prediction. Here each $h_k(\mathbf{x})$ is a different function (transformation) of $\mathbf{x}$. For any given $\mathbf{x}$, $\mathbf{z} = \{z_k\}_1^M$ represents a point in a M–dimensional space where $M >> \dim(\mathbf{x}) = n$. Thus, the number of "variables" used for classification is dramatically expanded. The procedure constructs simple *linear* classifier in $\mathbf{z}$–space

$$\begin{aligned}\hat{y}(\mathbf{z}) &= sign\left(\beta_0 + \sum_{k=1}^{M} \beta_k \, z_k\right) \\ &= sign\left(\beta_0 + \sum_{k=1}^{M} \beta_k \, h_k\,(\mathbf{x})\right).\end{aligned}$$

This is a highly *non*–linear classifier in $\mathbf{x}$–space owing to the nonlinearity of the derived transformations $\{h_k(\mathbf{x})\}_1^M$.

An important ingredient for calculating such a linear classifier is the inner product between the points representing two observations i and j

$$\begin{aligned}\mathbf{z}_i^T \mathbf{z}_j &= \sum_{k=1}^{M} z_{ik} z_{jk} \qquad (15) \\ &= \sum_{k=1}^{M} h_k\,(\mathbf{x}_i)\, h_k\,(\mathbf{x}_j) \\ &= H(\mathbf{x}_i, \mathbf{x}_j).\end{aligned}$$

This (highly nonlinear) function of the $\mathbf{x}$–variables, $H(\mathbf{x}_i, \mathbf{x}_j)$, defines the simple bilinear inner product $\mathbf{z}_i^T \mathbf{z}_j$ in $\mathbf{z}$–space.

Suppose for example, the derived variables (14) were taken to be all d–degree polynomials in the original predictor variables $\{x_j\}_1^n$. That is $z_k = x_{i_1(k)} x_{i_2(k)} \cdot\cdot\cdot x_{i_d(k)}$, with k labelling each of the $M = (n{+}1)^d$ possible sets of d integers, $0 \leq i_j(k) \leq n$, and with the added convention that $x_0 = 1$ even though x_0 is not really a component of the vector $\mathbf{x}$. In this case the number of derived variables is $M = (n + 1)^d$, which is the order of computation for obtaining $\mathbf{z}_i^T \mathbf{z}_j$ directly from the $\mathbf{z}$ variables. However, using

$$\mathbf{z}_i^T \mathbf{z}_j = H(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i^T \mathbf{x}_j + 1)^d \qquad (16)$$

reduces the computation to order n, the much smaller number of originally measured variables. Thus, if for any particular set of derived variables (14), the function $H(\mathbf{x}_i, \mathbf{x}_j)$ that defines the corresponding inner products $\mathbf{z}_i^T \mathbf{z}_j$ in terms of the original $\mathbf{x}$–variables can be found, computation can be considerably reduced.

As an example of a very simple linear classifier in $\mathbf{z}$–space, consider one based on nearest–means.

$$\hat{y}(\mathbf{z}) = sign(\,||\,\mathbf{z} - \bar{\mathbf{z}}_-\,||^2 - ||\,\mathbf{z} - \bar{\mathbf{z}}_+\,||^2). \qquad (17)$$

Here $\bar{\mathbf{z}}_\pm$ are the respective means of the $y = +1$ and $y = -1$ observations

$$\bar{\mathbf{z}}_\pm = \frac{1}{N_\pm} \sum_{y_i = \pm 1} \mathbf{z}_i.$$

For simplicity, let $N_+ = N_- = N/2$. Choosing the midpoint between $\bar{\mathbf{z}}_+$ and $\bar{\mathbf{z}}_-$ as the coordinate system origin, the decision rule (17) can be expressed as

$$\begin{aligned}\hat{y}(\mathbf{z}) &= sign(\mathbf{z}^T(\bar{\mathbf{z}}_+ - \bar{\mathbf{z}}_-)) \qquad (18) \\ &= sign\left(\sum_{y_i=1} \mathbf{z}^T \mathbf{z}_i - \sum_{y_i=-1} \mathbf{z}^T \mathbf{z}_i\right) \\ &= sign\left(\sum_{i=1}^{N} y_i \mathbf{z}^T \mathbf{z}_i\right) \\ &= sign\left(\sum_{i=1}^{N} y_i H(\mathbf{x}, \mathbf{x}_i)\right).\end{aligned}$$

Comparing this (18) with (12) (13), one sees that ordinary kernel rule ($\{\alpha_i = 1\}_1^N$) in $\mathbf{x}$–space is the nearest–means classifier in the $\mathbf{z}$–space of derived variables (14) whose inner product is given by the kernel function $\mathbf{z}_i^T \mathbf{z}_j = K(\mathbf{x}_i, \mathbf{x}_j)$. Therefore to construct an (implicit) nearest means classifier in $\mathbf{z}$–space, all computations can be done in $\mathbf{x}$–space because they only depend on evaluating inner products. The explicit transformations (14) need never be defined or even known.

4.1.2. Optimal separating hyperplane

Nearest–means is an especially simple linear classifier in $\mathbf{z}$–space and it leads to no compression: $\{\alpha_i = 1\}_1^N$ in (13). A support vector machine uses a more "realistic" linear classifier in $\mathbf{z}$–space, that can also be computed using only inner products, for which often many of the coefficients have the value zero ($\alpha_i = 0$). This classifier is the "optimal" separating hyperplane (OSH).

We consider first the case in which the observations representing the respective two classes are linearly separable in $\mathbf{z}$–space. This is often the case since the dimension M (14) of that (implicitly defined) space is very large. In this case the OSH is the unique hyperplane that separates two classes while maximizing the distance to the closest points in each class. Only this set of closest points equidistant to the OSH are required to define it. These closest points are called the support points (vectors). Their number can range from a minimum of two to a maximum of the training sample size N. The "margin" is defined to be the distance of support points from OSH. The $\mathbf{z}$–space linear classifier is given by

$$\hat{y}(\mathbf{z}) = sign\left(\beta_0^* + \sum_{k=1}^{M} \beta_k^* z_k\right) \tag{19}$$

where $(\beta_0^*,\ \beta^* = \{\beta_k^*\}_1^M\ )$ define the OSH. Their values can be determined using standard quadratic programming techniques.

An OSH can also be defined for the case when the two classes are not separable in $\mathbf{z}$–space by allowing some points to be on wrong side of their class margin. The amount by which they are allowed to do so is a regularization (smoothing) parameter of the procedure. In both the separable and non separable cases the solution parameter values (β_0^*, β^*) (19) are defined only by points close to boundary between the classes. The solution for β^* can be expressed as

$$\beta^* = \sum_{i=1}^{N} \alpha_i^*\, y_i\, \mathbf{z}_i$$

with $\alpha_i^* \neq 0$ only for points on, or on the wrong side of, their class margin. These are the support vectors. The SVM classifier is thereby

$$\begin{aligned}\hat{y}(\mathbf{z}) &= sign\left(\beta_0^* + \sum_{i=1}^{N} \alpha_i^*\, y_i\, \mathbf{z}^T \mathbf{z}_i\right) \\ &= sign\left(\beta_0^* + \sum_{\alpha_i^* \neq 0} \alpha_i^*\, y_i\, K(\mathbf{x}, \mathbf{x}_i)\right).\end{aligned}$$

This is a weighted kernel classifier involving only support vectors. Also (not shown here), the quadratic program used to solve for the OSH involves the data only through the inner products $\mathbf{z}_i^T \mathbf{z}_j = K(\mathbf{x}_i, \mathbf{x}_j)$. Thus, one only needs to specify the kernel function to implicitly define $\mathbf{z}$–variables (kernel trick).

Besides the polynomial kernel (16), other popular kernels used with support vector machines are the "radial basis function" kernel

$$K(\mathbf{x}, \mathbf{x}') = \exp(-\,||\,\mathbf{x} - \mathbf{x}'\,||^2/2\sigma^2), \tag{20}$$

and the "neural network" kernel

$$K(\mathbf{x}, \mathbf{x}') = \tanh(a\,\mathbf{x}^T \mathbf{x}' + b). \tag{21}$$

Note that both of these kernels (20) (21) involve additional tuning parameters, and produce infinite dimensional derived variable (14) spaces ($M = \infty$).

4.1.3. Penalized learning formulation

The support vector machine was motivated above by the optimal separating hyperplane in the high dimensional space of the derived variables (14). There is another equivalent formulation in that space that shows that the SVM procedure is related to other well known statistically based methods. The parameters of the OSH (19) are the solution to

$$(\beta_0^*, \beta^*) = \arg\min_{\beta_0, \beta} \sum_{i=1}^{N} [1 - y_i(\beta_0 + \beta^T \mathbf{z}_i)]_+ + \lambda \cdot\, ||\,\beta\,||^2. \tag{22}$$

Here the expression $[\eta]_+$ represents the "positive part" of its argument; that is, $[\eta]_+ = \max(0, \eta)$. The "regularization" parameter λ is related to the SVM smoothing parameter mentioned above. This expression (22) represents a penalized learning problem where the goal is to minimize the empirical risk on the training data using as a loss criterion

$$L(y, F(\mathbf{z})) = [1 - yF(\mathbf{z})]_+, \tag{23}$$

where

$$F(\mathbf{z}) = \beta_0 + \beta^T \mathbf{z},$$

subject to an increasing penalty for larger values of

$$||\,\beta\,||^2 = \sum_{j=1}^{n} \beta_j^2. \tag{24}$$

This penalty (24) is well known and often used to regularize statistical procedures, for example linear least squares regression leading to ridge–regression (Hoerl and Kannard 1970)

$$(\beta_0^*, \beta^*) = \arg\min_{\beta_0, \beta} \sum_{i=1}^{N} [y_i - (\beta_0 + \beta^T \mathbf{z}_i)]^2 + \lambda \cdot\, ||\,\beta\,||^2. \tag{25}$$

The "hinge" loss criterion (23) is not familiar in statistics. However, it is closely related to one that is well known in that field, namely conditional negative log–likelihood associated with logistic regression

$$L(y, F(\mathbf{z})) = -\log[1 + e^{-yF(\mathbf{z})}]. \tag{26}$$

In fact, one can view the SVM hinge loss as a piecewise–linear approximation to (26). Unregularized logistic regression is one of the most popular methods in statistics for treating binary response outcomes (9). Thus, a support vector machine can be viewed as an approximation to *regularized* logistic regression (in $\mathbf{z}$–space) using the ridge–regression penalty (24).

This penalized learning formulation forms the basis for extending SVMs to the regression setting where the response variable y assumes numeric values $y \in R^1$, rather than binary values (9). One simply replaces the loss criterion (23) in (22) with

$$L(y, F(\mathbf{z})) = (|y - F(\mathbf{z})| - \varepsilon)_+. \tag{27}$$

This is called the "ε–insensitive" loss and can be viewed as a piecewise–linear approximation to the Huber 1964 loss

$$L(y, F(\mathbf{z})) = \begin{cases} |y - F(\mathbf{z})|^2/2 & |y - F(\mathbf{z})| \le \varepsilon \\ \varepsilon(|y - F(\mathbf{z})| - \varepsilon/2) & |y - F(\mathbf{z})| > \varepsilon \end{cases} \tag{28}$$

often used for robust regression in statistics. This loss (28) is a compromise between squared–error loss (25) and absolute–deviation loss $L(y, F(\mathbf{z})) = |y - F(\mathbf{z})|$. The value of the "transition" point ε differentiates the errors that are treated as "outliers" being subject to absolute–deviation loss, from the other (smaller) errors that are subject to squared–error loss.

4.1.4. SVM properties

Support vector machines inherit most of the advantages of ordinary kernel methods discussed in Section 2. In addition, they can overcome the computation problems associated with prediction, since only the support vectors ($\alpha_i \neq 0$ in (13)) are required for making predictions. If the number of support vectors is much smaller that than the total sample size N, computation is correspondingly reduced. This will tend to be the case when there is small overlap between the respective distributions of the two classes in the space of the original predictor variables $\mathbf{x}$ (small Bayes error rate).

The computational savings in prediction are bought by dramatic increase in computation required for training. Ordinary kernel methods (4) require no training; the data set is the model. The quadratic program for obtaining the optimal separating hyperplane (solving (22)) requires computation proportional to the *square* of the sample size (N^2), multiplied by the number of resulting support vectors. There has been much research on fast algorithms for training SVMs, extending computational feasibility to data sets of size $N \lesssim 30,000$ or so. However, they are still not feasible for really large data sets $N \gtrsim 100,000$.

SVMs share some of the disadvantages of ordinary kernel methods. They are a black–box procedure with little interpretive value. Also, as with all kernel methods, performance can be very sensitive to kernel (distance function) choice (5). For good performance the kernel needs to be matched to the properties of the target function $F^*(\mathbf{x})$ (2), which are often unknown. However, when there is a known "natural" distance for the problem, SVMs represent very powerful learning machines.

4.2. Boosted trees

Boosting decision trees was first proposed by Freund and Schapire 1996. The basic idea is rather than using just a single tree for prediction, a linear combination of (many) trees

$$F(\mathbf{x}) = \sum_{m=1}^{M} a_m T_m(\mathbf{x}) \tag{29}$$

is used instead. Here each $T_m(\mathbf{x})$ is a decision tree of the type discussed in Section 3 and a_m is its coefficient in the linear combination. This approach maintains the (statistical) advantages of trees, while often dramatically increasing accuracy over that of a single tree.

4.2.1. Training

The recursive partitioning technique for constructing a single tree on the training data was discussed in Section 3. Algorithm 1 describes a forward stagewise method for constructing a prediction machine based on a linear combination of M trees.

Algorithm 1

Forward stagewise boosting

```
1 $F_0(\mathbf{x}) = 0$
2 For $m = 1$ to $M$ do:
3       $(a_m, T_m(\mathbf{x})) = \arg\min_{a,T(\mathbf{x})}$
4             $\sum_{i=1}^{N} L(y_i, F_{m-1}(\mathbf{x}_i) + aT(\mathbf{x}_i))$
5       $F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}_i) + a_m T_m(\mathbf{x})$
6 EndFor
7 $F(\mathbf{x}) = F_M(\mathbf{x}) = \sum_{m=1}^{M} a_m T_m(\mathbf{x})$
```

The first line initializes the predicting function to everywhere have the value zero. Lines 2 and 6 control the M iterations of the operations associated with lines 3–5. At each iteration m there is a current predicting function $F_{m-1}(\mathbf{x})$. At the first iteration this is the initial function $F_0(\mathbf{x}) = 0$, whereas for $m > 1$ it is the linear combination of the $m - 1$ trees induced at the previous iterations. Lines 3 and 4 construct that tree $T_m(\mathbf{x})$, and find the corresponding coefficient a_m, that minimize the estimated prediction risk

(1) on the training data when $a_m T_m(\mathbf{x})$ is added to the current linear combination $F_{m-1}(\mathbf{x})$. This is then added to the current approximation $F_{m-1}(\mathbf{x})$ on line 5, producing a current predicting function $F_m(\mathbf{x})$ for the next $(m+1)$st iteration.

At the first step, $a_1 T_1(\mathbf{x})$ is just the standard tree build on the data as described in Section 3, since the current function is $F_0(\mathbf{x}) = 0$. At the next step, the estimated optimal tree $T_2(\mathbf{x})$ is found to add to it with coefficient a_2, producing the function $F_2(\mathbf{x}) = a_1 T_1(\mathbf{x}) + a_2 T_2(\mathbf{x})$. This process is continued for M steps, producing a predicting function consisting of a linear combination of M trees (line 7).

The potentially difficult part of the algorithm is constructing the optimal tree to add at each step. This will depend on the chosen loss function $L(y, F)$. For squared–error loss

$$L(y, F) = (y - F)^2$$

the procedure is especially straight forward, since

$$\begin{aligned} L(y, F_{m-1} + aT) &= (y - F_{m-1} - aT)^2 \\ &= (r_m - aT)^2. \end{aligned}$$

Here $r_m = y - F_{m-1}$ is just the error ("residual") from the current model F_{m-1} at the mth iteration. Thus each successive tree is built in the standard way to best predict the *errors* produced by the linear combination of the previous trees. This basic idea can be extended to produce boosting algorithms for any differentiable loss criterion $L(y, F)$ (Friedman 2001).

As originally proposed the standard tree construction algorithm was treated as a primitive in the boosting algorithm, inserted in lines 3 and 4 to produced a tree that best predicts the current errors $\{r_{im} = y_i - F_{m-1}(\mathbf{x}_i)\}_1^N$. In particular, an optimal tree size was estimated at each step in the standard tree building manner. This basically assumes that each tree will be the last one in the sequence. Since boosting often involves hundreds of trees, this assumption is far from true and as a result accuracy suffers. A better strategy turns out to be (Friedman 2001) to use a constant tree size (J regions) at each iteration, where the value of J is taken to be small, but not too small. Typically $4 \leq J \leq 10$ works well in the context of boosting, with performance being fairly insensitive to particular choices.

4.2.2. Regularization

Even if one restricts the size of the trees entering into a boosted tree model it is still possible to fit the training data arbitrarily well, reducing training error to zero, with a linear combination of enough trees. However, as is well known in statistics, this is seldom the best thing to do. Fitting the training data too well can increase prediction risk on future predictions. This is a phenomenon called "over–fitting". Since each tree tries to best fit the errors associated with the linear combination of previous trees, the training error monotonically decreases as more trees are included. This is, however, not the case for *future* prediction error on data not used for training.

Typically at the beginning, future prediction error decreases with increasing number of trees M until at some point M^* a minimum is reached. For $M > M^*$, future error tends to (more or less) monotonically increase as more trees are added. Thus there is an optimal number M^* of trees to include in the linear combination. This number will depend on the problem (target function (2), training sample size N, and signal to noise ratio). Thus, in any given situation, the value of M^* is unknown and must be estimated from the training data itself. This is most easily accomplished by the "early stopping" strategy used in neural network training. The training data is randomly partitioned into learning and test samples. The boosting is performed using only the data in the learning sample. As iterations proceed and trees are added, prediction risk as estimated on the test sample is monitored. At that point where a definite upward trend is detected iterations stop and M^* is estimated as the value of M producing the smallest prediction risk on the test sample.

Inhibiting the ability of a learning machine to fit the training data so as to increase future performance is called a "method–of–regularization". It can be motivated from a frequentist perspective in terms of the "bias–variance trade–off" (Geman, Bienenstock and Doursat 1992) or by the Bayesian introduction of a prior distribution over the space of solution functions. In either case, controlling the number of trees is not the only way to regularize. Another method commonly used in statistics is "shrinkage". In the context of boosting, shrinkage can be accomplished by replacing line 5 in Algorithm 1 by

$$F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + (\nu \cdot a_m)\, T_m(\mathbf{x}). \quad (30)$$

Here the contribution to the linear combination of the estimated best tree to add at each step is reduced by a factor $0 < \nu \leq 1$. This "shrinkage" factor or "learning rate" parameter controls the rate at which adding trees reduces prediction risk on the learning sample; smaller values produce a slower rate so that more trees are required to fit the learning data to the same degree.

Shrinkage (30) was introduced in Friedman 2001 and shown empirically to dramatically improve the performance of all boosting methods. Smaller learning rates were seen to produce more improvement, with a diminishing return for $\nu \lesssim 0.1$, provided that the estimated optimal number of trees $M^*(\nu)$ for that value of ν is used. This number increases with decreasing learning rate, so that the price paid for better performance is increased computation.

4.2.3. Penalized learning formulation

The introduction of shrinkage (30) in boosting was originally justified purely on empirical evidence and the reason for its success was a mystery. Recently, this mystery has been solved (Hastie, Tibshirani and Friedman 2001 and Efron, Hastie, Johnstone and Tibshirani 2002). Consider a learning machine consisting of a linear combination of *all* possible (J–region) trees:

$$\hat{F}(\mathbf{x}) = \sum \hat{a}_m T_m(\mathbf{x}) \tag{31}$$

where

$$\{\hat{a}_m\} = \arg\min_{\{a_m\}} \sum_{i=1}^{N} L\left(y_i, \sum a_m T_m(\mathbf{x}_i)\right) + \lambda \cdot P(\{a_m\}). \tag{32}$$

This is a penalized (regularized) linear regression, based on a chosen loss criterion L, of the response values $\{y_i\}_1^N$ on the predictors (trees) $\{T_m(\mathbf{x}_i)\}_{i=1}^N$. The first term in (32) is the prediction risk on the training data and the second is a penalty on the values of the coefficients $\{a_m\}$. This penalty is required to regularize the solution because the number of all possible J–region trees is infinite. The value of the "regularization" parameter λ controls the strength of the penalty. Its value is chosen to minimize an estimate of future prediction risk, for example based on a left out test sample.

A commonly used penalty for regularization in statistics is the "ridge" penalty

$$P(\{a_m\}) = \sum a_m^2 \tag{33}$$

used in ridge–regression (25) and support vector machines (22). This encourages small coefficient absolute values by penalizing the l_2–norm of the coefficient vector. Another penalty becoming increasingly popular is the "lasso" (Tibshirani 1996)

$$P(\{a_m\}) = \sum |\, a_m \,|. \tag{34}$$

This also encourages small coefficient absolute values, but by penalizing the l_1–norm. Both (33) and (34) increasingly penalize larger average absolute coefficient values. They differ in how they react to dispersion or variation of the absolute coefficient values. The ridge penalty discourages dispersion by penalizing variation in absolute values. It thus tends to produce solutions in which coefficients tend to have equal absolute values and none with the value zero. The lasso (34) is indifferent to dispersion and tends to produce solutions with a much larger variation in the absolute values of the coefficients, with many of them set to zero. The best penalty will depend on the (unknown population) optimal coefficient values. If these have more or less equal absolute values the ridge penalty (33) will produce better performance. On the other hand, if their absolute values are highly diverse, especially with a few large values and many small values, the lasso will provide higher accuracy.

As discussed in Hastie, Tibshirani and Friedman 2001 and rigorously derived in Efron *et al* 2002, there is a connection between boosting (Algorithm 1) with shrinkage (30) and penalized linear regression on all possible trees (31) (32) using the lasso penalty (34). They produce very similar solutions as the shrinkage parameter becomes arbitrarily small $\nu \to 0$. The number of trees M is inversely related to the penalty strength parameter λ; more boosted trees corresponds to smaller values of λ (less regularization). Using early stopping to estimate the optimal number M^* is equivalent to estimating the optimal value of the penalty strength parameter λ. Therefore, one can view the introduction of shrinkage (30) with a small learning rate $\nu \lesssim 0.1$ as approximating a learning machine based on all possible (J–region) trees with a lasso penalty for regularization. The lasso is especially appropriate in this context because among all possible trees only a small number will likely represent very good predictors with population optimal absolute coefficient values substantially different from zero. As noted above, this is an especially bad situation for the ridge penalty (33), but ideal for the lasso (34).

4.2.4. Boosted tree properties

Boosted trees maintain almost all of the advantages of single tree modelling described in Section 3.1 while often dramatically increasing their accuracy. One of the properties of single tree models leading to inaccuracy is the coarse piecewise constant nature of the resulting approximation. Since boosted tree machines are linear combinations of individual trees, they produce a superposition of piecewise constant approximations. These are of course also piecewise constant, but with many more pieces. The corresponding discontinuous jumps are very much smaller and they are able to more accurately approximate smooth target functions.

Boosting also dramatically reduces the instability associated with single tree models. First only small trees (Section 4.2.1) are used which are inherently more stable than the generally larger trees associated with single tree approximations. However, the big increase in stability results from the averaging process associated with using the linear combination of a large number of trees. Averaging reduces variance; that is why it plays such a fundamental role in statistical estimation.

Finally, boosting mitigates the fragmentation problem plaguing single tree models. Again only small trees are used which fragment the data to a much lesser extent than large trees. Each boosting iteration uses the entire data set to build a small tree. Each respective tree can (if dictated by the data) involve different sets of predictor variables. Thus, each prediction can be influenced by a large number of predictor variables associated with all of the trees involved in the prediction if that is estimated to produce more accurate results.

The computation associated with boosting trees roughly scales as $nN \log N$ with the number of predictor variables n and training sample size N. Thus, it can be applied to fairly large problems. For example, problems with $n \backsim 10^2$–10^3 and $N \backsim 10^5$–10^6 are routinely feasible.

The one advantage of single decision trees not inherited by boosting is interpretability. It is not possible to inspect the very large number of individual tree components in order to discern the relationships between the response y and the predictors $\mathbf{x}$. Thus, like support vector machines, boosted tree machines produce black–box models. Techniques for interpreting boosted trees as well as other black–box models are described in Friedman 2001.

4.3. Connections

The preceding section has reviewed two of the most important advances in machine learning in the recent past: support vector machines and boosted decision trees. Although motivated from very different perspectives, these two approaches share fundamental properties that may account for their respective success. These similarities are most readily apparent from their respective penalized learning formulations (Section 4.1.3 and Section 4.2.3). Both build linear models in a very high dimensional space of derived variables, each of which is a highly nonlinear function of the original predictor variables $\mathbf{x}$. For support vector machines these derived variables (14) are implicitly defined through the chosen kernel $K(\mathbf{x}, \mathbf{x}')$ defining their inner product (15). With boosted trees these derived variables are all possible (J–region) decision trees (31) (32).

The coefficients defining the respective linear models in the derived space for both methods are solutions to a penalized learning problem (22) (32) involving a loss criterion $L(y, F)$ and a penalty on the coefficients $P(\{a_m\})$. Support vector machines use $L(y, F) = (1 - y\,F)_+$ for classification $y \in \{-1, 1\}$, and (27) for regression $y \in R^1$. Boosting can be used with any (differentiable) loss criterion $L(y, F)$ (Friedman 2001). The respective penalties $P(\{a_m\})$ are (24) for SVMs and (34) with boosting. Additionally, both methods have a computational trick that allows all (implicit) calculations required to solve the learning problem in the very high (usually infinite) dimensional space of the derived variables $\mathbf{z}$ to be performed in the space of the original variables $\mathbf{x}$. For support vector machines this is the kernel trick (Section 4.1.1), whereas with boosting it is forward stagewise tree building (Algorithm 1) with shrinkage (30).

The two approaches do have some basic differences. These involve the particular derived variables defining the linear model in the high dimensional space, and the penalty $P(\{a_m\})$ on the corresponding coefficients. The performance of any linear learning machine based on derived variables (14) will depend on the detailed nature of those variables. That is, different transformations $\{h_k(\mathbf{x})\}$ will produce different learners as functions of the original variables $\mathbf{x}$, and for any given problem some will be better than others. The prediction accuracy achieved by a particular set of transformations will depend on the (unknown) target function $F^*(\mathbf{x})$ (2). With support vector machines the transformations are implicitly defined through the chosen kernel function. Thus the problem of choosing transformations becomes, as with any kernel method, one of choosing a particular kernel function $K(\mathbf{x}, \mathbf{x}')$ (“kernel customizing”).

Although motivated here for use with decision trees, boosting can in fact be implemented using any specified “base learner” $h(\mathbf{x}; \mathbf{p})$. This is a function of the predictor variables $\mathbf{x}$ characterized by a set of parameters $\mathbf{p} = \{p_{1,}p_2, \cdots\}$. A particular set of joint parameter values $\mathbf{p}$ indexes a particular function (transformation) of $\mathbf{x}$, and the set of all functions induced over all possible joint parameter values define the derived variables of the linear prediction machine in the transformed space. If all of the parameters assume values on a finite discrete set this derived space will be finite dimensional, otherwise it will have infinite dimension. When the base learner is a decision tree the parameters represent the identities of the predictor variables used for splitting, the split points, and the response values assigned to the induced regions. The forward stagewise approach can be used with any base learner by simply substituting it for the decision tree $T(\mathbf{x}) \rightarrow h(\mathbf{x}; \mathbf{p})$ in lines 3–5 of Algorithm 1. Thus boosting provides explicit control on the choice of transformations to the high dimensional space. So far boosting has seen greatest success with decision tree base learners, especially in data mining applications, owing to their advantages outlined in Section 3.1. However, boosting other base learners can provide potentially attractive alternatives in some situations.

Another difference between SVMs and boosting is the nature of the regularizing penalty $P\left(\{a_m\}\right)$ that they implicitly employ. Support vector machines use the “ridge” penalty (24). The effect of this penalty is to shrink the absolute values of the coefficients $\{\beta_m\}$ from that of the unpenalized solution $\lambda = 0$ (22), while discouraging dispersion among those absolute values. That is, it prefers solutions in which the derived variables (14) all have similar influence on the resulting linear model. Boosting implicitly uses the “lasso” penalty (34). This also shrinks the coefficient absolute values, but it is indifferent to their dispersion. It tends to produce solutions with relatively few large absolute valued coefficients and many with zero value.

If a very large number of the derived variables in the high dimensional space are all highly relevant for prediction then the ridge penalty used by SVMs will provide good results. This will be the case if the chosen kernel $K(\mathbf{x}, \mathbf{x}')$ is well matched to the unknown target function $F^*(\mathbf{x})$ (2). Kernels not well matched to the target function will (implicitly) produce transformations (14) many of which have little or no relevance to prediction. The homogenizing effect of the ridge penalty is to inflate estimates of their relevance while deflating that of the truly relevant ones, thereby reducing prediction accuracy. Thus, the sharp sensitivity of SVMs on choice of a particular kernel can be traced to the implicit use of the ridge penalty (24).

By implicitly employing the lasso penalty (34), boosting anticipates that only a small number of its derived variables are likely to be highly relevant to prediction. The regularization effect of this penalty tends to produce large coefficient absolute values for those derived variables that appear to be relevant and small (mostly zero) values for the others. This can sacrifice accuracy if the chosen base learner happens to provide an especially appropriate space of derived variables in which a large number turn out to be highly relevant. However, this approach provides considerable robustness against less than optimal choices for the base learner and thus the space of derived variables.

5. CONCLUSION

A choice between support vector machines and boosting depends on one's a priori knowledge concerning the problem at hand. If that knowledge is sufficient to lead to the construction of an especially effective kernel function $K(\mathbf{x}, \mathbf{x}')$ then an SVM (or perhaps other kernel method) would be most appropriate. If that knowledge can suggest an especially effective base learner $h(\mathbf{x}; \mathbf{p})$ then boosting would likely produce superior results. As noted above, boosting tends to be more robust to misspecification. These two techniques represent additional tools to be considered along with other machine learning methods. The best tool for any particular application depends on the detailed nature of that problem. As with any endeavor one must match the tool to the problem. If little is known about which technique might be best in any given application, several can be tried and effectiveness judged on independent data not used to construct the respective learning machines under consideration.

Acknowledgments

Helpful discussions with Trevor Hastie are gratefully acknowledged. This work was partially supported by the Department of Energy under contract DE-AC03-76SF00515, and by grant DMS–97–64431 of the National Science Foundation.

References

[1] Bellman, R. E. (1961). Adaptive Control Processes. Princeton University Press.

[2] Breiman, L. (1996). Bagging predictors. *Machine Learning* **26**, 123-140.

[3] Breiman, L. (2001). Random forests, random features. Technical Report, University of California, Berkeley.

[4] Breiman, L., Friedman, J. H., Olshen, R. and Stone, C. (1983). *Classification and Regression Trees.* Wadsworth.

[5] Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2002). Least angle regression. Annals of Statistics. To appear.

[6] Freund, Y and Schapire, R. (1996). Experiments with a new boosting algorithm. In *Machine Learning: Proceedings of the Thirteenth International Conference*, 148–156.

[7] Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics* **29**, 1189-1232.

[8] Geman, S., Bienenstock, E. and Doursat, R. (1992). Neural networks and the bias/variance dilemma. Neural Computation **4**, 1-58.

[9] Hastie, T., Tibshirani, R. and Friedman, J.H. (2001). *The Elements of Statistical Learning.* Springer–Verlag.

[10] Hoerl, A. E. and Kennard, R. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* **12**, 55-67

[11] Nadaraya, E. A. (1964). On estimating regression. *Theory Prob. Appl.* 10, 186-190.

[12] Quinlan, R. (1992). *C4.5: Programs for machine learning.* Morgan Kaufmann, San Mateo.

[13] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Royal Statist. Soc.* **58**, 267-288.

[14] Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory.* Springer.

[15] Wahba, G. (1990). *Spline Models for Observational Data.* SIAM, Philadelphia.

[16] Watson, G. S. (1964). Smooth regression analysis. *Sankhya Ser. A.* **26**, 359-372.

A Multivariate Method for Comparing N-dimensional Distributions

James D. Loudin and Hannu E. Miettinen
Rice University, Houston, TX 77005, U.S.A.

We propose a new multivariate method for comparing two N-dimensional distributions. We first use kernel estimation to construct probability densities for the two data sets, and then define two discriminant functions, one appropriate for the null hypothesis and another appropriate for the actual data. Distributions of the two discriminant functions at random test points are then compared using the one-dimensional K-S test. The performance of the method is illustrated with Monte Carlo data.

1. INTRODUCTION

A comparison of two distributions is often a crucial part of data analysis. We may want to know whether a measured distribution is consistent with some hypothesis, or we want to compare "signal" and "background" distributions to see if they are different. In case of one-dimensional distributions the Kolmogorov-Smirnov (K-S) test [1] provides a tried-and-true method for such a comparison. The test uses the maximum distance d between the cumulative distribution functions of two histograms or probability densities as a measure of their similarity. The K-S test is non-parametric and independent of the shapes of the underlying distributions. However, the K-S test does not generalize naturally to higher dimensions, and there is no widely accepted test for comparing N-dimensional distributions.

In this note we propose a method which combines the information contained in N-dimensional distributions with the simplicity of the K-S test in one dimension. We first construct the relevant N-dimensional probability densities by using kernel estimation. We then define two discriminant functions, one for the null hypothesis and another for the actual data, and their distributions at randomly selected test points are compared using the standard K-S test. The method is mathematically uncomplicated and it appears to work well, based on test results in two dimensions.

2. THE METHOD

Consider two data sets A and B containing n_A and n_B data points (or "events"), respectively. Let each set be described by N variables x_i which are combined into a feature vector $\mathbf{x} = (x_1, \ldots, x_N)$. We assume that the data have been binned in each dimension using some reasonable criteria so that N-dimensional scatter plots for both sets are available.

Our method, sketched in Figure 1, consists of three steps:

1. We first construct probability densities $f_A(\mathbf{x})$ and $f_B(\mathbf{x})$ using kernel estimation. In this study we have used the PDE method [2], whereby the densities are estimated by adding up N-dimensional Gaussians placed at data points. The width of the Gaussian in each dimension, h_i, is proportional to the standard deviation σ_i of the i^{th} variable: $h_i = h \cdot \sigma_i$. Here h is a global smoothing parameter which sets the overall scale for the widths of the Gaussians. The value of h is optimized by minimizing the error $\int [d - f(\mathbf{x})]^2$, where d is the number of data points in a given bin and f is the corresponding density estimate. Figure 2 illustrates the density estimation and the optimization of h in the case when the feature space is two-dimensional.

2. We define a discriminant function D associated with the data as

$$D(\mathbf{x}) = \frac{f_A(\mathbf{x})}{f_A(\mathbf{x}) + f_B(\mathbf{x})} \tag{1}$$

   Possible values of D are obviously in the range $0 \leq D \leq 1$. We similarly define another discriminant function D^* by replacing f_B with f_A^* which is obtained by generating n_B random data points distributed according to f_A and then constructing the density as described above. Thus D^* is a discriminant function associated with the null hypothesis that A and B come from the same underlying density. In order to reduce statistical fluctuations we actually form D^* a number of times and use their average $\langle D^* \rangle$ as the discriminant function appropriate for the null hypothesis. Distributions of D, D^*, and $\langle D^* \rangle$ are obtained by evaluating each function at randomly generated test points pulled from the density f_A. If A and B are similar, the distribution of D should peak near $D = 0.5$. The distribution of $\langle D^* \rangle$ is independent of B and should ideally be narrow and peaked at $\langle D^* \rangle = 0.5$.

3. We compute the cumulative distribution function F given by

$$F(x) = \int_0^x f_D(t)\, dt \tag{2}$$

   where f_D is the probability density for D, obtained from the distribution of D. Similar cumulative distribution functions F^* and $\langle F^* \rangle$

are computed from the distributions of D^* and $\langle D^* \rangle$. The K-S distance d between F and $\langle F^* \rangle$ is a measure of similarity between sets A and B. The distribution of K-S distances d^* between F^* and $\langle F^* \rangle$ tells us what to expect for the null hypothesis. We can then compute the significance function $S(x)$, defined as

$$S(x) = \int_x^1 f_{d^*}(t)\,dt \qquad (3)$$

where f_{d^*} is the density function for d^*. If we obtain a K-S distance d for the data, then $S(d)$ is the probability that the K-S distance would exceed d under the null hypothesis, i.e. the probability that purely random fluctuations could produce the observed value.

3. RESULTS

We generated the data sets A and B from two-dimensional Gaussian densities of equal widths in the two dimensions. We chose $n_A = 1000$ and $n_B = 50$, motivated by a "typical" analysis situation where we might have $\sim$ 50 interesting data events to be compared with a larger control sample of e.g. Monte Carlo events. The relevant discriminant functions were constructed as described above, and they were evaluated at 1000 random test points to get the associated distributions. Figure 3 shows the distribution of $\langle D^* \rangle$, averaged over 500 distributions of D^*. The distribution is narrow and peaked at $\langle D^* \rangle \simeq 0.5$ as it should be.

We have studied the shapes of the D^* distributions as a function of the parameter h which determines the widths of the Gaussians used in kernel estimation. We find that if h is too small, there will be "valleys" in the density f_A^* due to the small sample size, and these valleys give rise to values of D^* near 1. If h is too large, all densities are too flat, and values of D^* much above 0.5 cannot occur. It is therefore important to optimize the value of h fairly carefully, otherwise the distribution of $\langle D^* \rangle$ will be asymmetric.

Figure 4 shows the cumulative distribution function $\langle F^* \rangle$ associated with $\langle D^* \rangle$. We also show F for a data set where the widths of the Gaussians for sets A and B are equal. The K-S distance d is indicated in the figure.

Figure 5a shows the (normalized) distribution of d^* for the null hypothesis. The mean value is $\langle d^* \rangle \simeq 0.10$, and there are no distances beyond $d^* \simeq 0.2$. The associated significance curve is shown in Figure 5b. We have verified that the shape of the significance curve for the null hypothesis is nearly independent of the widths of the Gaussians used in kernel estimation.

In order to test the "resolving power" of the method we have applied it to data sets A and B (1000 and 50 events, respectively) pulled from two-dimensional Gaussian densities whose widths differ by 10%, 20%, and 50%. In each case the B set was generated 200 times. The distributions of the K-S distance d are shown in Figure 6.

We find that in the 10% case the mean value is $\langle d \rangle \simeq 0.12$, indicating that there is a $\sim$ 30% probability that the *average* d would exceed this value for two identical densities. Thus we cannot distinguish between the two data sets. In the 20% case we find $\langle d \rangle \simeq 0.15$, and the null hypothesis probability for the average d has dropped to $\sim$ 10%. However, nearly half the time d is below 0.15, meaning that such values *could* arise from random fluctuations, and one quarter of the time d is below 0.10, meaning that such values are *likely* to arise from random fluctuations. In the 50% case it is easy to distinguish between A and B most of the time, but there is still a 15-20% chance that random fluctuations could explain the observed value of d.

4. CONCLUSIONS

We have outlined a method for comparing N-dimensional distributions which combines a multivariate approach with the standard K-S test. The method provides a precise way of quantifying the degree of similarity between two distributions. We have tested the method in two dimensions by comparing two Gaussian distributions of different widths, and find that the method performs well even when one of the data sets is relatively small.

Acknowledgments

We thank Sang-Joon Lee for discussions and for a critical reading of the manuscript.

References

[1] F.J. Massey, J. Amer. Stat. Assoc. **46**, 68-78 (1951).

[2] L. Holmström, S.R. Sain, and H.E. Miettinen, Comp. Phys. Communications **88** (1995) 195-210.

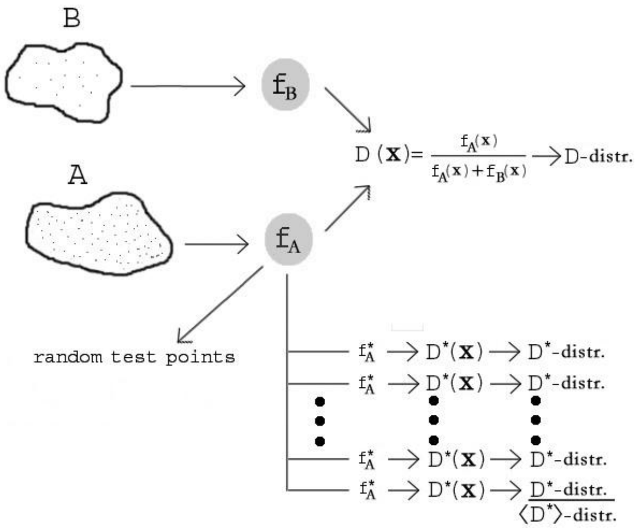

B
f_B
A
f_A
$D(\mathbf{x}) = \frac{f_A(\mathbf{x})}{f_A(\mathbf{x}) + f_B(\mathbf{x})} \rightarrow$ D-distr.
random test points
$f_A^* \rightarrow D^*(\mathbf{X}) \rightarrow D^*$-distr.
$f_A^* \rightarrow D^*(\mathbf{X}) \rightarrow D^*$-distr.
$f_A^* \rightarrow D^*(\mathbf{X}) \rightarrow D^*$-distr.
$f_A^* \rightarrow D^*(\mathbf{X}) \rightarrow D^*$-distr.
$\langle D^* \rangle$-distr.

Figure 1: Sketch of the analysis method.

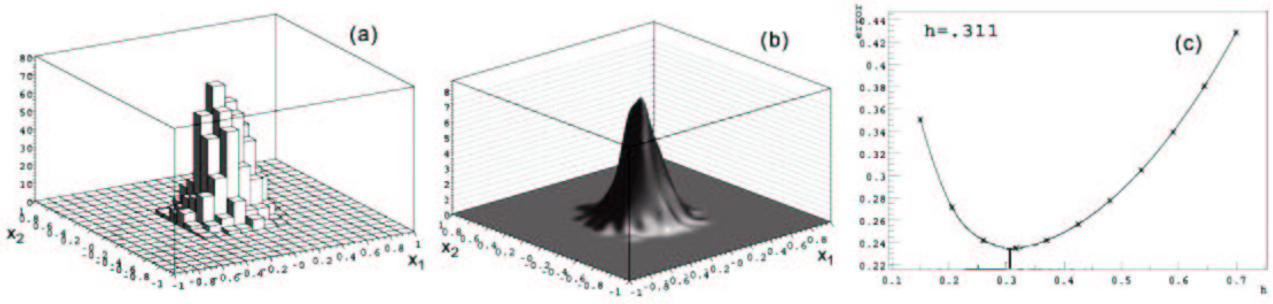


Figure 2: (a) Lego plot of a two-dimensional Gaussian. (b) The corresponding density estimate f. (c) Optimization curve for h.

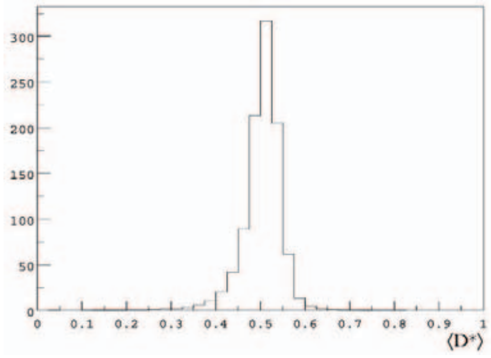


Figure 3: Distribution of $\langle D^* \rangle$, the discriminant function for the null hypothesis.

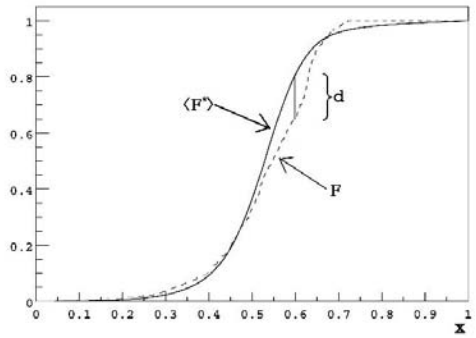


Figure 4: Cumulative distribution functions $\langle F^* \rangle$ for the null hypothesis (solid curve) and F for the data (dashed curve). The K-S distance d is also shown.

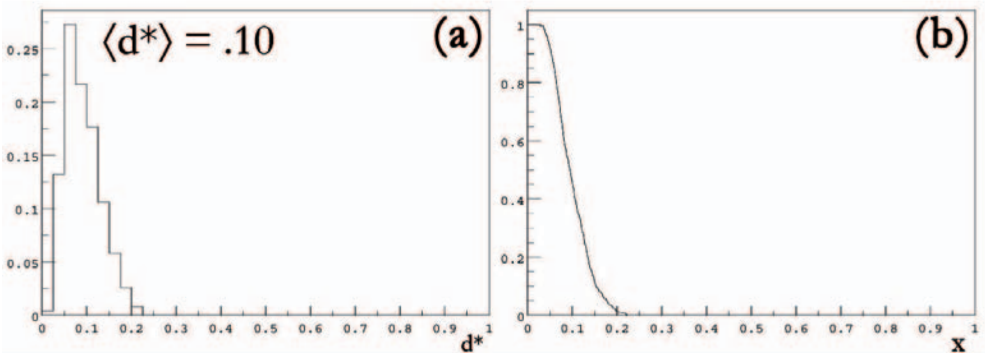


Figure 5: (a) Distribution of the K-S distance d^* for the null hypothesis. (b) The associated significance curve.

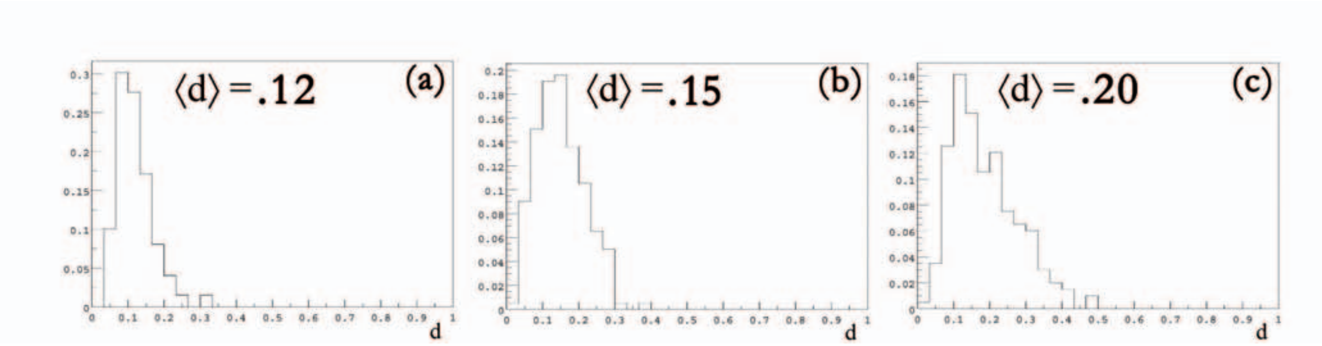


Figure 6: The distributions of the K-S distance d when the widths of the Gaussian densities for data sets A and B differ by (a) 10% (b) 20% (c) 50%.

Multivariate Analysis from a Statistical Point of View

K.S. Cranmer
University of Wisconsin-Madison, Madison, WI 53706, USA

Multivariate Analysis is an increasingly common tool in experimental high energy physics; however, many of the common approaches were borrowed from other fields. We clarify what the goal of a multivariate algorithm should be for the search for a new particle and compare different approaches. We also translate the Neyman-Pearson theory into the language of statistical learning theory.

1. INTRODUCTION

Multivariate Analysis is an increasingly common tool in experimental high energy physics; however, most of the common approaches were borrowed from other fields. Each of these algorithms were developed for their own particular task, thus they look quite different at their core. It is not obvious that what these different algorithms do internally is optimal for the the tasks which they perform within high energy physics. It is also quite difficult to compare these different algorithms due to the differences in the formalisms that were used to derive and/or document them. In Section 2 we introduce a formalism for a *Learning Machine*, which is general enough to encompass all of the techniques used within high energy physics. In Sections 3 & 4 we review the statistical statements relevant to new particle searches and translate them into the formalism of statistical learning theory. In the remainder of the note, we look at the main results of statistical learning theory and their relevance to some of the common algorithms used within high energy physics.

2. FORMALISM

Formally a Learning Machine is a family of functions $\mathcal{F}$ with domain I and range O parametrized by $\alpha \in \Lambda$. The domain can usually be thought of as, or at least embedded in, $\mathbb{R}^d$ and we generically denote points in the domain as x. The points x can be referred to in many ways (*e.g.* patterns, events, inputs, examples, . . .). The range is most commonly $\mathbb{R}$, $[0, 1]$, or just $\{0, 1\}$. Elements of the range are denoted by y and can be referred to in many ways (*e.g.* classes, target values, outputs, . . .). The parameters α specify a particular function $f_\alpha \in \mathcal{F}$ and the structure of $\alpha \in \Lambda$ depends upon the learning machine [1, 2].

In the modern theory of machine learning, the performance of a learning machine is usually cast in the more pessimistic setting of *risk*. In general, the risk, R, of a learning machine is written as

$$R(\alpha) = \int Q(x, y; \alpha)\; p(x, y) dx dy \qquad (1)$$

where Q measures some notion of *loss* between $f_\alpha(x)$ and the target value y. For example, when classifying events, the risk of mis-classification is given by Eq. 1 with $Q(x, y; \alpha) = |y - f_\alpha(x)|$. Similarly, for regression[1] tasks one takes $Q(x, y; \alpha) = (y - f_\alpha(x))^2$. Most of the classic applications of learning machines can be cast into this formalism; however, searches for new particles place some strain on the notion of risk.

3. SEARCHES FOR NEW PARTICLES

The conclusion of an experimental search for a new particle is a statistical statement – usually a declaration of discovery or a limit on the mass of the hypothetical particle. Thus, the appropriate notion of performance for a multivariate algorithm used in a search for a new particle is that performance measure which will maximize the chance of declaring a discovery or provide the tightest limits on the hypothetical particle. In principle, it should be a fairly straight-forward procedure to use the formal statistical statements to derive the most appropriate performance measure. This procedure is complicated by the fact that experimentalists (and statisticians) cannot settle on a formalism to use (*i.e.* Bayesians *vs.* Frequentists). As an example, let us consider the Frequentist theory developed by Neyman and Pearson [3]. This was the basis for the results of the search for the Standard Model Higgs boson at LEP [4].

The Neyman-Pearson theory (which we review briefly for completeness) begins with two Hypotheses: the null hypothesis H_0 and the alternate hypothesis H_1 [3]. In the case of a new particle search H_0 is identified with the currently accepted theory (*i.e.* the Standard Model) and is usually referred to as the "background-only" hypothesis. Similarly, H_1 is identified with the theory being tested usually referred to as the "signal-plus-background" hypothesis

[1] During the presentation, J. Friedman did not distinguish between these two tasks; however, in a region with $p(x, 1) = b$ and $p(x, 0) = 1 - b$, the optimal $f(x)$ for classification and regression differ. For classification, $f(x) = \{1 \text{ if } b > 1/2, \text{else } 0\}$, and for regression the optimal $f(x) = b$.

Next, one defines a region $W \in I$ such that if the data fall in W we accept the null hypothesis (and reject the alternate hypothesis)[2]. Similarly, if the data fall in $I - W$ we reject the null hypothesis and accept the alternate hypothesis. The probability to commit a Type I error is called the *size* of the test and is given by (note alternate use of α)

$$\alpha = \int_{I-W} p(x|H_0)dx. \tag{2}$$

The probability to commit a Type II error is given by

$$\beta = \int_{W} p(x|H_1)dx. \tag{3}$$

Finally, the Neyman-Pearson lemma tells us that the region W of size α which minimizes the rate of Type II error (maximizes the power) is given by

$$W = \left\{ x \;\middle|\; \frac{p(x|H_1)}{p(x|H_0)} > k_\alpha \right\}. \tag{4}$$

4. THE NEYMAN-PEARSON THEORY IN THE CONTEXT OF RISK

In Section 1 we provided the loss functional Q appropriate for the classification and regression tasks; however, we did not provide a loss functional for searches for new particles. Having chosen the Neyman-Pearson theory as an explicit example, it is possible to develop a formal notion of risk.

Once the size of the test, α, has been agreed upon, the notion of risk is the probability of Type II error β. In order to return to the formalism outlined in Section 2, identify H_1 with $y = 1$ and H_0 with $y = 0$. Let us consider learning machines that have a range $\mathbb{R}$ which we will compose with a step function $\tilde{f}(x) = \Theta(f_\alpha(x) - k_\alpha)$ so that by adjusting k_α we insure that the acceptance region W has the appropriate size. The region W is the acceptance region for H_0, thus it corresponds to $W = \{x|\tilde{f}(x) = 0\}$ and $I - W = \{x|\tilde{f}(x) = 1\}$. We can also translate the quantities $p(x|H_0)$ and $p(x|H_1)$ into their learning-theory equivalents $p(x|0) = p(x,0)/p(0) = \delta(y)p(x,y)/\int p(x,0)dx$ and $\delta(1-y)p(x,y)/\int p(x,1)dx$, respectively. With these substitutions we can rewrite the Neyman-Pearson theory as follows. A fixed size gives us the global constraint

$$\alpha = \frac{\int \Theta(f_\alpha(x) - k_\alpha)\ \delta(y)\ p(x,y))dxdy}{\int p(x,0)dx} \tag{5}$$

and the risk is given by

$$\begin{aligned} \beta &= \frac{\int [1 - \Theta(f_\alpha(x) - k_\alpha)]\ p(x,1)dx}{\int p(x,1)dx} \\ &\propto \int \Theta(f_\alpha(x) + k_\alpha)\ \delta(1-y)\ p(x,y)dxdy. \end{aligned} \tag{6}$$

Extracting the integrand we can write the loss functional as

$$Q(x,y;\alpha) = \Theta(f_\alpha(x) + k_\alpha)\ \delta(1-y). \tag{7}$$

Unfortunately, Eq. 1 does not allow for the global constraint imposed by k_α (which is implicitly a functional of f_α), but this could be accommodated by the methods of Euler and Lagrange. Furthermore, the constraint cannot be evaluated without explicit knowledge of $p(x,y)$.

4.1. Asymptotic Equivalence

Certain approaches to multivariate analysis leverage the many powerful theorems of statistics assuming one can explicitly refer to $p(x,y)$. This dependence places a great deal of stress on the asymptotic ability to estimate $p(x,y)$ from a finite set of samples $\{(x,y)_i\}$. There are many such techniques for estimating a multivariate density function $p(x,y)$ given the samples [5, 6]. Unfortunately, for high dimensional domains, the number of samples needed to enjoy the asymptotic properties grows very rapidly; this is known as the *curse of dimensionality.*

In the case that there is no (or negligible) interference between the signal process and the background processes one can avoid the complications imposed by quantum mechanics and simply add probabilities. This is often the case with searches for new particles, thus the signal-plus-background hypothesis can be rewritten $p(x,|H_1) = n_s p_s(x) + n_b p_b(x)$, where n_s and n_b are normalization constants that sum to unity. This allows us to rewrite the contours of the likelihood ratio as contours of the signal-to-background ratio. In particular the contours of the likelihood ratio $p(x|H_1)/p(x|H_0) = k_\alpha$ can be rewritten as $p_s(x)/p_b(x) = (k_\alpha - n_b)/n_s = k'_\alpha$.

The kernel estimation techniques described in this conference represent a particular statistical approach in which classification is achieved by cutting on a discriminant function $D(x)$ [7]. The discriminant function $D(x) = p_s(x)/(p_s(x) + p_b(x))$ is one-to-one with $p_s(x)/p_b(x)$ (which is in turn one-to-one with the likelihood ratio). These correspondences are only valid asymptotically, and the ability to accurately approximate $p(x)$ from an empirical sample is often far from ideal. However, for particle physics applications, up to 5-dimensional multivariate analyses have shown good performance [8]. Furthermore, they have the added benefit that they can be easily understood

[2] With m measurements, we should actually consider the data as $(x_1, \ldots, x_m) \in I^m$, but, for ease of notation, let us only consider $m = 1$.

4.2. Direct vs. Indirect Methods

The loss functional defined in Eq. 7 is derived from a minimization on the rate of Type II error. This is logically distinct from, but asymptotically equivalent to, approximating the likelihood ratio. In the case of no interference, this is logically distinct from, but asymptotically equivalent to, approximating the signal-to-background ratio. In fact, most multivariate algorithms are concerned with approximating an auxiliary function that is one-to-one with the likelihood ratio. Because the methods are not directly concerned with minimizing the rate of Type II error, they should be considered *indirect methods*. Furthermore, the asymptotic equivalence breaks down in most applications, and the indirect methods are no longer optimal. Neural networks, kernel estimation techniques, and support vector machines all represent indirect solutions to the search for new particles. The Genetic Programming (GP) approach presented in Section 6 is a *direct method* concerned with optimizing a user-defined performance measure.

5. STATISTICAL LEARNING THEORY

The starting point for statistical learning theory is to accept that we might not know $p(x, y)$ in any analytic or numerical form. This is, indeed, the case for particle physics, because only $\{(x, y)_i\}$ can be obtained from the Monte Carlo convolution of a well-known theoretical prediction and complex numerical description of the detector. In this case, the learning problem is based entirely on the training samples $\{(x, y)_i\}$ with l elements. The risk functional is thus replaced by the *empirical risk functional*

$$R_{\text{emp}}(\alpha) = \frac{1}{l}\sum_{i=1}^{l} Q(x_i, y_i; \alpha). \tag{8}$$

There is a surprising result that the true risk $R(\alpha)$ can be bounded independent of the distribution $p(x, y)$. In particular, for $0 \le Q(x, y; \alpha) \le 1$

$$R(\alpha) \le R_{\text{emp}}(\alpha) + \sqrt{\left(\frac{h(\log(2l/h) + 1) - \log(\eta/4)}{l}\right)}, \tag{9}$$

where h is the Vapnik-Chervonenkis (VC) dimension and η is the probability that the bound is violated. As $\eta \to 0$, $h \to \infty$, or $l \to 0$ the bound becomes trivial.

The VC dimension is a fundamental property of a learning machine $\mathcal{F}$, and is defined as the maximal cardinality of a set which can be shattered by $\mathcal{F}$. "A set $\{x_i\}$ can be shattered by $\mathcal{F}$" means that for each of the 2^h binary classifications of the points $\{x_i\}$, there exists a $f_\alpha \in \mathcal{F}$ which satisfies $y_i = f_\alpha(x_i)$. A set of three points can be shattered by an oriented line as illustrated in Figure 1. Note that for a learning machine with VC dimension h, not every set of h elements must be shattered by $\mathcal{F}$, but at least one.

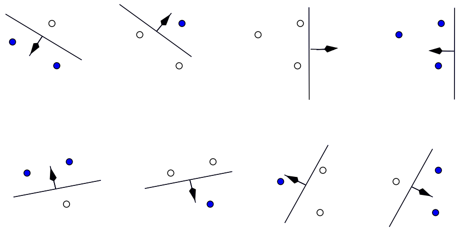


Figure 1: Example of an oriented line shattering 3 points. Solid and empty dots represent the two classes for y and each of the 2^3 permutations are shown.

Eq. 9 is a remarkable result which relates the number of training examples l, a fundamental property of the learning machine h, and the risk R independent of the unknown distribution $p(x, y)$. The bounds provided by Eq. 9 are relatively weak due to their stunning generality.

It is important to realize that with an independent testing sample one can evaluate the true risk arbitrarily well. This testing sample, by definition, is not known to the algorithm, so the bound is useful for the design of algorithms through *structural risk minimization*. However, neural networks and most other methods rely on an independent testing sample to aid in their design and validation. An independent testing sample is clearly a better way to assess the true risk of a multivariate algorithm; however, Eq. 9 does shed light on the issues of overtraining, suggests the number of training samples that are needed, and offers a tool to compare different algorithms.

5.1. VC Dimension of Neural Networks

In order to apply Eq. 9, one must determine the VC dimension of neural networks. This is a difficult problem in combinatorics and geometry aided by algebraic techniques. Eduardo Sontag has an excellent review of these techniques and shows that the VC dimension of neural networks can, thus far, only be bounded fairly weakly [9]. In particular, if we define ρ as the number of weights and biases in the network, then the best bounds are $\rho^2 < h < \rho^4$. In a typical particle physics neural network one can expect $100 < \rho < 1000$, which translates into a VC dimension as high as 10^{12}, which implies $l > 10^{13}$ for reasonable bounds on the risk. These bounds imply enormous numbers of training samples when compared to a typical training sample of 10^5. Sontag goes on to show that these shattered sets are incredibly special and that the set of all shattered sets of cardinality $\mu > 2\rho + 1$ is measure zero in general. Thus, perhaps a more relevant notion of the VC dimension of a neural network is given by μ.

6. GENETIC PROGRAMMING AND ALGORITHMS

Genetic Programming (GP) and Genetic Algorithms (GA) are based on a similar evolutionary metaphor in which “individuals” (potential solutions to the problem at hand) compete with respect to a user-defined performance measure. For new particle searches, the rate of Type II error, the significance, the exclusion potential, or G. Punzi’s suggestion [10] are all reasonable performance measures. Ideally, one would use as a performance measure the same procedure that will be used to quote the results of the experiment. For instance, there is no reason (other than speed) that one could not include discriminating variables and systematic error in the optimization procedure (in fact, the author has done both).

The use of GP for the classification is fairly limited; however, it can be traced to the early works on the subject by Koza [11]. To the best of the author’s knowledge, the first application of GP within particle physics will appear in [12]. The difference between the algorithms is that GAs evolve a bit string which typically encodes parameters to a pre-existing program, function, or class of cuts, while GP directly evolves the programs or functions. For example, Field and Kanev [13] used Genetic Algorithms to optimize the lower- and upper-bounds for six 1-dimensional cuts on Modified Fox-Wolfram “shape” variables. In that case, the phase-space region was a pre-defined 6-cube and the GA was simply evolving the parameters for the upper- and lower-bounds. On the other hand, GP algorithm is not constrained to a pre-defined shape or parametric form. Instead, the GP approach is concerned directly with the construction of an optimal, non-trivial phase space region (*i.e.* an acceptance region W) with respect to a user-defined performance measure. GPs which only produce polynomial expressions form a vector space, which allows for a quick approximation of their VC dimension [9].

7. CONCLUSIONS

Clearly multivariate algorithms will have an increasingly important role in high energy physics, which necessitates that the field develop a coherent formalism and carefully consider what it means for a method to be optimal. Statistical learning theory offers a formalism that is general enough to describe all of the common multivariate analysis techniques, and provides interesting results relating risk, the number of training samples, and the learning capacity of the algorithm. However, independent testing samples and the global constraint on the rate of Type I error places some strain on the risk formalism. Finally, when one takes into account limited training data and systematic errors it is not clear that indirect methods are truly optimizing an experiments sensitivity. Direct methods, such as Genetic Programming, force analysts to be more clear about what statistical statements they plan to make and remove an artificial boundary between the goals of the experiment and the optimization procedures of the algorithm.

Acknowledgments

This work was supported by a graduate research fellowship from the National Science Foundation and US Department of Energy Grant DE-FG0295-ER40896.

References

[1] V. Vapnik and A.J. Cervonenkis. The uniform convergence of frequencies of the appearance of events to their probabilities. *Dokl. Akad. Nauk SSSR*, 1968. in Russian.

[2] V. Vapnik. *The Nature of Statistical Learning Theory.* Springer, New York, 2nd edition, 2000.

[3] J.K Stuart, A. Ord and S. Arnold. *Kendall’s Advanced Theory of Statistics, Vol 2A (6th Ed.).* Oxford University Press, New York, 1994.

[4] Search for the standard model Higgs boson at LEP. *Phys. Lett.*, B565:61–75, 2003.

[5] D. Scott. *Multivariate Density Estimation: Theory, Practice, and Visualization.* John Wiley and Sons Inc., 1992.

[6] K. Cranmer. Kernel estimation in high-energy physics. *Comput. Phys. Commun.*, 136:198–207, 2001.

[7] A. Askew. Event selection with adaptive gaussian kernels. In *PhyStat2003*, 2003.

[8] L. Hölmstrom *et. al.* A new multivariate technique for top quark search. *Comput. Phys. Commun.*, 88:195–210, 1995.

[9] E. Sontag. VC dimension of neural networks. In C.M. Bishop, editor, *Neural Networks and Machine Learning*, pages 69–95, Berlin, 1998. Springer-Verlag.

[10] G. Punzi. Sensitivity of searches for new signals and its optimization. In *PhyStat2003*, 2003.

[11] J.R. Koza. *Genetic Programming: On the Programming of Computers by Means of Natural Selection.* MIT Press, Cambridge, MA, 1992.

[12] K. Cranmer and R.S. Bowman. PhysicsGP: A genetic programming approach to event selection. *submitted to Comput. Phys. Commun.*

[13] R. D. Field and Y. A. Kanev. Using collider event topology in the search for the six-jet decay of top quark antiquark pairs. *hep-ph/9801318*, 1997.

Event Selection Using an Extended Fisher Discriminant Method

Byron P. Roe
University of Michigan, Ann Arbor, Michigan 48109, USA

This note discusses the problem of choosing between hypotheses in a situation with many, correlated non-normal variables. A new method is introduced to shrink the many variables into a smaller subset of variables with zero mean, unit variance, and zero correlation coefficient between variables. These new variables are well suited to use in a neural net.

1. INTRODUCTION

At the Durham Statistics in Physics Conference (2002), S. Towers[1] noted some of the problems that occur when one uses many, correlated variables in a multivariate analysis and proposed a heuristic method to shrink the number. In this note a semi-automatic method is suggested help with this problem. The MiniBooNE experiment is faced with just such a problem, distinguishing ν_e events from background events given a large number of variables obtained from the event reconstructions.

2. FISHER DISCRIMINANT METHOD AND ITS EXTENSION

The Fisher discriminant method is a standard method for obtaining a single variable to distinguish hypotheses starting from a large number of variables. If the initial variables come from a multi-normal distribution, the Fisher variable encapsulates all of the discrimination information. However, in many problems the variables are not of this form and the Fisher variable, although useful, is not sufficient.

The Fisher discriminant method [2] finds the linear combination y of the initial variables x which maximizes

$$(\overline{y}_{\rm sig} - \overline{y}_{\rm bkg})^2/[{\rm var}_{\rm sig} + {\rm var}_{\rm bkg}],$$

where $\overline{y}$ is the mean value of the variable and var is the variance. If S is the correlation matrix for the original variables corresponding to the denominator, then the inverse of S dotted into $(\overline{x}_{\rm sig} - \overline{x}_{\rm bkg})$, gives the combination which maximizes the preceding expression.

If the distribution is not multi-normal, there is information still to be obtained after finding the Fisher variable. It is then useful to apply this method successively, firstly to the original variables and, afterwards, to several non-linear transformations of the variables. Presently, three transformations are chosen: the logarithms of the original variables, the exponentials of the original variables and the cube of the original variables. Together with the original variables, this is then four choices. The present note describes a work in progress. It is highly likely that these are not optimum and that better choices can and will be found. Indeed, it is likely that the optimum choice depends on the problem.

The method is also used with the original denominator, $[{\rm var}_{\rm sig} + {\rm var}_{\rm bkg}]$ and then with each individual variance in turn. This is done since some of the variables may be quite narrow for signal and wide for background or vice versa. (See Figure 1) There are then $4 \times 3 = 12$ variables obtained with this method.

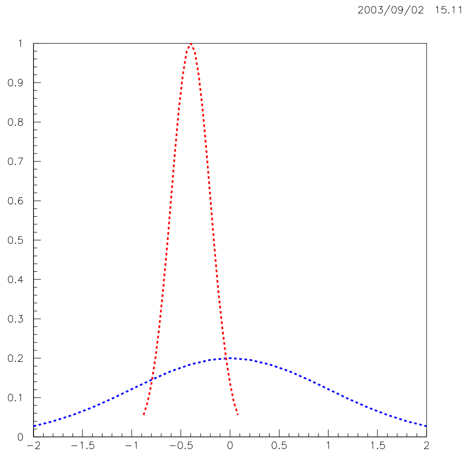


Figure 1: Different Width Normal Distributions

The procedure follows the following steps:

1. Start with equal Monte Carlo samples of signal and background events
2. Multiply and translate each variable to have an overall mean of zero and unit variance. It is useful to fold variables if necessary to maximize the difference in means. A few events, very far out on the tails of the distribution are clipped ($x > 6\sigma$).
3. Order the variables according to $|\overline{x}_{\rm sig} - \overline{x}_{\rm bkg}|$ divided by the smallest of the signal and background variances. At present, the ordering of variables is done only once.

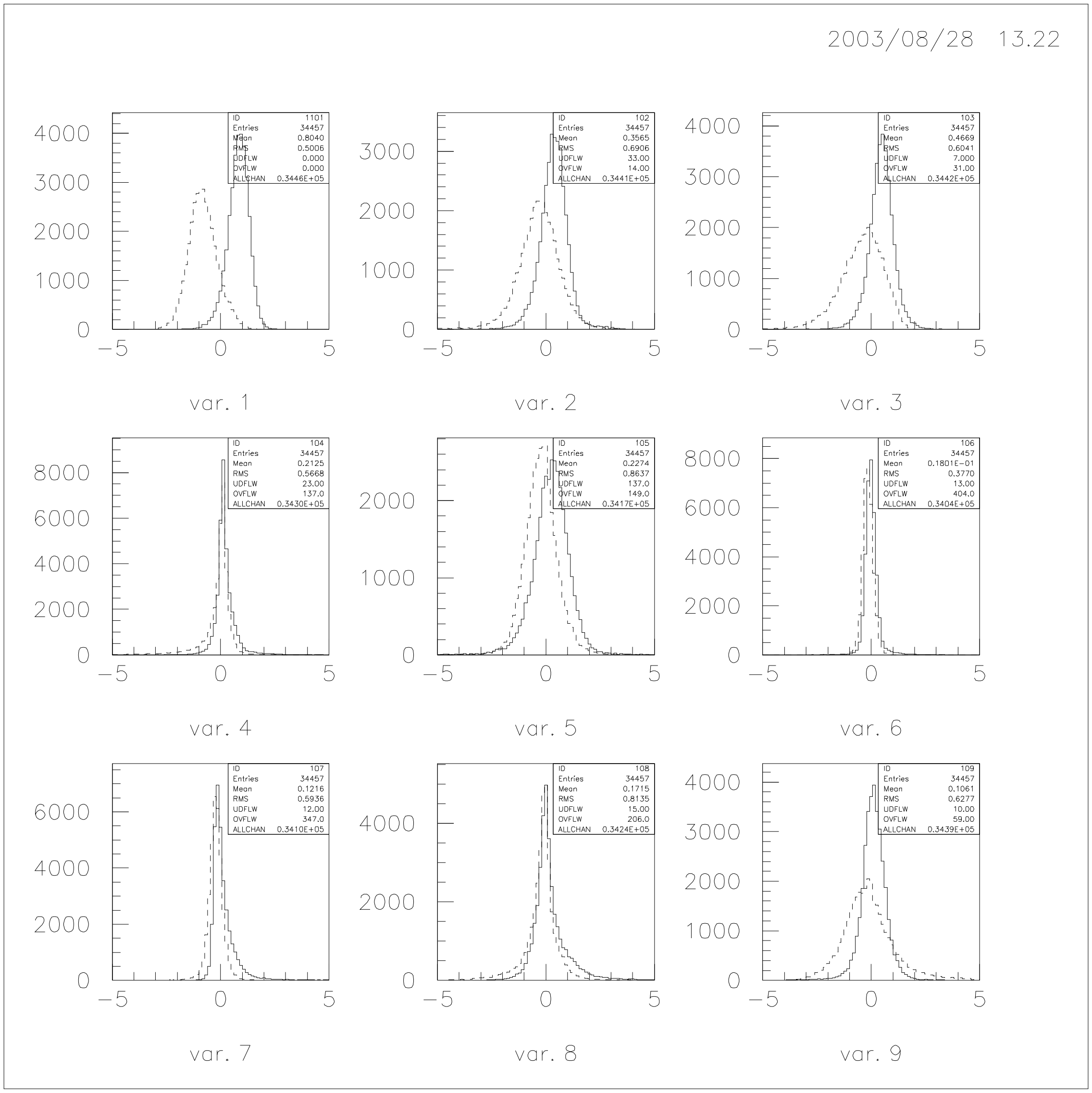


Figure 2: Plots of the first nine of the variables obtained. The solid lines are the ν_e signal and the dashed lines are the π^0 background.

4. Apply this extended Fisher method to the appropriate transformation of the variables.

5. Use the Gram-Schmidt procedure to make the other variables have zero correlation coefficient with the chosen linear combination.

6. The new variable is a linear combination of the original n variables. One variable must be discarded to have an independent set. Discard the least significant (by the criterion of step 3) of the original variables. Using the $n-1$ non-Fisher variables, go back to step 2 to get the next variable.

For MiniBooNE the roefitter reconstruction started with 49 particle identification variables. Using the steps outlined, these were reduced to 12 variables. When ν_e quasi-elastic events were compared with background neutral current π^0 events, the use of this procedure with a neural net kept 46% of the ν_e and reduced the π^0/ν_e ratio to 1.1% of its original value. The neural net was not hard to tune. The reconstruction–particle identification package is still being improved, so these numbers will improve further. The results obtained here are similar to those obtained using a more elaborate neural net on a subsample of 26 of the original 49 variables.

These 12 variables have zero correlation coefficients. Use of the neural net is simplified and it is convenient to look at the effect of cuts using these variables.

Plots of the first nine of the twelve variables are shown in Figure 2.

References

[1] S. Tower, *Benefits of Minimizing the Number of Discriminators Used in a Multivariate Analysis*, Durham Conference on Statistics (2002).

[2] Glen Cowan *Statistical Data Analysis*, Clarendon Press, Oxford (1998).

Optimizing the Limit Setting Potential of a Multivariate Analysis Using the Bayes Posterior Ratio

G.C. Hill[1], F. Lu[2], P. Desiati[1], G. Wahba[2]
1. Department of Physics, 2. Department of Statistics,
University of Wisconsin, Madison, 53706, USA

In this work we consider the problem of optimal cut selection in a multivariate analysis. When we wish to place an upper limit on the normalisation of a theoretical flux model, we show how the best detector sensitivity is found by optimizing the ratio of the average upper limit to the expected signal. In a multidimensional observable space, we find the constant Bayes posterior surface that defines an acceptance region of events yielding the best limit setting power. The calculation of the posterior using a penalized likelihood method is described.

1. INTRODUCTION

In this paper, we consider the problem of optimizing the limit setting power of a multivariate analysis. A limit optimisation technique, the *model rejection potential* [1] technique, has been proposed and used in various analyses (see e.g. ref. [2]). The best limiting power is found by minimizing the *model rejection factor*, the ratio of the average upper limit to the expected signal. In these analyses, event selection proceeds by first cutting on several variables, then performing the optimisation in the final variable with best discriminating power between signal and background. In this paper, we propose the simultaneous optimisation across all variables, by finding the multi-dimensional constant Bayes posterior hypersurface that yields the lowest model rejection factor. A method is proposed here (modified penalized likelihood estimation) for identifying the level curves of the posterior. This is an extension of previous penalized likelihood methods to the use of simulated training data that is drawn from biased distributions via importance sampling, a common practice in particle physics and astrophysics.

2. OPTIMIZING EXPERIMENTAL LIMIT SETTING POWER

We will use the example of setting a limit on a diffuse flux of extraterrestrial neutrinos in an underground detector to illustrate how limits are set and optimized in the field of particle astrophysics. Suppose one has a neutrino detector which observes atmospheric neutrinos, produced via interactions of cosmic rays in the earth's atmosphere. These neutrinos are produced with a power law spectrum that goes approximately as $E^{-3.7}$. Models of an extraterrestrial flux of neutrinos from the sum of all active galaxies have a somewhat flatter power law, with an energy dependence of E^{-2}. We will write $\phi_s\Phi_s(E)$ for the extraterrestrial signal neutrino flux and $\phi_b\Phi_b(E)$ for the background atmospheric neutrino flux, where $\Phi_s(E)$ and $\Phi_b(E)$ are p.d.f.s (i.e. $\int_E \Phi_s(E)dE = 1$). We denote the detector response to the flux of neutrinos by the probability $P(x \mid E)$, where x is a possibly multidimensional vector of observables describing an event. Then the p.d.f.s for signal and background in the event space are $h_s(x) = \int_E P(x \mid E)\Phi_s(E)dE$ and $h_b(x) = \int_E P(x \mid E)\Phi_b(E)dE$. Over the space of all possible events, we therefore expect ϕ_s signal and ϕ_b background events. Then the number of signal events, $N_s(r)$ expected in some yet to be defined subregion of x, denoted Ψ_r, can be written as

$$N_s(r) = (\phi_s + \phi_b)\pi_s \int_{x\in\Psi_r} h_s(x)dx \tag{1}$$

where the prior probability for signal is defined as $\pi_s = \phi_s/(\phi_s + \phi_b)$. After reducing the data by cutting on some of the variables, thereby leaving a subregion Ψ_r of events, we wish to set a limit on the normalisation scale factor ϕ_s. This involves determining an experimental signal event upper limit $\mu(N_{obs}(r), N_b(r))$, which is a function of the number of observed events, $N_{obs}(r)$, and expected background, $N_b(r)$, after the cuts are applied. The limit on the normalisation of the source flux will then be $\phi_{lim}(r) = \phi_s \times \mu(N_{obs}(r), N_b(r))/N_s(r)$. The choice of final cut is optimized before examining the data by minimizing the average "model rejection factor", where $\mathrm{MRF}(r) = \bar{\mu}(N_b(r))/N_s(r)$ [1], where the as yet unknown experimental event limit $\mu(N_{obs}(r), N_b(r))$ is replaced by the *average* upper limit $\bar{\mu}(N_b(r))$ [3]. Over an ensemble of hypothetical repetitions of the experiment, this choice of cut will lead to the best average limit $\bar{\phi}_{lim}$. Importantly, the choice of the optimal region Ψ_r is independent of the original assumption of the normalisation of the source flux to be tested (ϕ_s cancels in the expression for $\bar{\phi}_{lim}$ leaving $\bar{\phi}_{lim}(r) = \bar{\mu}(N_b(r))/\int_{x\in\Psi_r} h_s(x)dx$). This method has been applied to the analysis of data from the AMANDA-B10 detector [2], where preliminary cuts were made to isolate atmospheric neutrinos, then the model rejection potential method was applied to the most energy sensitive variable, the number of detector optical modules that had registered Cherenkov photons in a given event. The final region Ψ_r, is essentially a "rectangular" region in the space of the observables, with only the final cut optimized to give the best limit setting potential.

Rather than finally optimize with respect to a single variable, it is desired to find a region in the multidimensional variable space for which the inclusion of events leads to the optimized model rejection potential and best limit. We use the Neyman-Pearson Lemma as a guide to defining the optimal region. It states that the critical region defining the most powerful test of one hypothesis against an alternative is given by taking all events with a p.d.f. ratio greater than some constant. Following this reasoning, we only allow regions of the form Ψ_r containing all x such that $h_s(x)/h_b(x) \geq r$. For a particular cut determined by r, we can then calculate the limit on the flux normalisation $\phi_{lim}(r) = \phi_s \bar{\mu}(N_b(r))/N_s(r)$ as a function of r and seek to find r which minimizes it. In the next section, we discuss how a model of Ψ_r, equivalently the level curves of $h_s(x)/h_b(x)$, may be obtained using a penalized likelihood estimation method.

3. MODIFIED PENALIZED LIKELIHOOD ESTIMATION

Let x be a possibly multidimensional vector of event observables derived from a reconstructed event. Let $h_s(x)$ be the probability density function for signal vectors and $h_b(x)$ be the probability density for background vectors, and let π_s and π_b be prior probabilities of a signal and background observation, respectively. Then the posterior probability that x is a signal vector is $p(x) = \pi_s h_s(x)/(\pi_b h_b(x) + \pi_s h_s(x))$. The logit $f(x)$ is defined as $\log[p(x)/(1-p(x))] \equiv \theta + \log[h_s(x)/h_b(x)]$, where $\theta = \log \pi_s/\pi_b$. We will estimate the unbounded $f(x)$ (rather than $p(x)/(1-p(x))$ or the bounded $p(x)$) for a particular (implicit) value of θ, but since the end result is to obtain level curves of f, the particular value of θ is not important for the calculations. A modified form of the penalized likelihood estimate [4–6] will be used.

Let y_i be a random variable that is 1 (signal) with probability $p(x_i)$ and 0 (background) with probability $1-p(x_i)$. Then the likelihood of a single observation y_i is: $\mathcal{L} = p(x_i)^{y_i}(1-p(x_i))^{1-y_i}$. The negative log likelihood of (independent) data $y_1, \cdots, y_n$ is then, in terms of the logit given by

$$Q(y,f) = \sum_{i=1}^{n}[log(1+e^{f(x_i)}) - y_i f(x_i)] \qquad (2)$$

We want to find $f \cong \sum c_k B_k \in H_K$ (a reproducing kernel Hilbert space (RKHS)[4, 5, 7]) which minimizes the penalized log-likelihood: $I_\lambda(c) = Q(y,f) + \lambda \parallel f \parallel^2_{H_K}$, where B_k's are basis functions in H_K and $\parallel \cdot \parallel_{H_K}$ is the function norm in H_K.

This is essentially the penalized log likelihood estimate of f proposed in O'Sullivan, Yandell and Raynor [8], and in common usage in some fields. Under rather general conditions, which include a proper choice of λ, penalized log likelihood estimates in many RKHS are known to converge to the "true" f as the sample size becomes large [9]. RKHS's are discussed in Aronszajn [7] and their use in statistical model building in Wahba [4] and elsewhere, and a wide variety of these spaces are available. An RKHS is characterized by a unique positive definite function $K(\cdot,\cdot)$, and once K is chosen, the exact minimizer of $I_\lambda(c)$ is known to be in the span of a certain set of basis functions determined from K [10]. In Section 4 below we will select a particular K, known to be a good general purpose choice, and and use an approximating subset of this set of basis functions. Estimating f rather than p directly gives a strictly convex optimization problem whose gradient and Hessian are simple to compute, which then makes the numerical analysis easier and suitable for very large data sets. It is possible to estimate p directly [11] but this estimate is harder to compute in large data sets and is believed to be not as accurate.

The form of the negative log likelihood in equation 2 applies where the simulated training data is distributed as $h_b(x)$ and $h_s(x)$ through sampling directly from the generating distributions $\Phi_s(\tilde{E})$ and $\Phi_b(\tilde{E})$, then processing the events $\tilde{E}$ through the simulation chain to give events x. Here, $\tilde{E}$ means a vector of generating parameters, e.g. neutrino energy, position and arrival direction. Often, we wish to emphasize more interesting regions of the event space, by e.g. sampling from biased energy and arrival directions, or by forcing events to occur close to the detector. Suppose these biased distributions in the generating parameters may be summarized as $g_b(\tilde{E})$ and $g_s(\tilde{E})$, or there may be a single biased sampling distribution, $g(\tilde{E})$. We "unbias" the events by applying weight factors throughout any subsequent procedure. The weight for a given signal event x_i will be $w_s(x_i) = \Phi_s(\tilde{E}_i)/g_s(\tilde{E}_i)$ and for a background event $w_b(x_i) = \Phi_b(\tilde{E}_i)/g_b(\tilde{E}_i)$. In the case that a single sampling spectrum is used each event is re-weighted to both signal and background energy spectra. In either case the weights satisfy $\sum_{i=1}^{n} w_s(x_i) = N_s$ and $\sum_{i=1}^{n} w_b(x_i) = N_b$, i.e. the predicted numbers of events from the weighted simulation is the same as that from an un-weighted simulation. Now, if we have multiple unbiased observations at some x_i as $y_{ij}, j = 1, \ldots, m(i)$, the likelihood of all these observations is: $\mathcal{L} = p(x_i)^{\sum_{j=1}^{m(i)} y_{ij}}(1-p(x_i))^{\sum_{j=1}^{m(i)}(1-y_{ij})}$. If the samplings at x_i are biased, then the exponent sums are weighted by $w_s(x_i)$ and $w_b(x_i)$ respectively leading to a modified likelihood

$$Q(w,f)=\sum_{i=1}^{n}\sum_{y_i=0}^{1} w_{y_i}[\ log(1+e^{f(x_i)})-y_i f(x_i)] \quad (3)$$

where $w_{y_i}=w_s(x_i)$ for $y_i=1$ and $w_{y_i}=w_b(x_i)$ for $y_i=0$. The incorporation of weighted events is thus simply accounted for by weighting the terms in the logarithmic likelihood sum. Further, we can substitute $w_s(x_i)$ and $w_b(x_i)$ to obtain an alternative form of the likelihood

$$Q(w,f)=\sum_{i=1}^{n}\{w_t(x_i)[\log(1+e^{f(x_i)})-\tilde{p}(x_i)f(x_i)]\} \quad (4)$$

where $w_t(x_i) = w_s(x_i) + w_b(x_i)$ and $\tilde{p}(x_i) = w_s(x_i)/w_t(x_i)$.

4. IMPLEMENTATION OF THE MODIFIED PLE METHOD

After getting the modified penalized likelihood formulation, we now can move on to look for a 'good' estimate of $f(x)$ whose level curves can be obtained. In our implementation, we use radial basis functions plus constant and linear terms. So,

$$f(x)=\beta_0+\beta^T x+\sum_{k=1}^{N} c_k K_\sigma(x,x_{i_k}), \quad (5)$$

where $K_\sigma(\cdot,\cdot)$ is the Gaussian kernel with isotropic variance σ^2, N is the total number of basis functions and the N $x_{i_k}, k=1,\cdots N$ will be chosen as a subset of the $x_i, i=1,\cdots,n$ as described below. Thus, f will be specified as long as all coefficients, i.e. β_0, β and the c_i's are determined (note that β is a vector). By letting $\lambda \parallel f \parallel^2_{H_K} = \lambda \sum_{i,j=1}^{N} c_i c_j K_\sigma(x_i,x_j)$, we put a penalty only on the c_i's.

We used a sequence of simulated data driven procedures to fit the model in the sense that we let the simulated data choose the 'best' combination of smoothing parameter λ, scale parameter σ and number of basis functions N. Five dimensional simulated data (x_i's) are first rescaled using their own sample weighted standard deviation after a log transformation. Then, the whole simulated data set is randomly divided into three subsets of almost the same size, one as training set, one as tuning set and the last one as testing set. After that, we randomly, but according to weights (large-weight simulated data points have higher chance to be selected), choose the N x_{i_k}'s which determine basis functions as a subset of the training set. We solve the minimization problem on a coarse 2-D parameter grid of λ (usually on a log scale) and σ^2 using the training set. For each parameter pair (each point on the grid), a Newton-Raphson iteration is used to solve this convex minimization problem [4]. After the algorithm converges we calculate the Kullback-Leibler (KL) distance between tuning simulated data and fitted model, which is essentially just the first term of $I_\lambda(c)$ for tuning simulated data with f replaced by $\hat{f}$. We then find the best parameter combination based on the KL distance over the coarse grid. Starting from there, a direct-searching simplex method [12] is used to search for a locally best parameter combination according to the KL distance criterion. The procedure is repeated using $2N$ bases, then $4N$ bases and so on, until the improvement on the KL distance is smaller than some preset threshold. We use the coefficients corresponding to the then-best combination of parameters to construct our final estimate of the logit function. Next, the testing set is used to check the goodness of fit of this final model and to determine the optimal cut on $p(x_i)$ and thus the limit setting power of the analysis. Finally, the real data can be analysed by applying the optimal $p(x_i)$ cut, and the limit on the signal model determined.

5. CONCLUSIONS

In this work, we have described how the limit setting potential of a multivariate analysis is optimized by choosing an acceptance region of all x for which the estimated posterior probability $p(x)$ is greater than some specified threshold. The main contribution of this paper is to introduce the well known penalized log likelihood estimation procedure for estimating f and hence p to an audience to which it is apparently unfamiliar, and to develop numerical algorithms for efficient computation and testing of the estimate that are appropriate for large multivariate data sets obtained via importance sampling.

Acknowledgments

This research was supported by the National Science Foundation under Grants DMS-0072292 (G. Wahba and F. Lu) and OPP-9980474 (G.C. Hill and P. Desiati).

References

[1] G. C. Hill and K. Rawlins, “Unbiased cut selection for optimal upper limits in neutrino detectors: the model rejection potential technique”, Astropart. Phys. **19**, 393, 2003.

[2] J. Ahrens *et al.* “Limits on diffuse fluxes of high energy extraterrestrial neutrinos with the

AMANDA-B10 detector", Phys. Rev. Lett. **90**, 251101, 2003.

[3] G. J. Feldman and R. D. Cousins, "Unified approach to the classical statistical analysis of small signals", Phys. Rev. D **57**, 3873, 1998.

[4] G. Wahba, "Spline Models for Observational Data", CBMS-NSF Regional Conference series in applied mathematics, **59**, 1990.

[5] G. Wahba, "Soft and Hard Classification by Reproducing Kernel Hilbert Space Methods ", Proceedings of the National Academy of Sciences, **99**, 16524-16530, 2002.

[6] G. Wahba, Y. Wang, C. Gu, R. Klein, and B. Klein, "Smoothing Spline ANOVA for Exponential Families, with Application to the Wisconsin Epidemiological Study of Diabetic Retinopathy ", Ann. Statist., **23**, 1865-1895, 1995.

[7] N. Aronszajn, "Theory of reproducing kernels "Trans. Am. Math. Soc., **68**, 337-404, 1950.

[8] F. O'Sullivan, B. Yandell and W. Raynor, "Automatic smoothing of regression functions in generalized linear models", J. Amer. Statist. Assoc., **81**, 96-103, 1986.

[9] D. Cox, and F. O'Sullivan, "Asymptotic analysis of penalized likelihood and related estimators" Ann. Statist. **18**, 1676-1695, 1990.

[10] G. Kimeldorf and G. Wahba, "Some results on Tchebycheffian spline functions", J. Math. Anal. Applic., **33**, 82-95, 1971.

[11] M. Villalobos and G. Wahba, "Inequality constrained multivariate smoothing splines with application to the estimation of posterior probabilities", J. Am. Statist. Assoc., **82**, 239-248, 1987.

[12] J.C. Lagarias, J. A. Reeds, M. H. Wright, and P. E. Wright, "Properties of the Nelder-Mead Simplex Method in Low Dimensions ", SIAM Journal of Optimization, **9**, Number 1, 112-147, 1998.

Point Processes and the Analysis of Collision Data

David S. Newman, Retired (Boeing Applied Mathematics and Statistics Group)
Seattle, Washington, USA

Heavy ion collisions have been investigated since the 1960's. The current series of experiments at the RHIC facility at Brookhaven began in 2000. Each pair of colliding nuclei will yield from several hundred to over 1000 particles, and a typical experimental run will involve tens of thousands of collisions. The current generation of detectors identifies and measures the momentum of the majority of the charged particles emitted in a collision. It is natural to regard the observed particles in a single collision as a point process, and a series of collisions as an independent (but not identically distributed) sample from that process.

While point process methods are implicit in the method heavy-ion physicists use for pion intensity interferometry, their approach to analysis would be improved by their explicit use. The potential relevance of point processes to more precise single particle collision data analysis and more generally to collision modelling is discussed briefly.

1. INTRODUCTION

At very high energies, collisions between leptons, hadrons and/or heavy ions of given species exhibit substantial variability in the outcome from collision to collision, even at fixed energy and impact parameter. There is stochastic variation not only in the angular distribution and momenta of the observed particles, but also in the numbers and proportions of emerging particle species. Currently used modeling and statistical practices do not adequately address this variation and are therefore unable to take advantage of it. This paper outlines how the use of point processes [1, 2] can clarify and possibly remedy this situation.

Once the raw experimental data from collisions has been "decoded" by track separation and particle identification algorithms, the most natural probabilistic and statistical description of the possible observed outcomes of a collider experiment is that of a point process. This framework accounts for all the variability in the number and mix of particles produced as well as the momenta of the individual particles, and represents the data completely, except for instrumentation issues.

Validation of theory with data from collider experiments is generally linked to quantum field theories via differential cross sections, which correspond to the joint probability density of collisions with a fully specified set of particles as the outcome. In the general theory, collisions with different particle mix outcomes are treated separately. In earlier years, experimental particle physics emphasized statistical models (e.g., the Fermi model) of the nucleus. An important refinement of Fermi's model made in 1960 by Goldhaber et al [3] when they introduced pion intensity interferometry, but statistical modeling was soon overshadowed by the discovery of partonic interactions. As the Standard Model developed and its experimental consequences were probed more deeply, increasing collision energies made partonic models more reliable, and statistical models were ignored. But recently the need for greater precision in reporting of results is creating new demands on data analysis. Theoretical results expressed only as differential cross sections for completely specified outcomes do not specify a complete probability distribution for the numbers and species of emerging particles. Partonic models further approximate collisions in the form "x + y ➔ z + anything." But they provide no specific guidance for the treatment of the "anything" in data analysis, so that stochastic dependence on any details of the "anything" is ignored. In practice there is no standard method for calculating cross sections for all possible particle mix outcomes (or all such outcomes with probabilities above some small threshold) from theory. This situation makes the use of point process based data analysis more appealing and perhaps essential.

2. BACKGROUND ON HEAVY ION COLLIDER EXPERIMENTS

Statistical models were not ignored in heavy-ion physics, and have developed extensively since Goldhaber et al's seminal paper. In addition to cross-section based partonic and more "fundamental" models, heavy-ion collision modeling requires a variety of statistical transport process, hydrodynamic flow, and other statistical models of collective behavior. Pion intensity interferometry plays a crucial role in linking these intermediate statistical models with the identified data as it presents itself, i.e. as a point process.

The current series of heavy ion experiments at the RHIC (Relativistic Heavy Ion Collider) facility at Brookhaven began in 2000. In addition to exploring the properties of hadronic and nuclear matter, these experiments are searching for new phenomena, especially the elusive quark-gluon plasma (QGP), at collision energies of up to 200 Gev per nucleon. Each pair of colliding nuclei will yield from several hundred to over 1000 particles, and a typical experimental run will involve tens of thousands of collisions. These experiments are also expected to shed light on such diverse phenomena as phase transitions in subnuclear matter, the early history of the "big bang," stellar evolution, neutron stars, and other astronomical and cosmological phenomena.

Some of the most recently developed detectors — for example, the STAR detector at RHIC — are capable of identifying and measuring the momenta of virtually all the charged particles emitted in a collision. This capability presents an opportunity to address new issues, both in theory and experimentally, which have received little attention in the past. It should be possible, for example, to determine if the occurrence of one phenomenon affects the probability of others, i.e., their occurrence or non-occurrence is statistically dependent. The momentum phase space distribution of mesons (especially pions) plays a key role in other aspects of the analysis. The point process structure is implicit in the current method of pion data analysis, but the failure to recognize it explicitly has led to some misunderstandings which will be discussed below.

Many of the pions produced in heavy-ion collisions are at low momentum. The phase space-time region in which they are produced is small enough so that even if they are assumed (as is generally done) to be produced independently, i.e. by unrelated partonic events, identical pions must be represented by a single, symmetrized multiparticle wave function, since they are bosons. As is the case with photons, this fact makes intensity interferometry — also known as HBT interferometry, for the work of Hanbury-Brown and Twiss [4] — possible. Its introduction to nuclear collisions is due to Goldhaber et al, as noted earlier.

From the observed momentum distribution, pion HBT interferometry reconstructs the position space-time distribution of pion generation, which describes the hadronic "freeze-out" space-time surface, and the equation of state of the ultra-dense matter which precedes the hadronic state [5]. A realistic treatment of this reconstruction requires the support of the other statistical models cited earlier. This freeze-out surface geometry is a vital prediction target for models that attempt to represent the early stages of the collision, especially those concerned with how the phase transition from a pre-hadronic state (whether QGP or some other non-nuclear state) to observed hadrons develops.

Space does not permit a more detailed discussion of the method used by the heavy-ion community for HBT interferometry, how it relates to point processes, and how it could be improved by an explicit recognition of the connection. This material is available in an extended draft of this paper, and will be presented elsewhere.

3. POINT PROCESSES AND COLLISION DATA

Point processes developed from a number of converging ideas, all of which may be thought of as generalizations of a Poisson process. Important early contributions were made by physicists working on the analysis of cosmic ray showers. Subsequent developments by Bogoliubov, together with original mathematical discoveries prompted by very different fields of application, form the core of the subject. A brief history of the subject is given in [1]. With the exception of the work of Bénard and Macchi [6], which made the point process basis of quantum optics explicit, it appears that there has been little or no contact between more recent mathematical developments and the physics community since the 1950's.

In heavy ion experiments, many important comparisons of data to theoretical predictions are based on momentum distributions of specific particles which can be directly related to cross-section calculations. As we shall see, these distributions are not probability distributions in the ordinary sense. In point process terminology, they are *first moment distributions*. The distinction is extremely important, for reasons related to our comments about the highly variable outcomes of high energy collisions.

In the simplified case where only a single particle species is of interest, and all collisions occur at the same energy and impact parameter, collisions are not only independent but identically distributed. The probability distribution $\{p_n\}$ for the number of particles produced in a given collision is the *multiplicity distribution*, with $p_n \geq 0$ for $n \geq 0$, and $\sum_{n=0}^{\infty} p_n = 1$. As noted earlier, it is reasonable to assume that collisions are statistically independent of one another. Given *exactly n* identical particles, the 3-momenta $(\mathbf{k}_1,\ldots,\mathbf{k}_n)$ of these particles have a joint probability density $P_n^{(n)}(\mathbf{k}_1,\ldots,\mathbf{k}_n)$. The double-index notation used here is that of Zimanyi and Csörgö [7]; they and several other authors refer to it as the *n*-particle "exclusive" probability distribution for fixed multiplicity *n*, the "exclusive" referring to the absence of other particles which may influence the distribution. These joint probability densities are symmetric under permutations of the particle indices $1,\ldots,n$, reflecting the symmetry or antisymmetry of underlying field operators for bosons or fermions respectively.

The momentum distribution used in collider data analysis is *not* $P_1^{(1)}(\mathbf{k})$. This momentum distribution is estimated from data by creating a 3-dimensional histogram, i.e., by defining a collection of *B* (3-dimensional) momentum bins $\{\Delta\mathbf{k}_b,\ldots,b=1,\ldots,B\}$ and counting the numbers of particles observed in each from a series of collisions. If one passes to the limit as the number of collisions approaches infinity and the bins shrink to zero in volume, the resulting histogram approaches the *moment density function* for the point process population from which the collisions form an independent sample. In terms of the multiplicity distribution and probability densities defined above it has the definition

$$N_1(\mathbf{k}) = \sum_{n=1}^{\infty} np_n P_1^{(n)}(\mathbf{k}) \ , \qquad (1)$$

where

$$P_1^{(n)}(\mathbf{k}) \ = \int_{K^{(n-1)}}\!\!\ldots\int P_n^{(n)}(\mathbf{k},\mathbf{k}_2,\ldots,\mathbf{k}_n)\ \mathbf{dk}_2\ldots\mathbf{dk}_n \ , (2)$$

which is the marginal probability density for the momentum of one particle, given that exactly n are present. $K^{(n-1)}$ is *3(n-1)* dimensional momentum space. Note that in equation (1) the number of particles is indefinite, reflecting the fact that the "binning" in the sample estimate of $N_1(\mathbf{k})$ is performed over collisions with a variable number of particles. It is a moment density because

$$\int_K N_1(\mathbf{k})\,\mathbf{dk} = \langle N \rangle = \sum_{n=0}^{\infty} np_n \qquad (3)$$

is the first *population* moment (or mean, or average) of *N*, the multiplicity, i.e., the random number of particles in a single collision. For comparison with data from a finite sample, $\int_{\Delta\mathbf{k}_b} N_1(\mathbf{k})\,\mathbf{dk}$ is estimated, for each *b*, by the number of particles observed in $\Delta\mathbf{k}_b$ in all collisions, divided by the number of collisions - in other words the *sample* mean number of particles, by bin. This is simply an application of the method of moments for estimation: the *sample* mean *estimates* the *population* mean. In fact one does not need to bin one's data. To estimate $N_1(A) = \int_A N_1(\mathbf{k})\,\mathbf{dk}$, where A is *any* subset of K, simply divide the number of particles observed in A in all collisions by the number of collisions. The standard theorems - the law of large numbers and the central limit theorem in particular - all apply to this estimate ($\hat{N}_1(A)$ in the notation of many statistics texts) of $N_1(A)$. When random variables are replaced by observed values, the estimat*or* becomes an estimat*e*. Please pardon the picky statistician terminology, but this basic sample/population conceptual framework is so quickly ignored!

Higher moment densities can also be defined, and are in fact used in heavy-ion data analysis for pion interferometry. The second *factorial* moment density for the *population* is defined by

$$N_2(\mathbf{k}_1, \mathbf{k}_2) = \sum_{n=2}^{\infty} n(n-1)\, p_n P_2^{(n)}(\mathbf{k}_1, \mathbf{k}_2), \qquad (4)$$

where

$$P_2^{(n)}(\mathbf{k}_1, \mathbf{k}_2) = \int_{K^{(n-2)}} P_n^{(n)}(\mathbf{k}_1, \mathbf{k}_2, \ldots, \mathbf{k}_n)\, \mathbf{dk}_3 \ldots \mathbf{dk}_n\ , \qquad (5)$$

the marginal density for two particles, given that *n* are present. The factorial rather than the ordinary moment density is used, to avoid counting particles paired with themselves. The estimator corresponding to (4) is $\hat{N}_2(A_1, A_2)$

$$= \frac{1}{C} \sum_{j=1}^{C} \int_{A_1} \int_{A_2} \left(\sum_{\substack{i,i'=1 \\ i \neq i'}}^{N_j} \delta(\mathbf{k}_1 - \mathbf{K}_{ij})\, \delta(\mathbf{k}_2 - \mathbf{K}_{i'j}) \right) \mathbf{dk}_1 \mathbf{dk}_2 .$$

Note that the inner sum has $N_j(N_j - 1)$ terms. The integral of the population moment (4) over momentum pair space $K^{(2)}$ is $\langle N(N-1) \rangle = \sum_{n=2}^{\infty} n(n-1)\, p_n$.

Higher factorial moment densities are similarly defined. One occasionally needs third or fourth moments in practice in conventional sample statistics. Some heavy-ion HBT investigators have suggested that these may contain valuable information. These moments are defined by

$$N_m(\mathbf{k}_1, \ldots, \mathbf{k}_m) = \sum_{n=m}^{\infty} [n]_m\, p_n P_m^{(n)}(\mathbf{k}_1, \ldots, \mathbf{k}_m), \qquad (6a)$$

and $\int_{K^m} N_m(\mathbf{k}_1, \ldots, \mathbf{k}_m)\, \mathbf{dk}_1 \ldots \mathbf{dk}_m = \sum_{n=m}^{\infty} [n]_m\, p_n = \langle [N]_m \rangle$. (6b) The notation $[n]_m = n(n-1)\ldots(n-m+1) = n!/(n-m)!$, widely used in numerical analysis, has been adopted, and $P_m^{(n)}$ is the marginal probability density for the momenta of the "first *m*" particles, given that exactly *n* are present, generalizing (2) and (5). $\langle [N]_m \rangle$ is the m^{th} factorial moment of the random multiplicity *N*, which has probability distribution $\{p_n\}$. From (6) it is clear that in the m^{th} factorial moment density, the number of particles present is at least *m* but otherwise indefinite.

It is worthwhile seeing how these formulas simplify for a Poisson process. Since the individual particles in a collision are now completely independent of one another, the basic probability densities factorize: $P_n^{(n)}(\mathbf{k}_1, \ldots, \mathbf{k}_n) = \prod_{i=1}^{n} P_1^{(1)}(\mathbf{k}_i)$. The number of events has a Poisson distribution, $p_n = e^{-\lambda}\lambda^n / n!$. Then the first moment density is, from (1),

$$N_1(\mathbf{k}) = \sum_{n=1}^{\infty} np_n P_1^{(n)}(\mathbf{k})$$

$$= P_1^{(1)}(\mathbf{k}) \sum_{n=0}^{\infty} ne^{-\lambda}\lambda^n / n! = \lambda P_1^{(1)}(\mathbf{k}),$$

so that when integrated over K one has $\langle N \rangle = \lambda$. This Poisson process is, in general, inhomogeneous (not constant in infinite momentum space), and the first moment density is just the probability density of position for a single particle times the mean number of particles per collision. In this case — *and only in this case* —the first moment density can be identified with the probability density. The higher moment densities simplify in like manner.

Returning to the general case, the factorial moment densities $N_m(\mathbf{k}_1, \ldots, \mathbf{k}_m)$ are called "inclusive distributions" in many places in the heavy-ion literature [e.g. 7,8]. They have the following probabilistic interpretation [1, p.133] which does not appear to be widely appreciated: $N_m(\mathbf{k}_1, \ldots, \mathbf{k}_m)\mathbf{dk}_1 \ldots \mathbf{dk}_m$ is the probability that there is a single particle in each of the *m* momentum space infinitesimal volume elements $\{[\mathbf{k}_1, \mathbf{k}_1 + \mathbf{dk}_1), \ldots, [\mathbf{k}_m, \mathbf{k}_m + \mathbf{dk}_m)\}$, *allowing for the presence of other particles elsewhere*. It describes the distribution of *m* of the particles in a sampled event. But it is not a probability distribution, because the infinitesimal events are not mutually exclusive. The term *inclusive* is appropriate since an indefinite number of additional particles, in addition to the *m* with momenta specified by N_m, may be present. Because of this, sample moment densities cannot be used directly as a basis for likelihood inference. It may be possible to approximate the likelihood using the inverse of (6a) which expresses $P_n^{(n)}$ in terms of N_m, see [1] for details.

Because these infinitesimal events are not mutually exclusive, dividing the function N_m by $\langle [N_m] \rangle$ in order to make it integrate to one does not create a probability distribution for the particles in question. Unfortunately, this practice is seen frequently in both theoretical and experimental physics. While the normalization technically creates a probability distribution, its interpretation is unclear. The implications of a careful treatment of factorial moment densities in statistical and quantum physics are a subject for future research, and could have far-reaching implications.

The point process framework generalizes to more than one particle type. When only a single particle species is of interest, the outcome of a single collision is characterized by its multiplicity and its vector of momentum vectors, i.e. the pair of quantities n and $(\mathbf{k}_1, \ldots, \mathbf{k}_n)$. For S particle species, one needs S copies of this structure to characterize a collision. If $n_s, s = 1, \ldots, S$ are the multiplicities of the individual species, denote the n_s-vector of momentum vectors of the s^{th} species by $\vec{\mathbf{k}}_s = (\mathbf{k}_{1,s}, \ldots, \mathbf{k}_{n_s,s}), s = 1, \ldots, S$, which is a $3n_s$ dimensional vector. The joint multiplicity distribution takes the form $p(n_1, \ldots, n_S) = p(\tilde{n})$, and the joint densities of momentum vectors given $\tilde{n} = (n_1, \ldots, n_S)$ are denoted by $P_{\tilde{n}}^{(\tilde{n})}(\vec{\mathbf{k}}_1, \ldots \vec{\mathbf{k}}_S)$. The latter are symmetric *within* the momentum components for a single particle species, reflecting the symmetry (for bosons) or antisymmetry (for fermions) of underlying field operators, as noted earlier. But symmetry *between* species is not required.

This general point process formulation appears quite complex, but models based on quantum field theory and statistical physics will "flow through" to the point process framework for data analysis. For single-hadron and lepton collisions and for "partonic events" (i.e. jets) which can be identified within more complex collisions, conservation of momentum and energy may reduce the number of degrees of freedom substantially, and other particles in large n events may perhaps be ignored, up to a point. But looking at marginal distributions of the particles in a collision which are of interest, averaged over those which are not, will give a clearer picture of what is happening than informal approximations.

Special techniques may be used to carefully identify jets and other specific phenomena within a collision. It may be possible to formalize these techniques within the point process context by using marginal and conditional probability arguments, or if not, to validate them as good exploratory methods *post hoc* using point process based statistical inference. Generally, for collisions in which large numbers of identical particles are produced by similar processes, further reductions in the complexity of the point process description occur where n-body interactions can be summarized by pairwise functions of the interaction. Bénard and Macchi [6] show that n-particle wave functions of identical bosons and fermions can be described in this way in many situations, a result that is likely to be very useful in particle interaction modeling.

References

[1] D.J. Daley and D. Vere-Jones, *An Introduction to the Theory of Point Processes*. New York: Springer-Verlag, 1988. Second edition, Vol. 1, 2002. Page and chapter references are to the first edition. This text has the best coverage of methods and techniques which will prove relevant to physicists. It is much more readable than most treatises on this subject, much of it being written in the style of William Feller's classic work on probability. There is little on statistical inference here, but that is true of almost all books on the subject, and most treatments are extremely abstract and axiomatic. Still, some may find [1] "too mathematical." In the new edition some of the "heavy mathematics" is postponed to Vol. 2. For a lighter mathematical treatment [2] is recommended as a starter, but it does not cover enough of the material physicists will need.

[2] D.R. Cox and V. Isham, *Point Processes.* New York: Chapman & Hall, 1980 (now published by CRC Press).

[3] G. Goldhaber, S. Goldhaber, W. Lee and A. Pais, *Phys. Rev.* **120** (1960), 300

[4] R. Hanbury Brown and R.Q. Twiss, *Phil. Mag.* **45** (1954) 663; *Nature* **177** (1956) 27; *ibid* **178** (1956) 1046; *Proc. Roy. Soc.* **A 242** (1957) 300; *ibid* **243** (1957) 291; **248** (1958) 199; and **248** (1958) 222.

[5] U. Heinz and U. A. Wiedemann, *Phys. Rep.* **319** (1999) 145, nucl-th/9901094

[6] O. Macchi, *C. Rendus Acad. Sci.* série A **272** (1971), 437; C. Bénard, *Phys. Rev.* A **2** (1970), 2140; O. Macchi, *IEEE Transactions on Information Theory* **17** (1971), 2; C. Bénard and O. Macchi, *J. Math. Phys.* **14** (1973), 155; C. Bénard, *J. Math. Phys.* **16** (1975), 710; O. Macchi, *Advances in Applied Probability* **7** (1975), 83

[7] J. Zimanyi and T. Csörgö, *Heavy Ion Phys.* **9** (1999), 241, hep-ph/9705432

[8] I.V. Andreev, M. Plümer and R.M. Weiner, *Intl. Jour. of Mod. Phys.* **A 8** (1993), 4577

A Simple Iterative Alignment Method Using Gradient Descending Minimum Search

G. Zech and T. Zeuner
University of Siegen, 57072 Siegen, Germany

Large tracking systems of present experiments in high energy physics consist of typically 10^4 individual detector elements. The alignment procedure has to fix up to 10^5 parameters. We propose a simple and robust iterative updating alignment method using reconstructed tracks. The geometrical parameters are adjusted using a gradient descending minimum search which does not require handling of large matrices and can be used online. Tracks with a common vertex like V^0s can be used for an efficient alignment of modules which are not linked by single tracks. The method has been successfully applied in the alignment of the HERA-B tracking system.

1. INTRODUCTION

Modern experiments in particle physics are equipped with tracking detectors covering typically volumes of $(10\ m)^3$ and 10^7 channels in typically 10^4 mechanically independent modules. The resolution of the measurements is of order μm and requires the precise adjustment with particle tracks of the 10^5 free parameters corresponding to the degrees of freedom of the modules. Blobel et al. have developed a least square fitting procedure [1] which estimate all parameters at the same time. The corresponding program package has very successfully been applied to the tracking system of H1 and the vertex detector of HERA-B.

In this paper we present an alternative approach which is based on gradient descending minimum search. It offers the following advantages:

- It is very simple, robust and fast and does not require handling of large matrices.
- It is applicable not only to single tracks but also to complex event structures. Magnet parameters can be included in the adjustment.
- It is updating and can be used on-line.

The method has been applied successfully to the alignment of the Inner Tracker of HERA-B.

2. GENERAL REMARKS

The alignment of detectors is based on the residuals, the differences between fitted and measured coordinates. Thus a crude alignment of the detectors is necessary before an automatized alignment procedure can start. The alignment of the whole tracking system in a single program is more efficient than a two-step procedure where subdetectors are aligned first.

Usually, single tracks are used to align the detector elements. These tracks connect mainly certain groups of detectors which are located near a radial line drawn from the interaction point. Thus residual plots are usually swamped by those tracks and may look perfect except for some outliers but hide a rather bad alignment of laterally located detectors. However, it is the alignment of laterally located detectors which is important for a good mass resolution. For the alignment of these detectors some detector overlap is very helpful. Even better is to use events consisting of at least two tracks with some kinematical constraint like a common vertex at the interaction point or $K^0 \to \pi^+\pi^-$ and $\Lambda \to p\pi^-$ decays to correlate detectors not traversed by the same particle. Using these events also helps to adjust overall scaling parameters (scaling, sharing, stretching of the complete detector) which are not accessible with single track fitting. In experiments with a spectrometer magnet it is sometimes difficult to disentangle magnet parameters (position and field) and geometrical detector constants. It is easy to include magnet constants in the gradient descending alignment procedure.

Contrary to what is frequently claimed, alignment usually does not require high statistics. The order of thousand tracks per individual detector should be sufficient if there is enough overlap between adjacent detectors.

3. OUTLINE OF THE METHOD

We assume that event or track parameters are adjusted by a χ^2 fit to the measured coordinates u_i of the individual detectors. The value of χ^2 depends on the difference $u_i - u_i^f$ of measured and fitted coordinates and on the position parameters of the detectors. The position parameter λ_k of detector k is modified in proportion to the derivative of χ^2 with respect to the parameter.

$$\delta\lambda_k = -\alpha_k \frac{\partial\chi^2}{\partial\lambda_k} = -\alpha_k \sum_i \frac{\partial}{\partial\lambda_k} \frac{(u_i - u_i^f)^2}{\Delta u_i^2}$$

Here α_k is a learning constant and Δu_i^2 the quadratic uncertainty used in the χ^2 calculation. The corre-

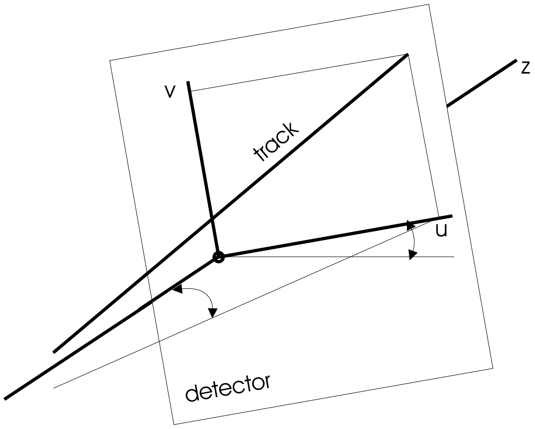


Figure 1: Coordinate definitions for a planar detector measuring the u coordinate.

sponding parameter shifts are

$$\delta\lambda_k = -2\alpha_k \sum_i \frac{u_i - u_i^f}{\Delta u_i^2} \frac{\partial(u_i - u_i^f)}{\partial\lambda_k}$$

where i runs over all measured coordinates.

An exact analytic expression for the derivative $\partial(u_i - u_i^f)/\partial\lambda_k$ can be given only in the simplest cases. There are two ways to compute it: i) It can be obtained directly from the fitting program from the observed change of χ^2 by repeating the fit with a small change $\delta\lambda_k$ of the parameter. ii) An analytic approximation is used.

The computation of a huge number of gradient components by program may require excessive computing time. The second method needs less computing power and is usually preferable.

It should be remarked that the gradient descending minimum search is insensitive to approximations as long as the steps proceed downwards in the χ^2 field. Since $u_i - u_i^f$ depends mostly on the geometrical parameters of the detector measuring u_i, all other derivatives can be neglected and $\partial(u_i - u_i^f)$ can be approximated by ∂u_i. From the six possible parameters of a planar detector usually only three are relevant. For a fixed target experiment (see Fig. 1) with one dimensional readout these are the position λ_u along the readout coordinate u, the azimuthal rotation angle λ_ϕ around the radius vector from the nominal interaction point to the chamber and the distance λ_z to the interaction point.

The three corresponding coefficients are simply:

$$\frac{\partial u}{\partial\lambda_u} = 1; \quad \frac{\partial u}{\partial\lambda_z} = \tan\gamma; \quad \frac{\partial u}{\partial\lambda_\phi} = -v$$

with $\tan\gamma$ the slope of the track in the $u - z$ plane (z is perpendicular to the nominal detector plane), and v the coordinate perpendicular to u measured in a coordinate system with the origin at $(0, 0)$ and the ϕ rotation around the z axis. If needed, the two other rotation parameters can easily be included in an analogous way.

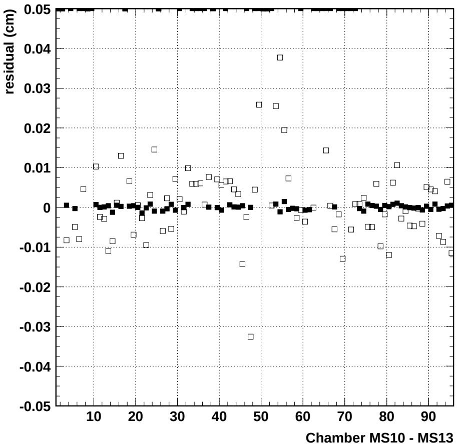


Figure 2: Mean residuals before (open sympols) and after the alignment (filled symbols). (The symbols at 0.5 cm correspond to missing detectors.)

Some remarks:

- Gradient descending minimum search often suffers from slow convergence and local minima. In our case, there are no local minima and due to the quadratic dependence of χ^2 on the difference, the χ^2 field is parabolic even far from the minimum.
- The learning constants have to be chosen such that the different parameter shifts are reasonable. The shift per event should be a small fraction of the corresponding resolution. The learning constant can be reduced during the alignment procedure.
- To minimize the computing time needed for updating data banks, the displacements of many events can be accumulated before updating.
- The same events can be used several times.
- Magnet constants can be handled in the same way as geometric parameters. The χ^2 dependence has to be estimated or computed by the reconstruction program. The additional computing time should be tolerable.

We have used the proposed method to align the Inner Tracker of the experiment HERA-B at Desy. It consists of 160 MSGC detectors. The same event sample of 10,000 events was used several times. The parameters converged after about 5 to 10 iterations. Fig. 2 shows the mean values of the residuals of part of the detectors before and after the alignment.

Acknowledgments

We acknowledge the finacial suport by the German Bundesministerium für Bildung und Forschung, Frderkennzeichen 05 HB1PSA/4

References

[1] V. Blobel and C. Kleinwort, “A new method for the high-precision alignment of track detectors”, Proc. of the Conf. on Advanced Statistical Techniques in Particle physics, ed. M. R. Walley, L. Lyons, Durham 2002.

MAGIC: Exact Bayesian Covariance Estimation and Signal Reconstruction for Gaussian Random Fields

Benjamin D. Wandelt*
Department of Physics and Department of Astronomy
University of Illinois at Urbana-Champaign, IL 61801, USA

In this talk I describe MAGIC [1], an efficient approach to covariance estimation and signal reconstruction for Gaussian random fields (MAGIC Allows Global Inference of Covariance). It solves a long-standing problem in the field of cosmic microwave background (CMB) data analysis but is in fact a general technique that can be applied to noisy, contaminated and incomplete or censored measurements of either spatial or temporal Gaussian random fields. In this talk I will phrase the method in a way that emphasizes its general structure and applicability but I comment on applications in the CMB context. The method allows the exploration of the full non-Gaussian joint posterior density of the signal and parameters in the covariance matrix (such as the power spectrum) given the data. It generalizes the familiar Wiener filter in that it automatically discovers signal correlations in the data as long as a noise model is specified and priors encode what is known about potential contaminants. The key methodological difference is that instead of attempting to evaluate the likelihood (or posterior density) or its derivatives, this method generates an asymptotically exact Monte Carlo sample from it. I present example applications to power spectrum estimation and signal reconstruction from measurements of the CMB. For these applications the method achieves speed-ups of many orders of magnitude compared to likelihood maximization techniques, while offering greater flexibility in modeling and a full characterization of the uncertainty in the estimates.

1. INTRODUCTION

Signal reconstruction from noisy data is one of the *raisons d'être* of applied statistics. If the signal is a Gaussian random field, and the signal and noise covariances are known in advance, Wiener filtering [2, 3] is the theoretically optimal method for estimating the signal from noisy data. In this simple case the solution is a linear operator that acts on the data vector and returns the minimum variance, maximum likelihood and maximum a posteriori estimator of the signal given the data.

What ought to be done, however, if the signal covariance is not known in advance, and the signal covariance must be estimated from the data? In fact there are applications where covariance estimation is the primary goal and signal reconstruction is secondary. These cases have traditionally been treated separately. For stationary signals, the covariance of the signal is best specified in the Fourier basis since this basis diagonalizes the covariance matrix. In these cases covariance estimation becomes power spectrum estimation. One such example is cosmic microwave background data (CMB) analysis which motivated this analysis. I will return to it in section 4. Other examples are time series analysis, spatial analysis of censored data, such as geological surveys, power spectrum estimation and signal reconstruction for helioseismology, image reconstruction based on a stochastic model of the form of pixel-pixel correlations, etc. The method described here generalizes the results of [3] and should therefore also be useful for the applications discussed there.

In this talk I will first review the common structure that underlies these apparently different statistical problems (section 2). I will then summarize the main advances realized by the new method in section 3. The subsequent section contains the results from the application of this new approach to the first all-sky CMB data set. Further details and examples can be found in our paper [1] and online materials at the conference WWW site [4].

The ideas in this paper were developed from a Bayesian perspective. There are pros and cons of Bayesian estimation. The pros are many: it maximizes the use of all available information and treats measurements, constraints and model on the same footing as information. The result of a Bayesian estimation is a probability density, not just a number, so one automatically obtains uncertainty information about the estimate. However, if Bayesian methods are implemented naively, these advantages come at the price of heavy computation especially for multivariate problems. However the results presented in this paper are an example that it is possible to overcome these computational challenges and make Bayesian techniques work in a highly multivariate ($D \sim 10^6$) problem.

2. SIGNAL RECONSTRUCTION AND COVARIANCE ESTIMATION

In this section I will review the problems of signal reconstruction and covariance estimation from a

*NCSA Faculty Fellow

Bayesian perspective. First, some notation. Let us assume that the data were taken according to the model equation

$$d = A(s+f) + n \tag{1}$$

where the n_d-vector d contains the data samples, the $(n_d \times n_s)$ matrix A is the observation matrix, the n_s-vector s is the (discretized) signal, the n_s-vector f represents any contaminants ("foregrounds") one has to contend with, and the n_d-vector n is the instrumental noise. I model the signal stochastically (vs. a deterministic functional form) and "infer" its covariance properties from the data. In particular, the signal is modeled through its covariance properties, encoded in $S \equiv \langle ss^T \rangle$, the signal covariance matrix.

Then I can write the Bayesian posterior as

$$P(s,f,S|d,N) \propto P(d|s,f,N)P(s|S)P(f)P(S) \tag{2}$$

where N is the noise covariance matrix $\langle nn^T \rangle$. I will now discuss the various terms in Eq. 2. The likelihood $P(d|s,f,N)$ specifies how the data is related to the quantities in the model. Given the model equation Eq. 1 specifies that $P(d|s,f,N) = G(d - A(s+f), N)$[1].

The other terms in Eq. 2 specify information about the components of the model. The term $P(s|S)$ contains information about the covariance of s. If s is a Gaussian random field with zero mean (examples from cosmology are the CMB or other probes of the density fluctuations of matter on cosmological scales) $P(s|S) = G(s,S)$. Note that it is not assumed that S is known.

Partial knowledge (or ignorance) about S is quantified in terms of the prior $P(S)$. For a stationary field $P(S)$ might simply represent the fact that I parameterize the covariance matrix in terms of power spectrum coefficients. Eq. 2 also assumes that the signal, noise and the contaminants are stochastically independent of each other. Further, the equations as written are conditioned on perfect knowledge of the noise covariance.[2]

Lastly, $P(f)$ encodes the knowledge or ignorance about foregrounds. Note that from a Bayesian perspective all that is required is that $P(f)$ accurately represents knowledge about f. Therefore assuming a Gaussian form for f does not assume that f actually has Gaussian statistics. In particular the mode of the Gaussian corresponds to the most probable (a priori) foreground model and the covariance to the uncertainty in the model. The ability to specify uncertainties in the foregrounds (which will then be taken into account when the method is applied) is a key feature of this approach which guards against biases from including incorrect foreground templates without the ability to account for the uncertainty in these templates.

Having specified the forms of the various terms on the right hand side of Eq. 2, the task is to explore the joint posterior density $P(s,f,S|d,N)$. However, traditionally the problem is treated in three different limits. If, as an expression of prior ignorance, I take $P(f) = const.$ and $P(s) = \int P(s|S)P(S)dS = const.$ then all the information is in the likelihood $P(d|s,f,N)$. In this case the best one can do if n is Gaussian, is to summarize what is known about $s+f$ in terms of the maximum likelihood estimate

$$m \equiv (A^T N^{-1} A)^{-1} A^T N^{-1} d \tag{4}$$

and quote the associated noise covariance matrix $C_N =< mm^T >= (A^T N^{-1} A)^{-1}$. In the CMB literature the process of obtaining m and C_N from the data are known as "map making."

If on the other hand, the signal covariance S is perfectly known and foregrounds are neglected then the joint posterior becomes $P(s,S|d,N) \propto P(s|d,N,S)$ where

$$P(s|d,N,S) = G(s - S(S+C_N)^{-1}m, S(S+C_N)^{-1}C_N). \tag{5}$$

This posterior for s peaks at s_{WF}, the well-known Wiener Filter reconstruction of s, so this is known as "Wiener Filtering."

In the third limit, "power spectrum estimation," one does not know S but have some information about how it is parameterized, namely that in the Fourier basis S is diagonal with the diagonal elements equal to the power spectrum coefficients C_l. If we ignore foregrounds again and set $P(S(C_l)) = const$ we can integrate out ("marginalize over") s and obtain the usual starting point for maximum likelihood power spectrum estimation

$$P(S(C_l)|d,N) = G(m, S(C_l) + C_N). \tag{6}$$

The density $P(S(C_l)|d,N)$ is considered as a multivariate function of all the power spectrum coefficients up to some band limit l_{max}. It represents all the information about $S(C_l)$ contained in the data. One can again summarize what is known about S by quoting the set of power spectrum estimates $\hat{C}_l$ for which $P(S|d,N)$ is maximum (equivalent to the maximum likelihood estimates) and include a summary of the width of the marginal distribution of $P(S|d,N)$ for each power spectrum coefficient.

[1] I use $G(x,X)$ as a shorthand for the multivariate Gaussian density

$$G(x,X) = \frac{1}{\sqrt{|2\pi X|}} \exp\left(-\frac{1}{2}x^T X^{-1} x\right). \tag{3}$$

[2] This assumption may not hold in practice and can in fact be relaxed. The resulting question whether both S and N can be usefully obtained from the data is determined by the structure of the observation matrix A.

However, in this case for any n_s larger than a few thousand this procedure is computationally prohibitive. Since the determinant in Eq. 6 depends on S, it needs to be evaluated if the shape of the likelihood is to be explored. Determinant evaluation scales as n_s^3. As a result, to evaluate Eq. 6 just once for a million pixel map would take several years, even if one achieved perfect parallelization across thousands of processors on the most powerful supercomputing platforms in the world. To find its maximum in a parameterization of 1000 power spectrum coefficients and compute marginalized confidence intervals for each C_l by integrating out all others is a lost cause.

The maximum likelihood techniques that are currently described in the literature [5, 6] avoid the determinant calculation in Eq. 6 by finding the zero of the first derivative of $P(S|d,N)$ using an approximate Newton-Raphson iteration scheme. However, for realistic data, the computational complexity is not reduced because the first derivative contains traces of matrix products that also require of order n_s^3 operations. In these treatments the error bars on the power spectrum coefficients are approximated by the second derivative of the likelihood at the peak even though the likelihood of S is non-Gaussian. This second derivative is again hard to compute, requiring of order n_s^3 operations.

Even these expensive methods do not provide a way of accurately summarizing and publishing the "data product," $P(S|d,N)$. There are various approximate techniques for doing this in the literature [7, 8] but it is not well understood how good these approximations are away from the peak of the likelihood [9].

3. METHOD: DO NOT EVALUATE, SAMPLE!

How does one overcome these computational challenges? The answer I propose is to *sample* from the full joint density $P(s,f,S|d,N)$. This may seem even more challenging, since this a function of millions arguments and general techniques of generating samples from complicated multivariate densities are very computationally intensive. However, the special structure of the Gaussian priors in Eq. 2 allows exact sampling from the conditional densities of $P(s,f,S|d,N)$. Exact sampling is made possible by solving systems of equations using the preconditioned conjugate gradient method [10]. This means the *Gibbs sampler* [11] can be used to construct a Markov Chain which will converge to sampling from $P(s,f,S|d,N)$. The Gibbs sampler is an iterative scheme for generating samples from a joint posterior density by iterating over the components of the density (such as s, S, and f) and sampling each of them in turn from their conditional distributions while keeping the other components fixed. Given a set of Monte Carlo samples from the joint posterior, any desired feature of the posterior density can be computed with accuracy only limited by the sample size.

After having obtained a sample from the joint posterior $P(s,f,S|d,N)$, it is trivial to generate samples from the marginal posteriors $P(s|d,N)$ or $P(S|d,N)$. Integration over a sampled representation of a function just corresponds to ignoring the dimensions that are being integrated over! For the problem at hand things are even better than this, since the conditional density $P(S|s)$ has a very simple analytical form. As a result, one can compute an analytical approximation to $P(S|d,N)$ using the Monte Carlo samples s_i

$$P(S|d,N) = \int ds P(S|s)P(s|d,N) \approx$$

$$(1/n_{MC}) \sum_{i=1}^{i=n_{MC}} P(S|s_i). \qquad (7)$$

This is known as the Blackwell-Rao estimator of $P(S|d,N)$ which is guaranteed to have lower variance than a binned estimator. In fact one can show that for perfect data (complete and without noise) this approximation is exact for a Monte Carlo sample of size 1! For realistic data, the approximation converges to the true power spectrum posterior given enough samples.

My collaborators and I call the approach and the set of tools we have developed to implement this approach the "MAGIC" method, since MAGIC Allows Global Inference from Correlated data. We give a detailed description of the technique in the context of CMB covariance analysis in [1]. Figure 1 shows the performance of MAGIC compared to power spectrum estimation techniques (which do not include the signal reconstruction and foreground separation features of MAGIC).The main advantages of the MAGIC method are the following:

1. Massive speed-up compared to brute force methods. For an (unrealistic) pre-factor of 1 a single n_p^3 operation would take 3×10^{10} seconds on a 1 GFlop computer. An unoptimized implementation running in the background on a desktop AthlonXP1800+ CPU currently requires less than 10^5 seconds per sample.

2. Massive reduction in memory use: since we only need to compute matrix-vector products (not matrix-matrix products, matrix inverses or determinants) only the parametrizations of the covariance matrices need to be stored (e.g. noise power spectrum for N and the signal power spectrum for S). This reduces the memory requirements from order $n_p \times n_p$ to at most order n_d which is usually many orders of magnitude less.

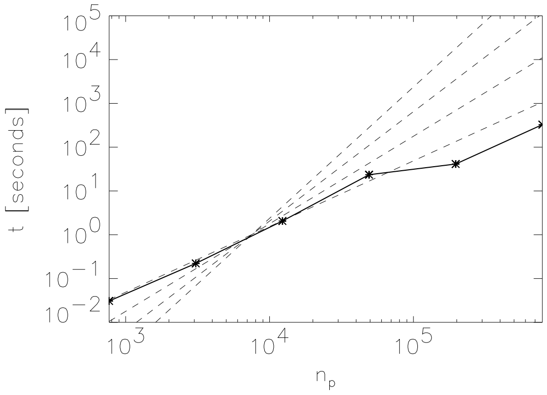


Figure 1: Average computing time (without code optimization) required for one iteration of the Gibbs sampler as a function of the number of pixels in the map. These timings are for a single AthlonXP 1800+ CPU. Solid line: actual timings. Dashed lines show n_p^x for $x \in \{3, 5/2, 2, 3/2\}$ from the top to the bottom on the right side of the figure. Brute force methods require $t \sim O(n_p^3)$ and approximate methods require $t \sim O(n_p^{(3/2)})$ computational time. For the WMAP data $n_p \sim 3 \times 10^6$ pixels.

3. Allows modeling realistic observational strategies and instruments.

4. Straightforward parallelization (run several MAGIC codes on separate processors to generate several times the number of samples in the same time).

5. Allows treating the statistical inference problem globally, that is it keeps the full set of statistical dependencies in the joint posterior given the data.

6. Generalizes Wiener Filter signal reconstruction to situations where the signal covariance is not known a priori but automatically discovered from the data at the same time as the actual signal is reconstructed.

7. Allows computing marginal credible intervals, either for individual power spectrum estimates or for combinations of any set of dimensions in the very high dimensional parameter space.

8. Allows incorporating uncertainties (e.g. about the foregrounds) in the analysis in such a way that they are propagated correctly through to the results.

9. Makes it possible to build in physical constraints in a straightforward way.

10. Generates an unbiased functional approximation to $P(C_l|d)$, as shown in Eq. 6. It has the advantage of being a controlled and improvable approximation and removes the need for parametric fitting functions such as the offset lognormal or hybrid approximations.

11. Generates a *sampled* representation of the joint posterior Eq. 2, which simplifies further statistical analyses.

Since MAGIC is a Markov Chain method, one also has to discuss the issue of burn-in and correlations of subsequent steps in the chain. Steps in the power spectrum coefficients C_l are proportional to the width of the perfect data posterior [1]. In other words, the number of steps it takes to generate two uncorrelated power spectrum samples is proportional to $(S/N)_l^2$ where $(S/N)_l$ is the rms signal to noise ratio for the l^{th} power spectrum coefficient. Conveniently, the samples are nearly uncorrelated over the range in l where the data is informative. In numerical experiments with the WMAP data it took about 15-20 steps for the chain to burn-in (for the range in l where $(S/N)_l \sim 1$ or greater) from a wildly wrong initial guess of the power spectrum ($C_l = const.$).

4. EXAMPLE APPLICATIONS TO THE COSMIC MICROWAVE BACKGROUND

In the online materials for this talk [4] I present the results of applying the MAGIC method to a synthetic data set which covers an unsymmetrically shaped part on the celestial sphere. I used MAGIC to reconstruct the signal on the full sky and to make movies of the Gibbs sampler iterations. This is an example where the signal is automatically discovered in the data by the algorithm, without specification of the signal covariance.

Figures 2 and 3 show the results of analyzing the COBE-DMR data [12], one of the most analyzed astronomical data sets. This allowed us to perform consistency checks between the MAGIC method, other methods and the recent results from the Wilkinson Microwave Anisotropy Probe (WMAP) [13].

I am also very interested in evaluating claims that the WMAP data favors theories which predict a lack of large scale fluctuation power in the CMB. This claim, if true, would have far-reaching consequences for our understanding of the Universe. Since cosmologists only have one sky to study, we have to be very careful to account for our limited ability to know the ensemble averaged power spectrum on large scales. The WMAP team estimated the fluctuation power on large scales using several techniques and consistently found it to be low. However, in all of these techniques, the variance of the estimates was computed in an approximate way (e.g. in terms of the curvature at the peak) and

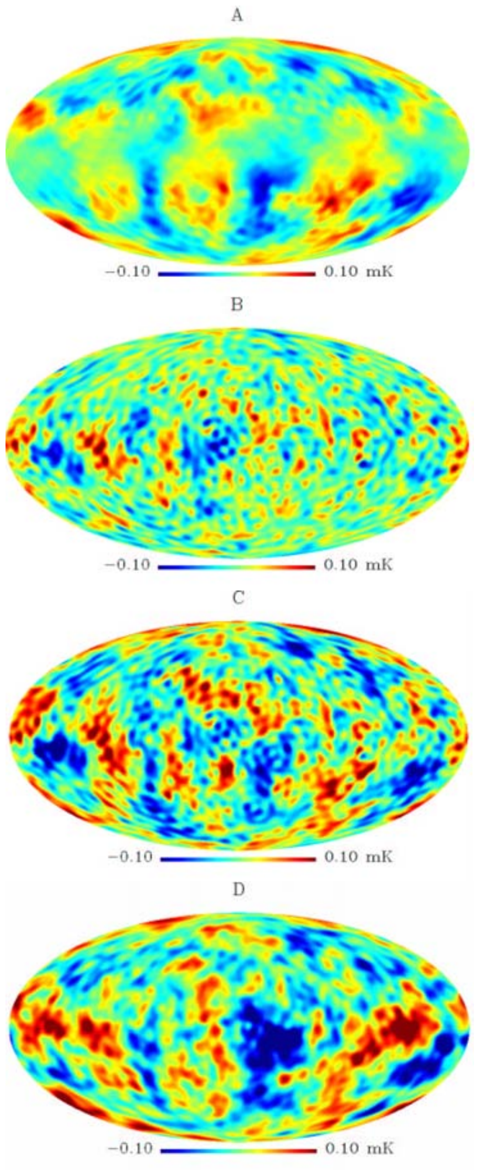


Figure 2: Reconstructed signal maps in Galactic coordinates. A: The posterior mean signal map ($\int dssP(s|d)$) for the COBE-DMR data. This is a generalized Wiener filter which does not require knowing the signal covariance a priori. B: One sample drawn from the conditional posterior $P(s|C_l, f, d)$. The posterior mean signal map, shown in panel A, has been removed. C: The sample pure signal sky at the same iteration. This is the pixel-by-pixel sum of the maps in panels A and B. D: The WMAP data smoothed to 5 degrees (less than A, more than C). The corresponding features in parts A and D are clearly visible.

relies on theory for the assessment of statistical significance. Using MAGIC one can easily integrate over the posterior density of the power spectrum given the data. Therefore it is easy to compute the probability for the power spectrum coefficients in any given l-range to be smaller than any given value.

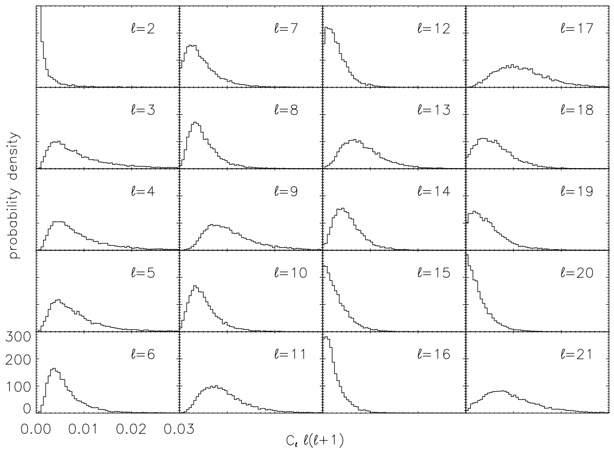


Figure 3: Marginal posterior densities for each individual C_ℓ from the COBE-DMR data. At each ℓ the fluctuations in the C_ℓ at all other ℓ were integrated out. The axis ranges are the same for all panels.

Using the MAGIC method it was straightforward to generate a preliminary sample of the power spectrum coefficients from the WMAP posterior using only the W1 channel, one of the cleanest channels in the WMAP data, in terms of systematic error estimates. For the cleaned W1 data and masking regions of galactic emission (mask *Kp0* in the WMAP data release) the quadrupole and octopole power is not obviously discrepant from theoretical expectations. Choosing a more aggressive mask could change this since that reduces the sampling variance. One should bear in mind that the power spectrum likelihood $P(C_l|d)$ has infinite variance for $l = 2$ even for perfect all-sky data, unless a prior is put on C_2's value. Therefore, in an exact assessment of the quadrupole issue claims of a significant discrepancy ought to be based on the actual shapes of posterior density, not a chi-square test (compare the detailed discussion of cosmic variance in [1]). I will address the issue of low power in the low cosmological multipoles in a future publication.

5. FUTURE DIRECTIONS

Of course, if desired, additional prior information about our Universe can be added to the analysis. For example instead of viewing the power spectrum as the quantity of interest, its shape could be parameterized as a function of the ~ 10 cosmological parameters which span the space of cosmological theories. Then instead of sampling from the power spectrum coefficients given the signal, one would run a short Metropolis-Hastings Markov chain at each Gibbs iteration to obtain a sample from the space of cosmological parameters given the data. These parameter samples, in turn define a density over the space of power spectra with considerably tighter error bars. The result is the non-linearly optimal filter for reconstructing the mean of the power spectrum incorporating physical information about the origin of the CMB anisotropies.

Another important direction is the analysis of image distortions. The treatment as detailed so far does not allow for the CMB to be lensed gravitationally by the mass distribution through which it streams on its way to us. This distortion itself contains very valuable cosmological information. Extending the formalism to account for lensing of the CMB and estimate the statistical properties of the lensing masses from the lensed CMB would be an important extension of this approach.

Acknowledgments

I thank my students and collaborators Arun Lakshminarayanan, David Larson, and Ian O'Dwyer, as well as Tom Loredo for his suggestions. This work has been partially supported by the National Computational Science Alliance under grant number AST020003N and the University of Illinois at Urbana-Champaign.

References

[1] Wandelt, B.D., Larson, D., and Lakshminarayanan, A., astro-ph/0310080, PRD submitted.

[2] N. Wiener, "Extrapolation, Interpolation, and Smoothing of Stationary Time Series with Engineering Applications," MIT Press, Cambridge, MA,1949.

[3] G. B. Rybicki and W. H. Press, "Interpolation, Realization, and Reconstruction of Noisy, Irregularly Sampled Data," ApJ 398, 169 (1992)

[4] PowerPoint slides of this talk and two AVI movies are at `http://www-conf.slac.stanford.edu /phy-stat2003/talks/wandelt/contributed/`

[5] M. Tegmark, Phys. Rev. D55, 5895 (1997)

[6] J. R. Bond, A. H. Jaffe, and L. Knox, Physical Review D 57, 2117 (1998)

[7] J. R. Bond, A. H. Jaffe, and L. Knox, Astrophys. J. 533, 19 (2000)

[8] L. Verde *et al.*, Ap.J.Suppl. 148, 195 (2003)

[9] Lewis, A., astro-ph/0310186

[10] William H. Press, *et al.*, *Numerical recipes*, Cambridge University Press, Cambridge, UK. (1992)

[11] Tanner, *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*, Springer Verlag, Heidelberg, Germany. (1996)

[12] C. L. Bennett *et al.*, Ap. J. 464, L1 (1996)

[13] C. L. Bennett *et al.*, Ap.J.Suppl. 148, 1 (2003)

Comments on Likelihood Fits with Variable Resolution

Giovanni Punzi
Scuola Normale Superiore and INFN, 56100 Pisa, Italy

Unbinned likelihood fits are frequent in Physics, and often involve complex functions with several components. We discuss the potential pitfalls of situations where the templates used in the fit are not fixed but depend on the event observables, as it happens when the resolution of the measurement is event–dependent, and the procedure to avoid them.

When several categories of events are present in the same data sample, an unbinned Maximum Likelihood fit is often used to determine the proportion and the properties of each class of events. This procedure makes use of "templates", representing the probability distribution of the observables used in the fit for each class of events. In the simplest cases the templates are completely determined by the values assigned to the parameters of the fit, but frequently a more sophisticated approach is chosen where templates vary on an event by event basis, according to the resolution of the measurement for that particular event. These variations are due to the dependence of resolution on extra variables, that change on an event-by-event basis . This may happen, for instance, when events are recorded by a detector that has different resolutions in different regions within its acceptance.

A common example of this kind of fit in HEP is given by lifetime and/or mass fits (see [1] for a sample list of recent experimental papers), where variations in resolution occur as a consequence of different configuration of each individual decay. The same kind of issue however is likely to arise in other situations.

The purpose of this short paper is to point out some potential pitfalls in this kind of fitting procedure. I will illustrate the point with reference to a simple toy problem.

1. A TOY PROBLEM

Consider an experiment in which two types of events, A and B, can occur. Let f be the fraction of type–A events, that is, the probability of a generic event to be of type A. We want to extract a measurement of f from a given sample of data. In order to do this, we measure the value of an observable x, having the following probability distributions:

$$p(x|A) = N(0, \sigma)$$
$$p(x|B) = N(1, \sigma)$$

Where σ is a known constant and $N(\mu, \sigma)$ is the normal distribution

This problem is easily solved using an "unbinned Likelihood fit". This consists of maximizing the Likelihood function:

$$L(f) = \prod_i (fN(x_i, 0, \sigma) + (1 - f)N(x_i, 1, \sigma)) \quad (1)$$

with respect to the required parameter f (here $N(x, \mu, \sigma)$ indicates the gaussian function in the variable x). This is very simple to perform with the help of a numerical maximization program.

Let's make a specific numeric example, where $f = 1/3$ and $\sigma = 1$ (see illustration in Fig. 1), and the size of the data sample is 150 events. By repeatedly generating MC samples of 150 events each, we obtain the distribution of the Maximum Likelihood estimator of f, which is shown in Fig. 2.

Its mean is 0.3368 ± 0.0041 and $SD = 0.083$, in agreement with expectations of 0.3333 and 0.088 respectively (the latter coming from Fisher information).

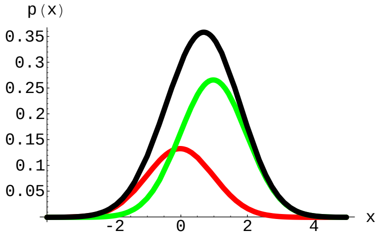


Figure 1: Probability distribution of x for the toy problem described in the text. Contribution of type–A and type–B events are also shown.

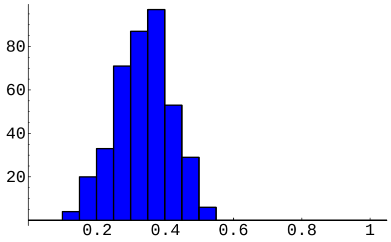


Figure 2: Distribution of ML estimate of the fraction f of type-A events (see text)

2. A TOY PROBLEM, WITH VARIABLE RESOLUTION

Let's now suppose that the resolution of x is not constant, but rather depends on the event: we are assuming that each event x_i comes together with an individual value of σ (let it be σ_i). This situation is encountered in many real–life problems, and the common approach found in the literature is to simply modify the Likelihood function as follows:

$$L(f) = \prod_i fN(x_i, 0, \sigma_i) + (1 - f)N(x_i, 1, \sigma_i)) \quad (2)$$

This looks like a pretty obvious generalization of expression (1). To test it in our toy problem, we modified our toy MC from previous example, by making σ fluctuate at each event within an arbitrarily chosen range (1.0 to 3.0), and again made repeated simulated experiments of 150 events each, maximizing the Likelihood expression (2) to estimate f. The result of this test is shown in Fig. 3, and rather surprisingly, shows a very large bias with respect to the true value of f.

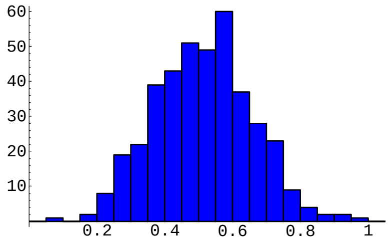


Figure 3: Distribution of ML estimate of the fraction f of type-A events, obtained from a "conditional Likelihood"

This may seem really odd, until one realizes that this new problem is very different from the previous one. Our problem now has actually two observables: each observation consists of the pair of values (x_i, σ_i) rather than just x_i, and its probability density depends on both. This means that the Likelihood must now be written based on the probability distributions of the (x_i, σ_i) *pair*:

$$L(f) = \prod_i fp(x_i, \sigma_i|A) + (1 - f)p(x_i, \sigma_i|B) \quad (3)$$

Remembering that $p(x_i, \sigma_i|X) = p(x_i|\sigma_i, X)p(\sigma_i|X)$ we can write the correct expression of the Likelihood for our problem as:

$$L(f) = \prod_i fN(x_i, 0, \sigma_i)p(\sigma_i|A) + (1 - f)N(x_i, 1, \sigma_i)p(\sigma_i|B) \quad (4)$$

where $p(\sigma_i|X)$ is the *pdf* of σ_i for events of type X, an element that was absent in eq. (2); in fact, comparing the two expressions shows that (2) is actually the *conditional* probability distribution $p(x_i|\sigma_i, f)$ (one might call it "conditional Likelihood") rather than the full distribution $p(x_i, \sigma_i|f)$. The difference matters for fitting unless it happens that the distribution of σ_i is the same for all types of events: $p(\sigma_i|A) = p(\sigma_i|B)$. In that case, $p(\sigma_i)$ can be factored out, and the incomplete Likelihood of eq. (2) differs from the true Likelihood just by a factor independent of f, that does not affect the maximization.

In the specific MC test reported above, we simulated a resolution 1.5 times worse for events of type B than for type A, setting the σ_i distribution as flat between 1 and 2 for A-type events, and flat between 1.5 and 3 for type-B events. We intentionally avoided saying this explicitly before, in order to put the reader in the typical situation encountered in reality, where no attention is payed to the distribution of those resolutions for the different classes of events considered in the fit. It turns out from our example that this may lead to very biased results.

In summary, expression (2) simply does not work for fitting, and by a large amount: it can be said to belong to that particular class of solutions nicely defined in [2] as 'SNW solutions'.

Conversely, if we use in fitting the correct expression of the Likelihood (eq. 4) we get the result shown in fig. 4, showing a negligible bias. The resolution of the fit is also much better, as the difference in the distributions of the σ themselves gets exploited in separating the two samples; this however is a minor point in comparison with the bias issue.

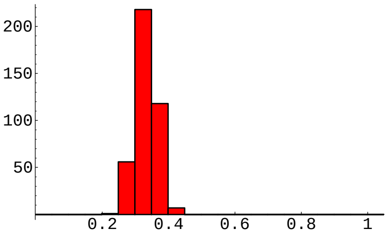


Figure 4: Distribution of ML estimate of the fraction f of type-A events, using the full Likelihood function

3. ADDITIONAL TESTS

One may wonder at what features of the distributions make for a large bias. Table I shows results for a few variants of the original problem. Tests include:

- Equal-width ranges of σ.
- Disjoint σ ranges.
- Constant, but different σ for A and B.
- Constant, and close σ's for A and B.
- Same-mean σ distribution with different widths.

Table I Results of MC fitting tests.

Resolutions		"conditional" L (2)		True Likelihood	
σ_A	σ_B	$\tilde{f}_A$	$\sigma(\tilde{f}_A)$	$\hat{f}_A$	$\sigma(\hat{f}_A)$
1.0	1.0	0.336 ± 0.003	0.08		
$[1.0, 2.0]$	$[1.5, 3.0]$	0.514 ± 0.007	0.14	0.335 ± 0.002	0.03
$[1.0, 2.0]$	$[1.5, 2.5]$	0.474 ± 0.007	0.14	0.335 ± 0.002	0.03
$[1.0, 2.0]$	$[2.0, 3.0]$	0.579 ± 0.008	0.15	0.333 ± 0.000	0.00
1.0	2.0	0.645 ± 0.006	0.12	0.333 ± 0.000	0.00
1.0	1.1	0.374 ± 0.004	0.08	0.333 ± 0.000	0.00
$[0.5, 3.5]$	$[1.5, 2.5]$	0.330 ± 0.006	0.12	0.332 ± 0.002	0.03
1.0	$[1.0, 2.0]$	0.482 ± 0.009	0.09	0.333 ± 0.000	0.00
(σ_A actually = 1.)		modified L (5)		True Likelihood	
1.0	$[1.0, 2.0]$	0.374 ± 0.004	0.08	0.333 ± 0.000	0.00
$[0.5, 3.5]$	$[1.0, 2.0]$	0.414 ± 0.004	0.08	0.332 ± 0.003	0.03

- Only one type of events has variable sigma.

In almost every tried situation we found expression (2) to return largely biased results. The exception occurs when the average σ is the same; the resolution on f is however much worse than with the correct expression. It looks like the most important element is the difference between the average values of σ for the different samples; the actual variability within each sample seems less important.

A simpler situation exists, that is pretty common in practice, where one has just one signal component over a background, and the signal distribution contains a variable sigma, while the background is represented just by a fixed template. In this case, expression (2) becomes:

$$L(f) = \prod_i fN(x_i, 0, 1) + (1 - f)N(x_i, 1, \sigma_i)) \qquad (5)$$

This expression of L is of course still incorrect, but it better describes reality at least for one of the two event categories by incorporating explicitly the information that it has a fixed sigma. Here a variable template appears just in one component, and being this the simplest configuration with a variable template, it is interesting to ask whether it yields a reasonable approximation of the correct results.

If we apply this new Likelihood expression to the last tested case, ($\sigma_A = 1.0$ and $\sigma_B \in [1.0, 2.0]$), we find that the result is still biased, although to a lesser extent (Tab. I). This shows that the distribution of σ must be kept into account even in the simplest situation, where it appears in only one component of the fit.

The mechanism underlying this problem is easier to see by looking at a variant of the previous case. Suppose that resolutions are the same as above, but for events of type-A the variable σ_i is distributed over a wide range (0.5-3-5); this is not the actual value of the resolution for those events, that is still fixed at 1, so for type-A events it represents just an additional meaningless number. This is a definite possibility in a real case, where the nature of type–A events may be so different from type–B to produce meaningless values for the resolution estimator σ_i, that was designed to work for type–B events – remember that the distribution of A is given as fixed. Note that the expression used (5) does know that much, and correctly disregards the value of σ_i in the A hypothesis. For events of type B, the variable σ_i correctly represents the sigma, event by event, of the observable x, and the L function correctly accounts for this, too.

It may come as a surprise that the result is largely biased. The reason for this rather spectacular failure is that the second piece of L, related to B-type events, gets confused by the presence of the events of type–A with meaningless values of sigma: they unavoidably enter both terms of L during the calculation. The conclusion is: whenever you include σ_i in your Likelihood expression, even for just *one* class of events, you must also account for its distribution, and you must do so for *all* event classes.

4. CONCLUSIONS

Whenever the templates used in a multi-component fit depend on additional observables, one should always use the correct, complete Likelihood expression (4), including the explicit distributions of all observables for all classes of events. This is necessary even if just one of the components is based on a variable σ. The simpler expressions that are commonly used should be considered unreliable unless one can show that the distribution of the variable σ is the same for all components.

A more general consideration suggested by the examples discussed above is that one should always be wary of "intuitive" modifications of a Likelihood function. For every given problem there is only one correct expression for the Likelihood (up to a multiplicative constant factor), and it is crucial to verify in every case that the expression used is the right one, rather than rely on intuition.

References

[1] Recent examples of variable-resolution fits in HEP can be found in: B. Aubert et al. [Babar Collab.], Phys. Rev. Lett. 91 (2003) 121801. K.Abe et al. [Belle Collab.], Phys. Rev. Lett. 88 (2002) 171801. P. Abreu et al. [Delphi Collab.] , Eur.Phys. J. C16 (2000) 555. ALEPH Collab., Phys.Lett. B492 (2000) 275-287. M. Paulini, Int. J. Mod.Phys. A14 (1999) 2791-2886.

[2] J. Heinrich, these proceedings.

Setting Confidence Intervals in Coincidence Search Analysis

L. Baggio and G.A. Prodi
INFN and Trento Univ., Povo, TN 38050, Italy

The main technique that has been used to estimate the rate of gravitational wave (gw) bursts is to search for coincidence among times of arrival of candidate events in different detectors. Coincidences are modelled as a (possibly non-stationary) random time series background with gw events embedded in it, at random times but constant average rate. It is critical to test whether the statistics of the coincidence counts is Poisson, because the counts in a single detector often are not. At some point a number of parameters are tuned to increase the chance of detection by reducing the expected background: source direction, epoch vetoes based on sensitivity, goodness-of-fit thresholds, etc. Therefore, the significance of the confidence intervals itself has to be renormalized. This review is an insight of the state-of-the-art methods employed in the recent search performed by the International Gravitational Event Collaboration for the worldwide network of resonant bar detectors.

1. INTRODUCTION

When a detector is pushed to its limits in order to reveal faint sources, every slight deviation of noise models from ideality can severely jeopardize the robustness of a detection claim. In fact, when the signal-to-noise ratio (SNR) is low, most goodness-of-the-fit tests have poor discrimination power. On the other hand, in the long run, the outliers add up and constitute a background which can be much larger than the isolated signals possibly present in the data.

Working with a network of detectors optimized for coincidence analysis allows to reduce the background and –most of all– to estimate reliably the background itself, which is essential to set reliable upper limits.

A gravitational wave (gw) resonant detector is built around a mechanically isolated massive resonant body. Cylindric 3m-long 2.3 ton aluminum alloy bars have been until now a widely adopted solution. Any planar (transverse) gravitational wave impinging on the bar with an angle θ relative to its axis excites the longitudinal mechanical mode, with amplitude proportional to $\sin^2\theta$. With respect to burst signals, the presently working resonant detectors are sensitive in a narrow ($\sim 1 - 10Hz$) frequency range near the resonance ($\sim 900Hz$). [3]

A candidate event is defined as the output of an automated max-hold algorithm based on two adaptive thresholds: one on the SNR of the peak amplitude (it has to be great enough to be identified without ambiguity, i.e. low timing error) and one on the minimum delay between consecutive events (in order to generate independent events it must be greater than a few times the autocorrelation of the processed data). Even with no outliers, this algorithm would produce random accidental events as samples of the extreme distribution for an (almost) Gaussian stochastic process.

The International Gravitational Events Collaboration (IGEC) [1–3] was founded in order to take up the task of assessing the detection of gw's from the candidate event lists compiled by the single detectors. The only requirement for member groups has been that the exchanged information should include:

i) event amplitudes and times of arrival (along with their estimated errors)

ii) minimal detectable amplitude –i.e. the sensitivity threshold of the detector– defined by requirement of unbiasedness of amplitude estimates and unambiguous timing.

The IGEC analysis is based on time coincidence search, and in the first 4 year run (1997-2000) the five detectors of the collaboration were purposely aligned to be as parallel as possible, in order to maximize the efficiency of the network. The analysis as it was recently performed is still not optimal in many respects. Moreover, because the gw source amplitude distribution and polarization are unknown, the detection efficiency is not completely determined. However, with respect to past and recent proposals, this analysis improves the control of probability of false dismissal of candidate gw signals and provides the detailed computation of the probability of accidental detection.

In the IGEC analysis many selections and tests are applied to the data, in order to enhance the chances of gw detection as a function of the amplitude and direction of target gw signals. The selections may enhance accidental detections as well, therefore a record of all the attempts has to be compiled. When a complete account is given for all the operations on the data, and assuming that their statistics is known, the probability that any of the observed results is due to chance can be well accounted for within the frequentist framework.

2. DATA CUTS AND COINCIDENCE SEARCH

Hereafter the focus will be a source located in the direction of the galactic center, as it is likely that the present sensitivity of bar detectors limits the observation range to sources within the Milky Way. The times of arrival are supposed to be already corrected for the light travel time delay for detectors at different positions. Moreover, as discussed in Fig. 1, the measured amplitude of events has been corrected for the angular sensitivity factor.

A (twofold) *coincidence* is defined when the time of arrival t_i and t_j of two events from different detectors satisfies the inequality

$$|t_i - t_j| < k\sqrt{\sigma_i^2 + \sigma_j^2} \tag{1}$$

where σ_i and σ_j are the standard deviation of the time error, and k depends on the target false dismissal. The timing error is not Gaussian, and its standard deviation is strongly dependent on the signal-to-noise ratio of the event amplitude (it ranges from a second down to milliseconds; see for instance Fig. 1 in Ref. [4]). A conservative value for k is given by the Bienaymè-Tchebycheff inequality: the probability that the absolute value of a zero mean random variable is greater than k times its standard deviation σ is $P(|x| > k\sigma) \leq k^{-2}$. For instance, $k \sim 4.5$ guarantees a false dismissal less than 5%.

In general, an M-fold coincidence is defined as the simultaneous coincidence in the $M(M-1)$ distinct couples out of M detectors. In this case, for a target false dismissal probability P_T, one has to set $k = (1 - (1-P_T)^{2/[M(M-1)]})^{-1/2}$. As for the rate of accidental coincidences, it is proportional to k^{M-1} and to the rate of events in each individual detector[4].

The IGEC adopted the following data selection scheme (see Fig. 1):
i) fix a common (absolute) threshold A_{th};
ii) cut the time spans when the minimal detectable amplitude of each detector was greater than A_{th};
iii) within these periods, include only those events with amplitude greater than A_{th}.

We investigated different results from many values of A_{th}, and consequently we accounted for the increased probability of false alarm[1] (see Sec. 4).

3. BACKGROUND ESTIMATE

The IGEC uses resampling methods to estimate the rate of uncorrelated background coincidences. Approximately randomized samples of the coincidence counts can be obtained by rigidly shifting the times of arrival of the original event time series of individual detectors relative to each other. With this new data set, the whole analysis is repeated: amplitude modulation, data selection and coincidence search.

The choice of a rigid time shift instead of reshuffling or swapping is due to the presence of structures in the autocorrelation of the single detector event time series, with characteristic timescales from a few seconds to one minute (see Fig. 8 in Ref. [3]) –i.e. the time series are not Poisson. Moreover, the angular modulation and the common amplitude thresholding applied to the data conspire to produce further event clustering (see Fig. 1). A rigid time shift guarantees that all these structures are not smoothed out when generating resampled counts.

In order to obtain *independent* resampled counts, the time series were always shifted more than the maximum *time window* (i.e. the right side of Eq. 1) ever used (in practice, few seconds).

To test that the resampled counts come from the same statistic, and that the latter is Poisson, the histograms of coincidence counts were fitted with a Poisson probability density profile. The one-tail χ^2-test has been performed on every network configuration (provided that at least one degree of freedom was available), and the histogram of the computed p-levels was in agreement with uniform density, which is the expected one if the model of the background statistic is good.

Strictly speaking, what has been verified is just the coherence of the resampling approximation –all resampled counts due to the same statistic. This result holds up to timescales of the order of one hour, which translates in a few thousands of independent resampled coincidence counts. The statistical error for the resampled background rate is then about 3%.

However, in order to conclude that the resampled statistics is also identical to the statistic of the unshifted original data, one has to be confident that no source of correlated background events exists. This ansatz is assumed without proof.

4. CONFIDENCE INTERVALS

The results of IGEC search are frequentist, i.e. the quoted confidence level or coverage are meant to be –at least conservatively– the probability that the confidence interval contains the true value. This approach is also unified in that it prescribes how to set a confidence interval automatically leading to a claim of detection or an upper limit. The construction of the confidence belt however does not proceed *à la* Feldman and Cousins [5], or proposed modifications, where the coverage is kept as fixed as possible for any source strength. Instead, the confidence interval bounds are independently derived from the likelihood function. This inevitably leads to variable coverage, and we shall show briefly how the minimum of the coverage is related to the integral of the likelihood.[6, 7]

Let $N_c = N_b + N_\Lambda$ where N_c are the counted coincidences, N_b those due to background, N_Λ those due to

[1] Actually, in Refs. [2, 3] it is a common practice to perform the analysis separately on disjoint subsets of the data, each one pertaining to a different configuration of the network –i.e. different combinations of detectors in common operation. Eventually, the data are re-aggregated per equal amplitude threshold.

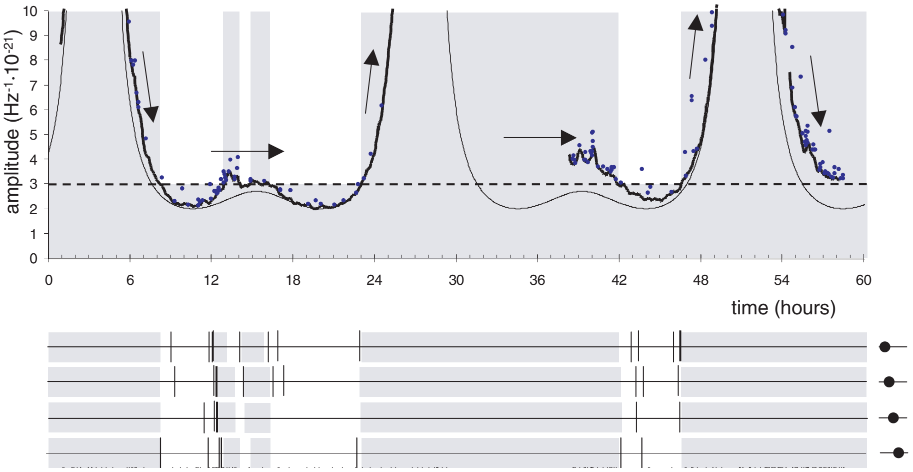


Figure 1: (*above*) Example of data selection in a time span of a few hours. The amplitude is given in terms of spectral density of the gw strain at a frequency about $900Hz$, assuming one specific direction (galactic center), and neglecting polarization. From the original event time series (*dots*) after angular sensitivity modulation (*solid smooth curve*) only those are retained whose amplitude is above a fixed common absolute threshold (*dashed line*). Correspondingly, the periods when the local detector threshold (*solid crispy curve*) is above the common threshold are removed from the observation time (*vertical solid shadows*). This generates the "on-source" time series (*below, top row*). To obtain resampled (and "off-source") event selections (*below, under the first row*), the local time coordinate at the detector site is shifted by a proper amount (*arrows*). It is worth noticing that the background event density drops exponentially toward greater amplitudes. The density of event amplitude *relative to the local threshold* is more or less the same at all times, but *relative to the fixed common threshold* it is highly nonstationary. In fact almost all events are cut out by the selection mechanism except for when the local threshold approaches closely the common threshold from below –i.e. near the edges of the live time spans. The angular sensitivity modulation (which is similar in parallel detectors) enhances this mechanism of artificial clustering, and it generates a remarkable cross-correlation of event rates between detectors. The described resampling procedure preserves the correlation pattern.

a hypothetical flux of gw's with mean rate Λ; let also μ_c, μ_b and μ_Λ be their mean values, respectively. The probability density function under the hypothesis of a Poisson statistic for both N_b and N_Λ is

$$f(N_c;\mu_\Lambda,\mu_b) = \frac{e^{-(\mu_b+\mu_\Lambda)}}{N_c!}\left(\mu_b+\mu_\Lambda\right)^{N_c} \qquad (2)$$

and the likelihood function is defined as usual as $\ell(\mu_\Lambda; N_c, \mu_b) \equiv f(N_c;\mu_\Lambda,\mu_b)$. Let I be a parameter from 0 to 1; one has to solve for $0 \leq N_{\mathrm{inf}} < N_{\mathrm{sup}}$ the equations

$$\begin{cases} \ell(n_{\mathrm{inf}}; N_c, \mu_b) = \ell(N_{\mathrm{sup}}; N_c, \mu_b) \\ N_{\mathrm{inf}} = \max(n_{\mathrm{inf}}, 0) \\ I = \left[\int_0^\infty \ell(\mu; N_c, \mu_b) d\mu\right]^{-1} \int_{N_{\mathrm{inf}}}^{N_{\mathrm{sup}}} \ell(\nu; N_c, \mu_b) d\nu \end{cases} \qquad (3)$$

The interval for μ_Λ, delimited by N_{inf} and N_{sup}, maximizes the integral of the likelihood in the physical domain $\mu_\Lambda \geq 0$, hence it belongs to a set which can be derived by a Bayesian procedure assuming constant prior for $\mu_\Lambda \geq 0$. However, we would give to this intervals frequentist interpretation, by computing the coverage

$$C(\mu_\Lambda) \equiv \sum_{N_c | N_{\mathrm{inf}} < \mu_\Lambda < N_{\mathrm{sup}}} f(N_c;\mu_\Lambda,\mu_b) \qquad (4)$$

The sum runs over the possible outcomes N_c for which the interval $N_{\mathrm{inf}} - N_{\mathrm{sup}}$ covers the given value of μ_Λ. The coverage depends on μ_Λ, hence to be conservative we refer to the coverage C_{min} at the least covered value of μ_Λ: $C_{min} \equiv \min_{\mu>0} C(\mu)$. In Fig. 2 the relation between I and C_{min} has been computed numerically, for various values of μ_b.

The choice of this procedure for IGEC analysis was first announced in Ref. [4], but in that paper the effective coverage of the procedure is not pointed out. Ref. [6] describes the same approach, but it also suggests *ad hoc* modifications to improve the relation between the coverage and the integral of the likelihood. We think that this modification could jeopardize robustness, in particular when errors in the estimated background are not completely negligible. Fig. 3 in Ref. [4] shows a sample confidence belt originating from this method, and the uncertainty on confidence interval bounds due to uncertainty on N_b.

When N_{inf} and N_{sup} have been computed, one divides them by the length of the selected observation time, obtaining the bounds Λ_{inf} and Λ_{sup} on the flux of gw bursts whose *measured amplitude* is above the common threshold. This limit is obviously cumulative, as lower flux is expected at higher thresholds. The details on how to unfold the results in terms of the *true amplitude* go beyond the scope of this paper.

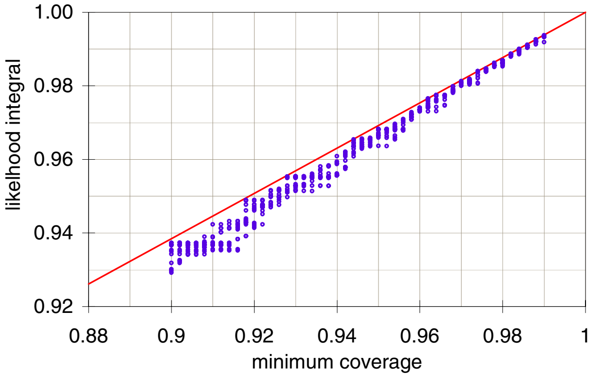


Figure 2: Integral of Poisson likelihood I vs minimum coverage of μ_Λ, for various choices of the background: $\mu_b \in \{0.01, 0.02, 0.05, 0.1, ..., 20, 50\}$. For any chosen value of I and μ_b, each dot was obtained by scanning a range of source rates, computing the coverage at each one and then taking the minimum. The relation between I and μ_b depends weakly on the background and is approximately linear.

Many selection thresholds were tried, and all of these selections happened to be independent, as we shall say in a moment. As a result, the coverage of a single confidence interval does not tell the whole story. On one hand, a confidence interval set at lower selection threshold reinforces the confidence of the exclusion region resulting from a higher threshold where the exclusion regions overlap. On the other hand, even if there are actually no true gw events, after many trials a confidence interval excluding $\Lambda = 0$ will eventually come out accidentally, as the coverage probability for $\mu_\Lambda = 0$ –i.e. $C(0)$– is not 1. This would lead to falsely reject the null hypothesis.

In order to compute correctly the probability of false claim (defined as *at least* one interval not containing $\Lambda = 0$) two methods were investigated.

First, if one assumes that the measures coming from different selections are independent random variables, then the probability of an accidental claim in case the null hypothesis is true is given by $1-\prod_i C^{(i)}(0)$, where the index i runs over all different data selections. Notice that in the Poisson case $C^{(i)}(0) > C^{(i)}_{min}$ always, and $C^{(i)}(0)$ *depends* on the background $\mu^{(i)}_b$.

Another method, which requires less assumptions, consists in resampling the entire list of results using the same randomizing procedure described above. In other words, the confidence intervals are computed on time-shifted data, for which we do not expect any genuine disagreement with the null result. From the resampled population of the would-be claims one can compute directly the chance of false alarm.

The two methods gave consistent results, which is in turn an evidence for independence of the different data selections[2].

In this way the interpretation of the measure has two layers. We start from the bare confidence intervals, and count the ones which individually would deny the null hypothesis. Then we compare this number with the expected false claims. In the end, we get a confidence interval on the number of *true* claims –if it includes zero, then we assess that no significant deviation from the null hypothesis was observed.

As a final remark, one should be aware that the number of papers quoting “95%” results *just in the gw search field* has grown such that it would not be surprising to find a positive result among them by chance. If a sequence of negative results has just been observed, the first false positive is coming from the last –supposedly better– experiment. It is really tempting to forget about the many previous null attempts (even easier if they were not published). However, a similar configuration can be just accidental (and much more than 5% likely). This should be kept in mind when hurrying to claim the first non-null result in a series of many independent attempts –it is perhaps advisable to wait until it has been confirmed by successive experiments. Another solution would be to quote “99%” (or higher) confidence results, which give lower probability of a false claim. But this is not always possible because of limitations in the degree of accuracy of the noise models (in our case, it would require a more powerful test on the tails of the density function of N_c).

Acknowledgments

The authors wish to acknowledge the useful discussions within IGEC, and with J. Linnemann.

References

[1] http://igec.lnl.infn.it/
[2] B.Z. Allen *et al*, *Phys. Rev. Lett.* **85** (2000) 5046
[3] P. Astone *et al*, *Phys. Rev. D* **68** (2003) 022001
[4] L. Baggio *et al*, *Class. Quantum Grav.* **19** (2002) 1541
[5] G.J. Feldman and R.D. Cousins, *Phys. Rev. D* **57** (1998) 3873
[6] P.B. Roe and M.B. Woodroofe, *Phys. Rev. D* **63** (2000) 090013
[7] F. Porter, *Nucl. Instr. Meth. A* **368** (1996) 793

[2] Of course, this depends on the coarseness of the chosen stepping for the common amplitude threshold. With finer steps one would expect correlation between nearby selections.

Optimal Use of Information to Measure Top Quark Properties

M. F. Canelli
University of Rochester, Rochester, NY 14627, USA

We present a method developed at DØ for extracting information from data through a direct calculation of a probability for each event. This probability, which is a function of any parameter of interest, is calculated by convoluting the differential cross section with the resolution and acceptance of the detector. The method is used to remeasure the mass of the top quark and to extract the fraction of longitudinal polarized W bosons in the lepton + jets $t\bar{t}$ sample, previously collected by the DØ experiment during Run I of the Fermilab Tevatron. The new method yields a top mass of $M_{top}(preliminary)$=180.1±3.6 (stat) ± 4.0 (syst) GeV/c^2, which corresponds to a significant reduction in the uncertainty on M_{top}. Assuming Standard Model coupling in the tbW vertex, we also extract the fraction of longitudinal W decays as $F_0(preliminary)$=0.56±0.31 (stat) ± 0.04 (syst).

1. INTRODUCTION

In proton-antiproton collisions top quarks are produced primarely in pairs, either via $q\bar{q}$ or gg fusion. At the Tevatron, the main contribution to the $t\bar{t}$ yield is from $q\bar{q}$ annihilation. This is purely the result of the fact that the parton distribution functions (PDFs) favor this channel at Run I $\sqrt{s}$=1.8 TeV and Run II $\sqrt{s}$=1.96 TeV. In fact, about 90% of the top quarks are produced through the quark interaction.

The top quark is detected indirectly via its decay products. It decays via the weak interaction, and according to the Standard Model (SM) is almost always expected to decay to a b quark and a W boson. This is followed by W decay into two quarks or a lepton and a neutrino. The final state of the $t\bar{t}$ system has different topological classifications that depend on the decay of the W boson. The results presented here use the lepton+jets channel, which corresponds to one W decaying leptonically (into a electron or a muon), and the other W hadronically. This channel has a branching fraction of about 30%.

Although its value is not predicted, M_{top} is a fundamental parameter in the Standard Model. The best value of the top quark mass found from combining all channels at the Tevatron is [1]

$$M_{top} = 174.4 \pm 5.1 \text{ GeV/c}^2 \tag{1}$$

The top-quark mass, along with the mass of the W boson, provides through radiative corrections the best indication for the value of the mass of the Higgs boson [2]. The measurement of M_W will improve significantly in the future, with an uncertainty of about 25 GeV/c^2 being a goal for Run II of the Tevatron. To be able to make maximum use of this precision measurement to constrain the mass of the Higgs, the top mass should be measured with an uncertainty of less than 3 GeV/c^2. This will yield a prediction for the Higgs mass with an uncertainty of 40%. It is therefore important to develop techniques for extracting the mass of the top quark that will optimize the use of Run II data.

The observation of the top quark at the Tevatron [3] has provided a new laboratory for examining the more subtle implications of the SM. The fact that the top quark is so massive has led to speculations that its interactions might be especially sensitive to the impact of symmetry breaking and any new physics that is expected to appear at the TeV energy scale. And, in fact, several pioneering studies of the decays of the top quark have already appeared int the literature [4, 5]. Although these have been severely limited by the low statistics of the data sample of Run I, they have nevertheless indicated that it is feasible to extract such information from the complex $t\bar{t}$ final states.

The standard top quark decays via V–A charged-current weak interaction. Hence, for massless b quarks [6], a top quark can decay to a left-handed W (negative helicity W_-) or a longitudinal W (zero helicity W_0). In the SM, top quarks decay to longitudinal W bosons with a branching ratio [7]:

$$F_0 = \frac{M_{top}^2}{M_{top}^2 + 2M_W^2} \approx 0.7 \tag{2}$$

where we take M_{top} = 174.3 GeV/c^2 and M_W = 80.4 GeV/c^2 [1].

We report a preliminary new measurement of the mass of the top quark and the longitudinal component of the helicity of the W boson in DØ data from Run I. The analysis is based on a new method of extracting parameters from hadron-collider data [8].

2. THE GENERAL METHOD

This method is similar to that suggested for $t\bar{t}$ dilepton decay channels, and used in previous mass analyses of dilepton events [9]. A similar approach was also suggested for the measurement of the mass of the W boson at LEP [10]. It compares each individual event with the differential cross section for $t\bar{t}$ production and decay. The luminosity used in this analysis corresponds to 125 events/pb, and the data was accumulated by the DØ experiment during Run I of the Tevatron. This analysis is based on the same sample that was used to extract the mass of the top quark in a previous publication [11]. A set of selections was introduced to improve acceptance for lepton+jets from $t\bar{t}$ relative to background. The standard requirements were: $E_T^{lepton} > 20$ GeV, $|\eta_e| < 2$, $|\eta_\mu| < 1.7$, $E_T^{jets} > 15$ GeV, $|\eta_{jets}| < 2$, $\not{E}_T > 20$ GeV, $|E_T^{lepton}| + |\not{E}_T| > 60$ GeV; $|\eta_{(lepton+\not{E}_T)}| < 2$. A total of 91 events remained after these selections. In our analysis we will use events that contain exactly four jets.

Given N events, a parameter α is estimated by maximizing the likelihood,

$$L(\alpha) = e^{-N\int P(x,\alpha)\mathrm{d}x} \prod_{i=1}^{N} P_m(x_i,\alpha) \tag{3}$$

where x is the set of variables needed to specify the measured event, $P_m(x,\alpha)$ is the probability of measuring that event, and α represents the parameters to be determined. The probability density can be written as a convolution of the calculable differential cross section and measurement resolution:

$$P(x,\alpha) = \frac{1}{\sigma}\int \mathrm{d}^n\sigma(y,\alpha)\mathrm{d}q_1\mathrm{d}q_2 f(q_1)f(q_2)W(x,y) \tag{4}$$

$W(y,x)$, our general transfer function, is the normalized probability that the measured set of variables x come from a set of partonic variables y, $d^n\sigma(y,\alpha)$ is the partonic differential cross section, and $f(q)$ are the parton distribution functions. Dividing by the total cross section σ for the process ensures $P_m(x,\alpha)$ is properly normalized. The integral in Eq. 4 sums over all possible parton states leading to what is observed in the detector.

The $t\bar{t}$ production probability is calculated as:

$$\begin{aligned} P_{t\bar{t}}(x;\alpha) =& \frac{1}{12\sigma_{t\bar{t}}}\int \mathrm{d}\rho_1 \mathrm{d}m_1^2 \mathrm{d}M_1^2 \mathrm{d}m_2^2 \mathrm{d}M_2^2 \\ & \times \sum_{\text{perm.},\nu} |\mathcal{M}_{t\bar{t}}|^2 \frac{f(q_1)f(q_2)}{|q_1||q_2|}\Phi_6 W_{jet}(E_y,E_x) \end{aligned} \tag{5}$$

where $|\mathcal{M}_{t\bar{t}}|^2$ is the leading order matrix element [12], $f(q_1)$ and $f(q_2)$ are the CTEQ4M parton distribution functions for the incident quarks [13], Φ_6 is the phase-space factor for the 6-object final state, and the sum is over all 12 permutations of the jets (the permutation of the jets from W decay was performed by symmetrizing the matrix element), and all possible longitudinal momenta of the neutrino solutions. The integration variables used in the calculation are the top masses ($m_{1,2}$), the W masses ($M_{1,2}$), and the energy of one of the quarks in the hadronic decay of the W bosons (ρ_1). Observed electron momenta are assumed to correspond to those of produced electrons. The angles of the jets are also assumed to reflect the angles of the partons in the final state, and we ignore any transverse momentum for the incident partons. $W_{jet}(E_y,E_x)$ corresponds to a function that parameterizes the mapping between parton-level energies E_y and energies measured in the detector E_x. A large Monte Carlo sample of $t\bar{t}$ events (generated with top masses between 140−200 GeV/c^2 in HERWIG [14], and processed through the DØ detector-simulation package) is used to determine $W_{jet}(E_y,E_x)$. For a final state with a muon, W_{jet} is expanded to include the known muon momentum resolution, and an integration over muon momentum is added to Eq. 5. Effects such as geometrical acceptance, trigger efficiencies, event selection, etc, are taken into account through a multiplicative function $A(x)$ that is independent of α. This function relates the production probability $P(x;\alpha)$ to the measured probability $P_m(x;\alpha)$: $P_m(x;\alpha) = A(x)P(x;\alpha)$. All processes that can contribute to the observed final state must be included in the probability. Therefore the final probability is written as $c_1P(x;\alpha) + c_2P_{\mathrm{bkg}}(x)$. The VECBOS [15] W+jets matrix element is used to calculate the background probability, which is integrated over the four quark energies and the W-boson mass, and later summed over the 24 jet permutations and neutrino solutions.

Since the method involves a comparison of data with a leading-order matrix element for the production and decay process, as mentioned above the sample is restricted to only four-jet events, thereby reducing the sample to 71 events. In order to increase the purity of signal, a selection is made on the probability that any event corresponds to background. This selection is required to minimize a bias introduced by the presence of background, and its imposition leaves a sample of only 22 events. Figure 1a) shows a comparison between the probability for a background interpretation of events calculated for a sample of Monte Carlo events (solid histogram) and for the 71 $t\bar{t}$ candidates (data points). The left-hatched (right-hatched) histogram shows the contribution from $t\bar{t}$ (W+4 jets) MC events. The ratio of $t\bar{t}$ to $W+4$ jets events in the MC is normalized to the ratio $S/B = 12/10$ observed in the data to the left of the vertical line. The selected value of the cuttoff $P_{\mathrm{bgd}} < 10^{-11}$ was based on MC studies carried out before applying the method to data, and, for a top mass of 175 GeV/c^2, it retains 70% of the signal while rejecting 70% of the background.

A discriminant $D = P_{t\bar{t}}/(P_{t\bar{t}} + P_{\mathrm{bkg}})$ was defined to compare the probability that an event corresponds to signal or background [11]. Since the signal probability depends on M_{top}, D was calculated with the signal probability taken at its most likely value. Figure 1b) shows a comparison of the discriminant calculated for data (points with error bars) and for MC (solid his-

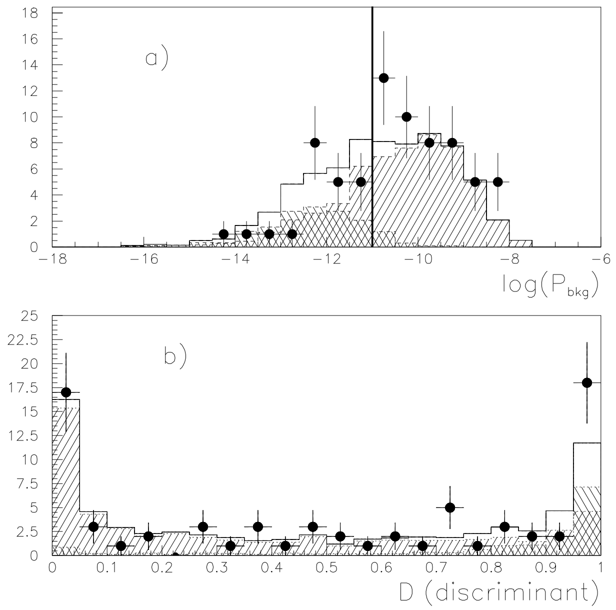


Figure 1: a) Distribution for probability of events being background, and b) discriminant $P_{t\bar{t}}/(P_{t\bar{t}}+P_{\rm bkg})$, calculated for the 71 $t\bar{t}$ candidates (data points). The data is compared with the results expected from MC-simulated samples (solid histogram). Only events with $P_{\rm bkg} < 10^{-11}$ are considered in the final analysis.

togram) events, with the MC normalized as in Figure 1a). The discriminant was not used to reject background, because (unlike the background probability) its value depends directly on M_{top}, and is shown simply to illustrate the level of discrimination of signal from background.

The probabilities are inserted into the likelihood function of Eq. 3, and the best estimate of M_{top} and F_0 is obtained by maximizing this likelihood function. (-ln$L$ was minimized with respect to the parameters c_1, c_2, and M_{top} or F_0.)

Figure 2a) shows the value of -ln$L$ as a function of M_{top} for the 22 events that pass all the selection criteria, 12 of which are signal and 10 background. Figure 2b) shows the likelihood normalized to its maximum value. The Gaussian fit in the figure yields M_{top}=179.6 GeV/c^2, and an uncertainty δM_{top}=3.6 GeV/c^2. Monte Carlo studies show that there is a shift to 0.5 GeV/c^2 in the extracted mass. Applying this shift, yields the new preliminary result:

$$M_{top}(preliminary) = 180.1 \pm 3.6(\mathrm{stat}) \pm 4.0(\mathrm{sys})\ \mathrm{GeV/c^2} \tag{6}$$

The main systematic uncertainties are due to the jet-energy scale (3.6 GeV/c^2), model for $t\bar{t}$ (1.5 GeV/c^2), model for background (1.0 GeV/c^2), noise and multiple interactions (1.3 GeV/c^2), parton distribution functions (0.2 GeV/c^2), and acceptance corrections (0.5 GeV/c^2).

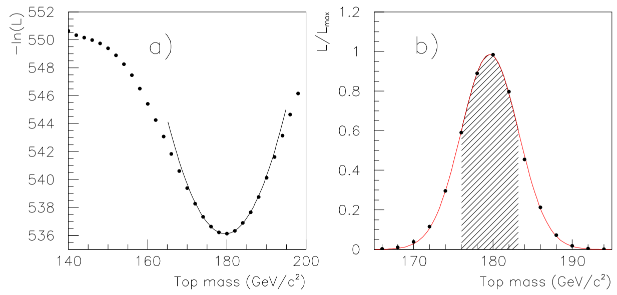


Figure 2: a) Negative of the log of the likelihood as a function of the mass of the top quark for the 22 $t\bar{t}$ candidates in our final sample. b) Likelihood normalized to the maximum value. The curves are Gaussian fits to the likelihood plot b). The hatched area corresponds to the 68.27% probability interval.

Figure 3a) shows the log of the likelihood as a function of F_0 for the 22 data events and Fig. 3b) shows the likelihood normalized to its maximum value. The shaded region corresponds to the most narrow 68.27% probability interval about the most probable value, and reflects the statistical error convoluted with the uncertainty on the top mass. The uncertainty on the top mass was included through an integration of the probability over the top mass, assuming a uniform prior. This likelihood also has a response correction to F_0, which was obtained from Monte Carlo studies. This probability is fitted to a 5^{th} order polynomial as a function of F_0. We use the most probable output value and the smallest 68.27% interval within the physical region to define our extracted value of F_0:

$$F_0(preliminary) = 0.56 \pm 0.31(\mathrm{stat}) \tag{7}$$

The other systematic uncertainties were calculated by varying their impact in the Monte Carlo or data, and added in quadrature. The systematic uncertainties are due to the jet-energy scale (0.014), model for $t\bar{t}$ (0.020), background (0.010), multiple interactions (0.009), parton distribution functions (0.007), $t\bar{t}$ spin correlations (0.008), and acceptance corrections (0.021). The final preliminary result is

$$F_0(preliminary) = 0.56 \pm 0.31(\mathrm{stat}) \pm 0.04(\mathrm{syst}) \tag{8}$$

consistent with expectations of the SM.

Acknowledgments

The author wishes to thank Juan Estrada, Tom Ferbel, and Gaston Gutierrez for their collaboration in this analysis.

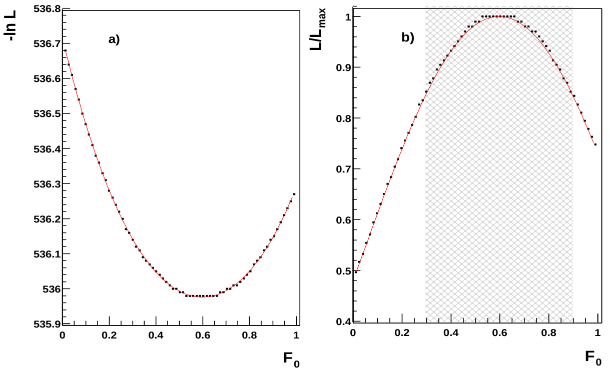


Figure 3: a) -ln L as a function of F_0 from the Run I data sample. b) Likelihood normalized to the maximum value. The curves are polynomial of 5th order fits to the likelihood plot b). The hatched area corresponds to the 68.27% probability interval.

References

[1] K. Hagiwara *et al*, Particle Data Group, Phys. Rev. D **66**, 010001 (2002).

[2] LEP Electroweak Working Group, *http://lepewwg.web.cern.ch*.

[3] S.Abachi *et al*, DØ Collaboration. Phys. Rev. Lett. **74**: 2632 (1995); F.Abe *et al*, CDF Collaboration. Phys. Rev. Lett. **74**: 2626 (1995).

[4] T. Affolder *et al*, CDF Collaboration. Phys. Rev. Lett. **84**: 216 (2000).

[5] B. Abbott *et al*, DØ Collaboration. Phys. Rev. Lett. **85**: 256 (2000).

[6] M. Fischer, S. Groote, J. G. Korner, M. C. Mauser. Phys. Rev. D **65**: 054036, (2002).

[7] I. Bigi *et al*, Phys. Lett. B **181**: 157 (1986); G. L. Kane, G. A. Ladinsky. Phys. Rev. D **45**, 124 (1992); T. Tait, C.-P. Yuan. Phys. Rev. D **63**: 014018, (2001).

[8] M. F. Canelli. Ph.D. Thesis, University of Rochester, (2003); J. C. Estrada. Ph.D. Thesis, University of Rochester, (2001).

[9] B. Abbott *et al*, DØ Collaboration. Phys. Rev D **60**, 052001 (1999); R. H. Dalitz and G. R. Goldstein, Proc. R. Soc. Lond. A **445**, 2803 (1999), and referenced therein; K. Kondo *et al*, J. Phys. Soc. Jap. **62**, 1177 (1993).

[10] F. A. Berends, C. G. Papadopoulos, and R. Pittau, Phys. Rev. Lett. B **411**, 133 (1997).

[11] B. Abbott *et al*, DØ Collaboration. Phys. Rev. D **58**, 052001 (1998).

[12] G. Mahlon, S. Parke, Phys. Lett. B **411**, 133 (1997); G. Mahlon, S. Parke, Phys. Rev. D **53**, 4886 (1996).

[13] H. L. Lai *et al*, Phys. Rev. D **51**, 4763 (1995).

[14] G. Marchesini *et al*, Comput. Phys. Commun. **67**, 467 (1992).

[15] F. A. Berends, H. Kuijf, B. Tausk and W. T. Giele. Nucl. Phys. **B357**, 32 (1991).

Temporal Bias in the Clustering of Massive Cosmological Objects

Evan Scannapieco
Kavli Institute for Theoretical Phys., Kohn Hall, UC Santa Barbara, Santa Barbara, CA 93106
Robert J. Thacker
Dept. of Phys. & Astron., McMaster Univ., 1280 Main St. West, Hamilton, Ontario, L8S 4M1, Canada

It is a well-established fact that massive cosmological objects exhibit a "geometrical bias" that boosts their spatial correlations with respect to the underlying mass distribution. Although this geometrical bias is a simple function of mass, this is only half of the story. We show using numerical simulations that objects that are in the midst of accreting material also exhibit a "temporal bias," which further boosts their clustering far above geometrical bias levels. These results may help to resolve a discrepancy between spectroscopic and clustering mass estimates of Lyman Break Galaxies, a population of high-redshift galaxies that are caught in the act of forming large numbers of new stars.

1. INTRODUCTION

Large-scale structure in the Universe is believed to have originated from a primordial Gaussian random field of matter fluctuations, the product of quantum fluctuations that were shifted to larger spatial scales during cosmological inflation [*e.g.* 1]. Cosmic Microwave Background observations show that when the Universe was 100,000 years old, the gaseous component was extremely smooth, with temperature variations $\sim 10^{-5}$. These tiny inhomogeneities, in concert with inhomogeneities in the unseen massive dark-matter, were amplified through gravitational instability, eventually forming the galaxies, clusters, and other cosmological objects we see today.

While the precise details of structure formation are highly complex, the simplicity of the initial random field allows us to easily compute the overall distributions of cosmological objects. Galaxies and galaxy clusters represent the peaks in the initial density distribution which accrete matter at the expense of the diffuse regions between them, and thus their number densities can be simply related to the number densities of the peaks of the initial random field. This technique has been applied most cleanly to galaxy clusters, whose densities and evolution provide strong constraints on the overall matter density [2].

Similarly, because the peaks in a random field are more clustered than the overall distribution, the spatial clustering of cosmological objects is stronger than the underlying mass distribution. Furthermore, this "geometrical bias" is a systematic function of the mass of these structures, an effect that has been well-studied analytically and numerically [*e.g.* 3, 4, 5].

Yet, this is only half of the story. Here we conduct a detailed numerical simulation that shows that the spatial correlation function of objects that are in the midst of accreting substantial amounts of material is significantly enhanced over that of the general population. This temporal bias causes them to mimic the properties of higher-mass structures, with important astrophysical implications as discussed below.

The structure of this work is as follows: In §2 we describe our numerical simulation, discuss our group-finding algorithms, and develop a robust definition of accreting groups. In §3 we present our results for the spatial correlation functions of these samples, and in §4 we discuss the astrophysical implications of our results. Further details of this study are given in [6].

2. SIMULATIONS AND GROUP FINDING

Our numerical simulation traced the growth of primordial density fluctuations by dynamically evolving a large number of point test particles. The distribution of these particles was then used to determine the nonlinear evolution of the spatial correlation function of massive objects at late times. Driven by measurements of the Cosmic Microwave Background, the number abundance of galaxy clusters, and high redshift supernova estimates [*e.g.* 7, 2, 8] we focused our attention on a Cold Dark Matter cosmological model with parameters $H = 70$ km/s/Mpc, $\Omega_0 = 0.3$, $\Omega_\Lambda = 0.65$, $\Omega_b = 0.05$, and $\sigma_8 = 0.87$, where H is the Hubble constant today, Ω_0, Ω_Λ, and Ω_b are the total matter, vacuum, and baryonic densities in units of the critical density, and σ_8^2 is the present variance of linear fluctuations on the $8 \times (100/70)$ Mpc scale (where 1 Mpc is 3.26×10^6 light years).

Periodic boundary conditions, which approximate large-scale homogeneity and isotropy, were taken, and the mass within our simulation volume was held fixed. Thus the overall box expanded along with the cosmological expansion, such that each side at any given redshift z was $73/(1+z)$ Mpc across. This box was populated with 350^3 dark matter particles that interact only gravitationally and represent the dominant mass component of the Universe. The mass of each particle, 4.3×10^8 solar masses ($M_\odot$), was chosen to match the observed mass density of the Universe, and the simulation was started at an initial redshift of $z = 49$. The simulation used a parallel OpenMP-based version of the HYDRA code [9, 10] with 64-bit precision.

To demonstrate the robustness of our results we have chosen two distinct group-finding approaches; the friends-of-friends approach [11] (FOF) and the HOP algorithm [12]. FOF works by linking together all pairs of particles within a fixed "linking-length" of each other, and then taking each such group of "friends" to be an identified cosmological object. Although it remains popular, the FOF mass estimates are known to have significant scatter due to a problem that can occur as small strings of particles fall within the linking length.

The HOP algorithm works by using the local density for each particle to trace ('hop') along a path of increasing density to the nearest density maxima, at which point the particle is assigned to the group defined by that local density maximum. As this process assigns all particles to groups, a 'regrouping' stage is needed in which a merger criterion for groups above a threshold density δ_{outer} is applied. This criterion merges all groups for which the boundary density between them exceeds δ_{saddle}, and all groups thus identified must have one particle that exceeds δ_{peak} to be accepted as a group (see [12] for explicit details).

Beginning from $z = 4.89$, we saved particle positions every 50 million years up to the final output at $z = 3$. For the final 5 outputs we found FOF groups using a linking parameter of $b = 0.18$, and HOP groups using the parameters: $N_{dens} = 48$, $N_{hop} = 20$, $N_{merge} = 5$, $\delta_{peak} = 160$, $\delta_{saddle} = 140$, and $\delta_{outer} = 80$. Visual inspection showed strong similarities between the two populations, with a small amount of unavoidable noise coming from groups around the 80 particle resolution limit (a group found by FOF at this limit may not be found by HOP and vice versa). We compare groups from one output to another by tracing back all particles with a given group index from the later output to the earlier output. Particles that show no membership to a group at the earlier time are regarded as 'smooth infall' while other non-null indices describe the merger history of the object.

To give a rough estimate of the accuracy of the group finding methods in Fig. 1 we plot the mass of the most massive progenitor at $t_1(z = 3.059)$, versus the mass at $t_2(z = 3)$, such that $\Delta t = 5 \times 10^7$ years. As compared to the FOF groups, the smaller fraction of HOP groups lying above the equal mass line shows that the HOP algorithm identifies groups that are more likely to be massive at later outputs. The effect of this difference is significant.

Our definition of accreting groups is similar to that of [13], except that we select the subset that grew by 20% from t_1 to t_2, which implicitly includes mass accretion via smooth infall and results in 545(980) HOP(FOF) groups if $\Delta t = 5 \times 10^7$ years. Note that the mass of each group is that at the end of each time interval, such that we tag all groups that *experienced* appreciable infall. The 20% value is arbitrary, but we selected it primarily because it appears to lie outside the central 'noise' band in the FOF data (see Fig 1). The groups corresponding to this cut appear in Fig 1, as points to the right of the dashed lines.

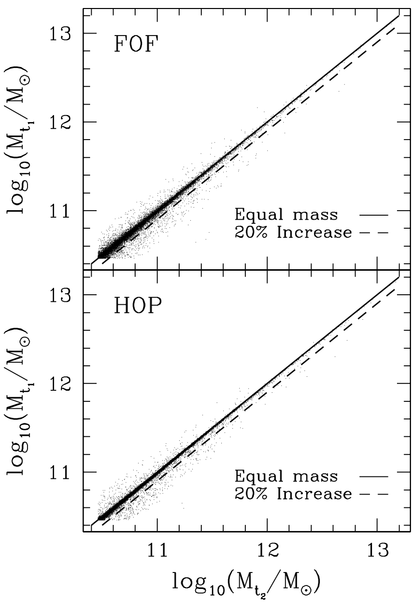


Figure 1: Comparison of group growth. The FOF algorithm exhibits a significant amount of scatter in mass estimates between outputs. Only 67% of groups grow from time t_1 to t_2, compared with 82% for HOP.

3. TEMPORAL BIAS

In Figure 2 we show the spatial correlation functions of the groups selected by both the HOP and FOF algorithms and compare them with $\xi(r)$ of the accreting groups. This function measures the excess probability of finding a pair of groups at a given separation r relative to a random distribution, and is calculated for a separation bin r_l as

$$1 + \xi(r_l) = N(r_l)/N_{\rm random}(r_l), \qquad (1)$$

where $N(r_l)$ is the number of pairs separated by distances between r_l and r_{l+1}, and $N_{\rm random}(r_l) = \frac{1}{2}N^2\frac{4\pi}{3}(r_{l+1}^3 - r_l^3)/V$, with N the total number of groups and V the volume of the simulation. In the accreting case we co-added the correlation functions calculated from the differences from the last four $\Delta t = 5 \times 10^7$ year intervals and the last two $\Delta t = 10 \times 10^7$ year intervals. Radial bins of 1/80 the simulation size, corresponding to 0.92 comoving Mpc, were taken throughout. For comparison, in each panel of Fig. 2 we also show the correlation function of all the groups in the next largest mass bin. The amplitudes of the correlation functions obtained using the full set of HOP and FOF groups agree with each other

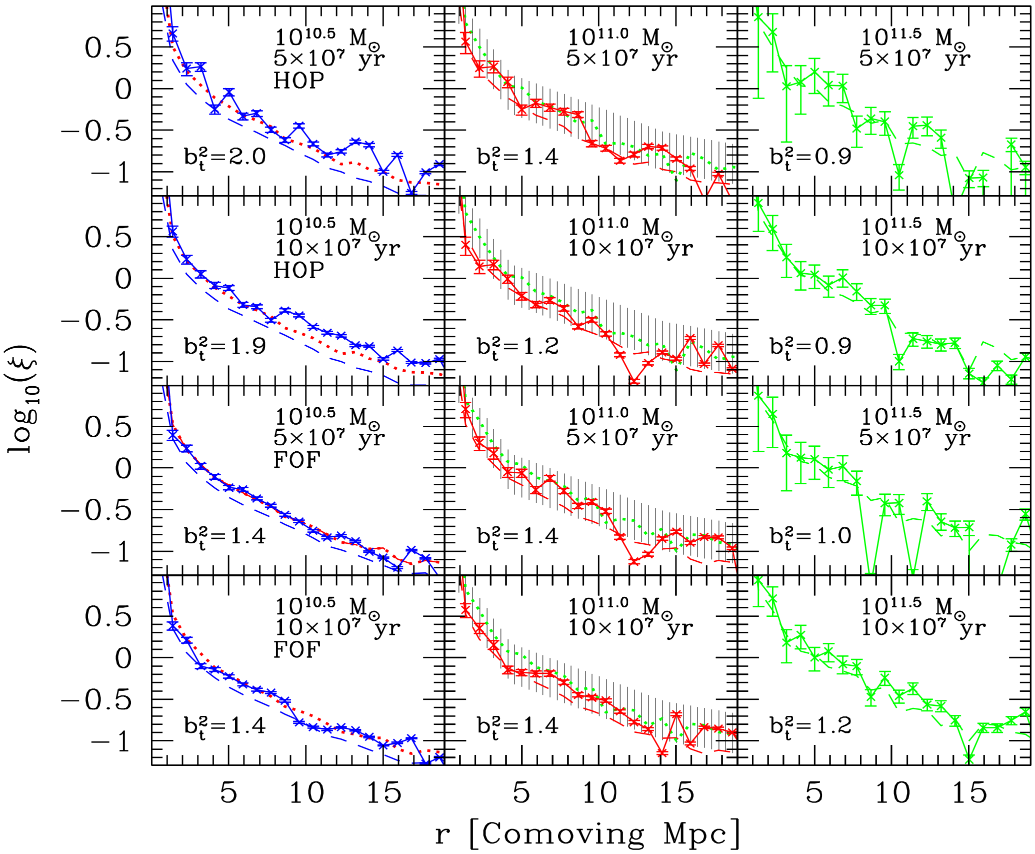


Figure 2: Spatial Correlation Functions. In each panel the dashed line shows the correlation function for all the groups, while the points connected by the solid line show $\xi(r)$ for groups that have accreted appreciable mass in the last Δt years. Panels are labeled by their mass range and Δt values, and in each panel, the dotted line shows the correlation function of all the groups in the next highest mass bin. The top two rows were generated from a set of groups selected by the HOP algorithm, while the groups in the lower two rows were selected using the FOF approach. The shaded region in the central panels represents the observed correlation function of $\mathcal{R}_{AB} \leq 25.5$ Lyman break galaxies as computed in [16] by inversion of the angular correlation function. A 10% accretion threshold is applied in the $10^{11.5} M_\odot$ case to increase the number of measured groups.

to within statistical uncertainties, as well as with analytical estimates.

The upper row demonstrates a clear enhancement of the clustering of accreting groups at both the $10^{10.5} M_\odot$ and $10^{11.0} M_\odot$ mass scales, with their correlation functions roughly matching those of objects three times greater in mass (no conclusion can be drawn from the high mass bin as the sample is too small). This "temporal biasing" arises from the fact that *both* objects accreting substructure as well as those experiencing considerable smooth infall tend to be found in the densest regions of space, which are themselves highly clustered. This conclusion is supported by the fact that the average local overdensity of groups in the $10^{10.5}(10^{11}) M_\odot$ mass bin is 0.82(0.87) (measured in 4 Mpc (comoving) spheres, corresponding to a mass scale of 1.2×10^{13} $M_\odot$), whereas the same mass bin for the entire population exhibits an overdensity of 0.60(0.73).

In the second row of Fig. 2 we take a longer interval of $\Delta t = 10 \times 10^7$ yr. This has a only slight dampening effect on temporal bias, which can not be definitively distinguished from statistical noise in our measurements. In the $\Delta t = 20 \times 10^7$ yr case, however, only a very weak enhancement of $\xi(r)$ was measured.

In the lower two rows of this figure, we repeat our analyses using the FOF group finder. Although this approach is more susceptible to statistical noise, the same trends are apparent as in the HOP case. If $\Delta t = 5 \times 10^7$ yr, this temporal bias is roughly equal to the geometrical bias of the groups three times more massive, while if $\Delta t = 10 \times 10^7$ yr, $\xi(r)$ is boosted to a slightly lesser degree.

Finally, to quantify our results, we have computed the effective temporal bias in each mass bin, Δt, and group finder. We define b_t^2 as the ratio of the correlation function of the accreting groups to the overall correlation function, weighted by the number of points in each bin in the overall function; $b_t^2 \equiv \sum_{i=0}^{20} \frac{\xi_{\mathrm{accreting},i} N_{\mathrm{all},i}}{\xi_{\mathrm{all},i} N_{\mathrm{all},i}}$, where the sum is carried out over all bins within $r \leq 20$ comoving Mpc. These values are labeled in each panel, and in the $\Delta t = 20 \times 10^7$ yr case, $b_t^2 = 1.1(1.0)$ in the $10^{10.5}(10^{11.0}) M_\odot$ HOP bins and 1.1(1.3) in the respective FOF bins.

4. ASTROPHYSICAL IMPLICATIONS

While temporal bias is a general property of the peaks of a gravitationally amplified Gaussian random field, our results have specific implications for the large sample of $z \sim 3$ galaxies made available by the Lyman-break color-selection technique [14]. Lyman break galaxies (LBGs) are observed to have enormous star-formation rates on the order of $\sim 50 M_\odot$ per year [15], implying that these objects are likely to be accreting large amounts of material. Furthermore, although the clustering of LBGs brighter than $\mathcal{R}_{AB} \leq 25.5$ is roughly that expected from the geometrical bias of $10^{12} M_\odot$ objects [*e.g.* 16, 17], the linewidths measured from a spectroscopic sample these galaxies correspond to total masses $\leq 10^{11} M_\odot$ [18].

To relate our result to LBGs we plot the spatial correlation function of $\mathcal{R}_{AB} \leq 25.5$ LBGs, as derived in [16], in the center column of Fig. 2. Although there are significant uncertainties involved in computing this quantity, since comparisons are more naturally conducted in angular coordinates, the shaded regions provide a guide to the range of $\xi(r)$ values consistent with observations. In these panels, we see that if $\Delta t = 5 \times 10^7$ yr is chosen, then temporal bias boosts the correlation function of $10^{11} M_\odot$ groups into reasonable agreement with observations.

This mass is marginally consistent with the upper mass bound inferred from the rotation curves of a somewhat bright ($\mathcal{R}_{AB} \lesssim 24$) spectroscopic subset of LBGs [18]. Furthermore, only $\sim 4\%$ of all groups exhibit appreciable accretion in each $\Delta t = 5 \times 10^7$ year time interval and the density of $10^{11} M_\odot$ groups is $\sim 2 \times 10^{-2}$ Mpc^3, at $z = 3$ in our assumed cosmology. Thus associating such objects with 5×10^7 year starbursts results in a density $\sim 5 \times 10^{-4}$ Mpc^3, comparable with that observed.

While quite suggestive, these comparisons are not meant as a complete model, and may not prove to be the final explanation of the discrepant mass estimates of LBGs. Kinematic models have been explored, for example, in which the observed velocity dispersions of LBGs are much less than the circular velocities of the groups in which they are contained [19]. What is clear however, is that this bias can not be ignored and must be carefully considered when interpreting the clustering of these objects. While perhaps only part of the story, temporal biasing represents an important factor that must be taken into account when studying the properties of Lyman break galaxies.

Acknowledgments

ES would like to express his sincere thanks for the hospitality shown to him by Jon Weisheit and the T-6 group at Los Alamos National Laboratory, where this work was initiated. We are grateful to Marc Davis for fruitful suggestions. ES was supported in part by an NSF MPS-DRF fellowship. RJT acknowledges funding from the Canadian Computational Cosmology Consortium and use of the CITA computing facilities. This work was supported by the National Science Foundation under grant PHY99-07949.

References

[1] Linde, A. D. 1983, Phys. Lett., 129B, 177

[2] Eke, V. R., Cole, S., & Frenk C. S. 1996, Monthly Notices of the Royal Astronomical Society, 282, 263

[3] Kaiser, N. 1984, Astrophysical Journal, 284, L9

[4] Mo, H. J. & White S. D. M., 1996, Monthly Notices of the Royal Astronomical Society, 282, 348

[5] Jing, Y. P. 1999, Astrophysical Journal, 515, L45

[6] Scannapieco, E. & Thacker, R. J. 2003, Astrophysical Journal Letters 590, 69

[7] Spergel, D. N. et al. 2003, Astrophysical Journal Supplement, 148, 175

[8] Perlmutter, S. et al. 1999, Astrophysical Journal, 517, 565

[9] Couchman, H. M. P., Thomas, P. A., & Pearce, F. R. 1995, Astrophysical Journal, 452, 797

[10] Thacker, R. J & Couchman, H .M. P. 2000, Astrophysical Journal, 545, 728

[11] Davis, M., Efstathiou, G., Frenk, C. S., & White, S. D. M. 1985, Astrophysical Journal, 292, 371

[12] Eisenstein, D. J. & Hut, P. 1998, Astrophysical Journal, 498, 137

[13] Percival, W. J., Scott, D., Peacock, J., A., & Dunlop, J. S. 2003, Monthly Notices of the Royal Astronomical Society, 338, L31

[14] Steidel, C. C., Adelberger, K. L., Dickinson, M., Giavalisco, M., Pettini, M., Kellogg, M. 1998, Astrophysical Journal, 492, 428

[15] Adelberger, K. L. & Steidel, C. C. 2000, Astrophysical Journal, 544, 218

[16] Wechsler, R. H., Somerville, R. S., Bullock, J. S.; Kolatt, T. S.; Primack, J. R.; Blumenthal, G. R.; Dekel, A. 2001, Astrophysical Journal, 554, 85

[17] Porciani, C. & Giavalisco, M. 2002, Astrophysical Journal, 565, 24

[18] Pettini, M et al. 2001, Astrophysical Journal, 554, 981

[19] Mo, H. J., Mao, S., & White, S. D. M. 1999, Monthly Notices of the Royal Astronomical Society, 304, 175

Asymmetric Errors

Roger Barlow
Manchester University, UK and Stanford University, USA

Errors quoted on results are often given in asymmetric form. An account is given of the two ways these can arise in an analysis, and the combination of asymmetric errors is discussed. It is shown that the usual method has no basis and is indeed wrong. For asymmetric systematic errors, a consistent method is given, with detailed examples. For asymmetric statistical errors a general approach is outlined.

1. ASYMMETRIC ERRORS

In the reporting of results from particle physics experiments it is common to see values given with errors with different positive and negative numbers, to denote a 68% central confidence region which is not symmetric about the central estimate. For example (one of many) the Particle Data Group[1] quote

$$B.R.(f_2(1270) \to \pi\pi) = (84.7^{+2.4}_{-1.3})\%.$$

The purpose of this note is to describe how such errors arise and how they can properly be handled, particularly when two contributions are combined. Current practice is to combine such errors separately, i.e. to add the σ^+ values together in quadrature, and then do the same thing for the σ^- values. This is not, to my knowledge, documented anywhere and, as will be shown, is certainly wrong.

There are two separate sources of asymmetry, which unfortunately require different treatments. We call these 'statistical' and 'systematic'; the label is fairly accurate though not entirely so, and they could equally well be called 'frequentist' and 'Bayesian'.

Asymmetric statistical errors arise when the log likelihood curve is not well described by a parabola [2]. The one sigma values (or, equivalently, the 68% central confidence level interval limits) are read off the points at which $\ln L$ falls from its peak by $\frac{1}{2}$ – or, equivalently, when χ^2 rises by 1. This is not strictly accurate, and corrections should be made using Bartlett functions[3], but that lies beyond the scope of this note.

Asymmetric systematic errors arise when the dependence of a result on a 'nuisance parameter' is non-linear. Because the dependence on such parameters – theoretical values, experimental calibration constants, and so forth – is generally complicated, involving Monte Carlo simulation, this study generally has to be performed by evaluating the result x at the $-\sigma$ and $+\sigma$ values of the nuisance parameter a (see [4] for a fuller account) giving σ_x^- and σ_x^+. ($a \pm \sigma$ gives $\sigma_x^\pm$ or $\sigma_x^\mp$ according to the sign of $\frac{dx}{da}$.)

This note summarises a full account of the procedure for asymmetric systematic errors which can be found in [5] and describes what has subsequently been achieved for asymmetric statistical errors. For another critical account see [6].

2. ASYMMETRIC SYSTEMATIC ERRORS

If σ_x^- and σ_x^+ are different then this is a sign that the dependence of x on a is non-linear and the symmetric distribution in a gives an asymmetric distribution in x. In practice, if the difference is not large, one might be well advised to assume a straight line dependence and take the error as symmetric, however we will assume that this is not a case where this is appropriate. We consider cases where a non-linear effect is not small enough to be ignored entirely, but not large enough to justify a long and intensive investigation. Such cases are common enough in practice.

2.1. Models

For simplicity we transform a to the variable u described by a unit Gaussian, and work with $X(u) = x(u) - x(0)$. It is useful to define the mean σ, the difference α, and the asymmetry A:

$$\sigma = \frac{\sigma^+ + \sigma^-}{2} \qquad \alpha = \frac{\sigma^+ - \sigma^-}{2} \qquad A = \frac{\sigma^+ - \sigma^-}{\sigma^+ + \sigma^-} \tag{1}$$

There are infinitely many non-linear relationships between u and X that will go through the three determined points. We consider two. We make no claim that either of these is 'correct'. But working with asymmetric errors must involve some model of the non-linearity. Practitioners must select one of these two models, or some other (to which the same formalism can be applied), on the basis of their knowledge of the problem, their preference and experience.

- Model 1: Two straight lines

  Two straight lines are drawn, meeting at the central value

$$\begin{aligned} X &= \sigma^+ u \qquad u \geq 0 \\ &= \sigma^- u \qquad u \leq 0. \end{aligned} \tag{2}$$

- Model 2: A quadratic function

  The parabola through the three points is

$$X = \sigma u + \alpha u^2 = \sigma u + A\sigma u^2. \tag{3}$$

These forms are shown in Figure 1 for a small asymmetry of 0.1, and a larger asymmetry of 0.4.

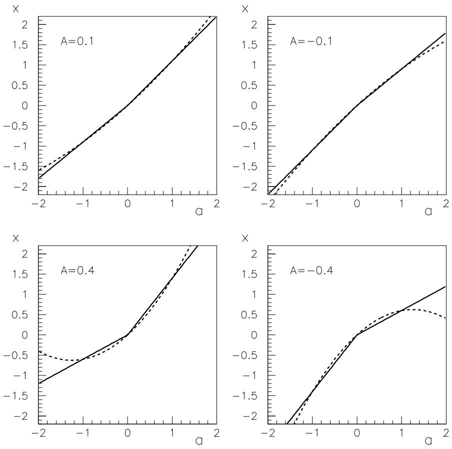


Figure 1: Some nonlinear dependencies

Model 1 is shown as a solid line, and Model 2 is dashed. Both go through the 3 specified points. The differences between them within the range $-1 \leq u \leq 1$ are not large; outside that range they diverge considerably.

The distribution in u is a unit Gaussian, $G(u)$, and the distribution in X is obtained from $P(X) = \frac{G(u)}{|dX/du|}$. Examples are shown in Figure 2. For Model 1 (again a solid line) this gives a dimidated Gaussian - two Gaussians with different standard deviation for $X > 0$ and $X < 0$. This is sometimes called a 'bifurcated Gaussian', but this is inaccurate. 'Bifurcated' means 'split' in the sense of forked. 'Dimidated' means 'cut in half', with the subsidiary meaning of 'having one part much smaller than the other' [7]. For Model 2 (dashed) with small asymmetries the curve is a distorted Gaussian, given by $\frac{G(u)}{|\sigma+2\alpha u|}$ with $u = \frac{\sqrt{\sigma^2+4\alpha X}-\sigma}{2\alpha}$. For larger asymmetries and/or larger $|X|$ values, the second root also has to be considered.

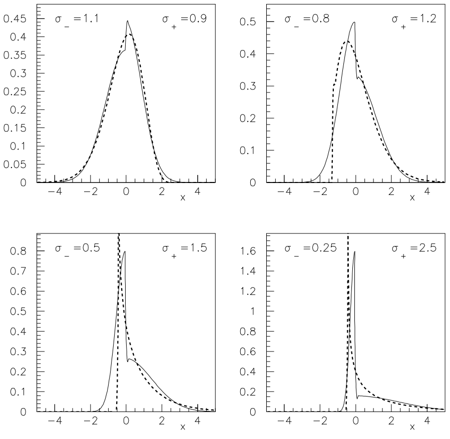


Figure 2: Probability Density Functions from Figure 1

It can be seen that the Model 1 dimidated Gaussian and Model 2 distorted Gaussian are not dissimilar if the asymmetry is small, but are very different if the asymmetry is large.

2.2. Bias

If a nuisance parameter u is distributed with a Gaussian probability distribution, and the quantity $X(u)$ is a nonlinear function of u, then the expectation $\langle X \rangle$ is not $X(\langle u \rangle)$.

For model 1 one has

$$< X >= \frac{\sigma^+ - \sigma^-}{\sqrt{2\pi}} \tag{4}$$

For model 2 one has

$$< X >= \frac{\sigma^+ - \sigma^-}{2} = \alpha \tag{5}$$

Hence in these models, (or any others), if the result quoted is $X(0)$, it is not the mean. It differs from it by an amount of the order of the difference in the positive and negative errors. It is perhaps defensible as a number to quote as the result as it is still the median - there is a 50% chance that the true value is below it and a 50% chance that it is above.

2.3. Adding Errors

If a derived quantity z contains parts from two quantities x and y, so that $z = x + y$, the distribution in z is given by the convolution:

$$f_z(z) = \int dx f_x(x) f_y(z-x) \tag{6}$$

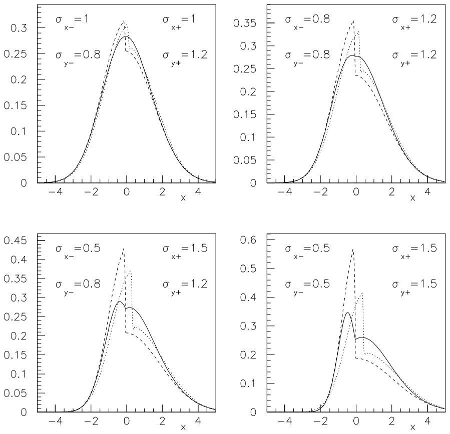


Figure 3: Examples of the distributions from combined asymmetric errors using Model 1.

With Model 1 the convolution can be done analytically. Some results for typical cases are shown in

Figure 3. The solid line shows the convolution, the dashed line is obtained by adding the positive and negative standard deviations separately in quadrature (the 'usual procedure'). The dotted line is described later.

The solid and dashed curves disagree markedly. The 'usual procedure' curve has a larger skew than the convolution. This is obvious. If two distributions with the same asymmetry are added the 'usual procedure' will give a distribution just scaled by $\sqrt{2}$, with the same asymmetry. This violates the Central Limit Theorem, which says that convoluting identical distributions must result in a combined distribution which is more Gaussian, and therefore more symmetric, than its components. This shows that the 'usual procedure' for adding asymmetric errors is inconsistent.

2.4. A consistent addition technique

If a distribution for x is described by some function, $f(x; x_0, \sigma^+, \sigma^-)$, which is a Gaussian transformed according to Model 1 or Model 2 or anything else, then 'combination of errors' involves a convolution of two such functions according to Equation 6. This combined function is not necessarily a function of the same form: it is a special property of the Gaussian that the convolution of two Gaussians gives a third. The (solid line) convolution of two dimidated Gaussians is not itself a dimidated Gaussian. Figure 3 is a demonstration of this.

Although the form of the function is changed by a convolution, some things are preserved. The semi-invariant cumulants of Thièle (the coefficients of the power series expansion of the log of the Fourier Transform) add under convolution. The first two of these are the usual mean and variance. The third is the unnormalised skew:

$$\gamma = < x^3 > -3 < x >< x^2 > +2 < x >^3 \qquad (7)$$

Within the context of any model, a consistent approach to the combination of errors is to find the mean, variance and skew: μ, V and γ, for each contributing function separately. Adding these up gives the mean, variance and skew of the combined function. Working within the model one then determines the values of σ_-, σ_+, and x_0 that give this mean, variance and skew.

2.5. Model 1

For Model 1, for which $\langle x^3 \rangle = \frac{2}{\sqrt{2\pi}}(\sigma_+^3 - \sigma_-^3)$ we have

$$\mu = x_0 + \frac{1}{\sqrt{2\pi}}(\sigma^+ - \sigma^-)$$

$$V = \sigma^2 + \alpha^2 \left(1 - \frac{2}{\pi}\right)$$

$$\gamma = \frac{1}{\sqrt{2\pi}}\Big[2({\sigma^+}^3 - {\sigma^-}^3) - \frac{3}{2}(\sigma^+ - \sigma^-)({\sigma^+}^2 + {\sigma^-}^2)$$

$$+\frac{1}{\pi}(\sigma^+ - \sigma^-)^3\Big] \qquad (8)$$

Given several error contributions the Equations 8 give the cumulants μ, V and γ of each. Adding these up gives the first three cumulants of the combined distribution. Then one can find the set of parameters σ^-, σ^+, x_0 which give these values by using Equations 8 in the other sense.

It is convenient to work with Δ, where Δ is the difference between the final x_0 and the sum of the individual ones. The parameter is needed because of the bias mentioned earlier. Even though each contribution may have $x_0 = 0$, i.e. it describes a spread about the quoted result, it has non-zero μ_i through the bias effect (c.f. Equations 4 and 5). The σ^+ and σ^- of the combined distribution, obtained from the total V and γ, will in general not give the right μ unless a location shift Δ is added. *The value of the quoted result will shift.*

Recalling section B, for the original distribution one could defend quoting the central value as it was the median, even though it was not the mean. The convoluted distribution not only has a non-zero mean, it also (as can be seen in Figure 3) has non-zero median. If you want to combine asymmetric errors then you have to accept that the quoted value will shift. To make this correction requires a real belief in the asymmetry of the error values. At this point practitioners, unless they are sure that their errors really do have a significant asymmetry, may be persuaded to revert to quoting symmetric errors.

Solving the Equations 8 for σ^-, σ^+ and x_0 given μ, V and γ has to be done numerically. A program for this is available on `http://www.slac.stanford.edu/~barlow`. Some results are shown in the dotted curve of Figure 3 and Table 1.

Table I Adding errors in Model 1

σ_x^-	σ_x^+	σ_y^-	σ_y^+	σ^-	σ^+	Δ
1.0	1.0	0.8	1.2	1.32	1.52	0.08
0.8	1.2	0.8	1.2	1.22	1.61	0.16
0.5	1.5	0.8	1.2	1.09	1.78	0.28
0.5	1.5	0.5	1.5	0.97	1.93	0.41

It is apparent that the dotted curve agrees much better with the solid one than the 'usual procedure' dashed curve does. It is not an exact match, but does an acceptable job given that there are only 3 adjustable parameters in the function. If the shape of the solid curve is to be represented by a dimidated Gaussian, then it is plausible that the dotted curve is the 'best' such representation.

2.6. Model 2

The equivalent of Equations 8 are

$$\begin{aligned}\mu &= x_0 + \alpha \\ V &= \sigma^2 + 2\alpha^2 \\ \gamma &= 6\sigma^2\alpha + 8\alpha^3\end{aligned} \qquad (9)$$

As with Method 1, these are used to find the cumulants of each contributing distribution, which are summed to give the three totals, and then Equation 9 is used again to find the parameters of the distorted Gaussian with this mean, variance and skew. The web program will also do these calculations

Some results are shown in Figure 4 and Table II. The true convolution cannot be done analytically but can be done by a Monte Carlo calculation.

Table II Adding errors in Model 2

σ_x^-	σ_x^+	σ_y^-	σ_y^+	σ^-	σ^+	Δ
1.0	1.0	0.8	1.2	1.33	1.54	0.10
0.8	1.2	0.8	1.2	1.25	1.64	0.20
0.5	1.5	0.8	1.2	1.12	1.88	0.35
0.5	1.5	0.5	1.5	1.13	2.07	0.53

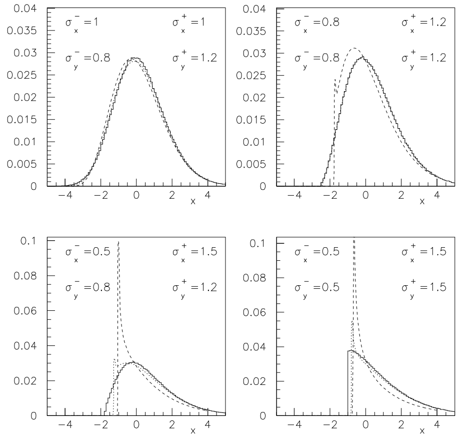


Figure 4: Examples of combined errors using Model 2.

Again the true curves (solid) are not well reproduced by the 'usual procedure' (dashed) but the curves with the correct cumulants (dotted) do a good job. (The sharp behaviour at the edge of the curves is due to the turning point of the parabola.)

2.7. Evaluating χ^2

For Model 1 the χ^2 contribution from a discrepancy δ is just $\delta^2/{\sigma^+}^2$ or $\delta^2/{\sigma^-}^2$ as appropriate. This is manifestly inelegant, especially for minimisation procedures as the value goes through zero.

For Model 2 one has

$$\delta = \sigma u + A\sigma u^2. \qquad (10)$$

This can be considered as a quadratic for u with solution which when squared gives u^2, the χ^2 contribution, as

$$u^2 = \frac{2 + 4A\frac{\delta}{\sigma} - 2(1 + 4A\frac{\delta}{\sigma})^{\frac{1}{2}}}{4A^2} \qquad (11)$$

This is not really exact, in that it only takes one branch of the solution, the one approximating to the straight line, and does not consider the extra possibility that the δ value could come from an improbable u value the other side of the turning point of the parabola. Given this imperfection it makes sense to expand the square root as a Taylor series, which, neglecting correction terms above the second power, leads to

$$\chi^2 = (\frac{\delta}{\sigma})^2\left(1 - 2A(\frac{\delta}{\sigma}) + 5A^2(\frac{\delta}{\sigma})^2\right). \qquad (12)$$

This provides a sensible form for χ^2 from asymmetric errors. It is important to keep the δ^4 term rather than stopping at δ^3 to ensure χ^2 stays positive! Adding higher orders does not have a great effect. We recommend it for consideration when it is required (e.g. in fitting parton distribution functions) to form a χ^2 from asymmetric errors

2.8. Weighted means

The 'best' estimate (i.e. unbiassed and with smallest variance) from several measurements x_i with different (symmetric) errors σ_i is given by a weighted sum with $w_i = 1/\sigma_i^2$. We wish to find the equivalent for asymmetric errors.

As noted earlier, when sampling from an asymmetric distribution the result is biassed towards the tail. The expectation value $\langle x \rangle$ is not the location parameter x. So for an unbiassed estimator one must take

$$\hat{x} = \sum w_i(x_i - b_i)/\sum w_i \qquad (13)$$

where

$$b = \frac{\sigma^+ - \sigma^-}{\sqrt{2\pi}} \quad \text{(Model 1)} \qquad b = \alpha \quad \text{(Model 2)} \qquad (14)$$

The variance of this is given by

$$V = \frac{\sum w_i^2 V_i}{\left(\sum w_i\right)^2} \qquad (15)$$

where V_i is the variance of the i^{th} measurement about its mean. Differentiating with respect to w_i to find the minimum gives

$$\frac{2w_iV_i}{\left(\sum w_j\right)^2} - \frac{2\sum w_j^2 V_j}{\left(\sum w_j\right)^3} = 0 \qquad \forall i \qquad (16)$$

which is satisfied by $w_i = 1/V_i$. This is the equivalent of the familiar weighting by $1/\sigma^2$. The weights are given, depending on the Model, by (see Equations 8 and 9)

$$V = \sigma^2 + (1 - \frac{2}{\pi})\alpha^2 \qquad \text{or} \qquad V = \sigma^2 + 2\alpha^2 \quad (17)$$

Note that this is not the Maximum Liklelihood estimator - writing down the likelihood in terms of the χ^2 and differentiating does not give a nice form - so in principle there may be better estimators, but they will not have the simple form of a weighted sum.

3. ASYMMETRIC STATISTICAL ERRORS

As explained earlier, (log) likelihood curves are used to obtain the maximum likelihood estimate for a parameter and also the 68% central interval – taken as the values at which $\ln L$ falls by $\frac{1}{2}$ from its peak. For large N this curve is a parabola, but for finite N it is generally asymmetric, and the two points are not equidistant about the peak.

The bias, if any, is not connected to the form of the curve, which is a likelihood and not a pdf. Evaluating a bias is done by integrating over the measured value not the theoretical parameter. We will assume for simplicity that these estimates are bias free. This means that when combining errors there will be no shift of the quoted value.

3.1. Combining asymmetric statistical errors

Suppose estimates $\hat{a}$ and $\hat{b}$ are obtained by this method for variables a and b. a could typically be an estimate of the total number of events in a signal region, and b the (scaled and negated) estimate of background, obtained from a sideband. We are interested in $u = a + b$, taking $\hat{u} = \hat{a} + \hat{b}$. What are the errors to be quoted on $\hat{u}$?

3.2. Likelihood functions known

We first consider the case where the likelihood functions $L_a(\vec{x}|a)$ and $L_b(\vec{x}|b)$ are given.

For the symmetric Gaussian case, the answer is well known. Suppose that the likelihoods are both Gaussian, and further that $\sigma_a = \sigma_b = \sigma$. The log likelihood term

$$\left(\frac{\hat{a} - a}{\sigma}\right)^2 + \left(\frac{\hat{b} - b}{\sigma}\right)^2 \qquad (18)$$

can be rewritten

$$\frac{1}{2}\left(\frac{\hat{a} + \hat{b} - (a + b)}{\sigma}\right)^2 + \frac{1}{2}\left(\frac{\hat{a} - \hat{b} - (a - b)}{\sigma}\right)^2 \quad (19)$$

so the likelihood is the product of Gaussians for $u = a + b$ and $v = a - b$, with standard deviations $\sqrt{2}\sigma$.

Picking a particular value of v, one can then trivially construct the 68% confidence region for u as $[\hat{u} - \sqrt{2}\sigma, \hat{u} + \sqrt{2}\sigma]$. Picking another value of v, indeed any other value of v, one obtains the same region for u. We can therefore say with 68% confidence that these limits enclose the true value of u, whatever the value of v. The uninteresting part of a and b has been 'parametrised away'. This is, of course, the standard result from the combination of errors formula, but derived in a frequentist way using Neyman-style confidence intervals. We could construct the limits on u by finding $\hat{u} + \sigma_u^+$ such that the integrated probability of a result as small as or smaller than the data be 16%, and similarly for σ_u^-, rather than taking the $\Delta \ln L = -\frac{1}{2}$ shortcut, and it would not affect the argument.

The question now is how to generalise this. For this to be possible the likelihood must factorise

$$L(\vec{x}|a, b) = L_u(\vec{x}|u)L_v(\vec{x}|v) \qquad (20)$$

with a suitable choice of the parameter v and the functions L_u and L_v. Then we can use the same argument: for any value of v the limits on u are the same, depending only on $L_u(\vec{x}|u)$. Because they are true for any v they are true for all v, and thus in general.

There are cases where this can clearly be done. For two Gaussians with $\sigma_a \neq \sigma_b$ the result is the same as above but with $v = a\sigma_b^2 - b\sigma_a^2$. For two Poisson distributions v is a/b. There are cases (with multiple peaks) where it cannot be done, but let us hope that these are artificially pathological.

On the basis that if it cannot be done, the question is unanswerable, let us assume that it is possible in the case being studied, and see how far we can proceed. Finding the form of v is liable to be difficult, and as it is not actually used in the answer we would like to avoid doing so. The limits on u are read off from the $\Delta \ln L(\vec{x}|u, v) = -\frac{1}{2}$ points where v can have any value provided it is fixed. Let us choose $v = \hat{v}$, the value at the peak. This is the value of v at which $L_v(v)$ is a maximum. Hence when we consider any other value of u, we can find $v = \hat{v}$ by finding the point at which the likelihood is a maximum, varying $a - b$, or a, or b, or any other combination, always keeping $a + b$ fixed. We can read the limits off a 1 dimensional plot of $\ln L_{max}(\vec{x}|u)$, where the 'max' suffix denotes that at each value of u we search the subspace to pick out the maximum value.

This generalises to more complicated situations. If $u = a + b + c$ we again scan the $\ln L_{max}(\vec{x}|u)$ function, where the subspace is now 2 dimensional.

3.3. Likelihood functions not completely known

In many cases the likelihood functions for a and b will not be given, merely estimates $\hat{a}$ and $\hat{b}$ and their asymmetric errors σ_a^+, σ_a^-, σ_b^+ and σ_b^-. All we can do is to use these to provide best guess functions $L_a(\vec{x}|a)$ and $L_b(\vec{x}|b)$. A parametrisation of suitable shapes, which for $\sigma^+ \sim \sigma^-$ approximate to a parabola, must be provided. Choosing a suitable parametrisation is not trivial. The obvious choice of introducing small higher-order terms fails as these dominate far from the peak. A likely candidate is:

$$\ln L(a) = -\frac{1}{2}\left(\frac{\ln(1+a/\gamma)}{\ln\beta}\right)^2 \qquad (21)$$

where $\beta = \sigma_+/\sigma_-$ and $\gamma = \frac{\sigma_+\sigma_-}{\sigma_+-\sigma_-}$. This describes the usual parabola, but with the x-axis stretched by an amount that changes linearly with distance. Figure 5 shows two illustrative results. The first is the Poisson likelihood from 5 observed events (solid line) for which the estimate using the $\Delta\ln L = \frac{1}{2}$ points is $\mu = 5^{+2.58}_{-1.92}$, as shown. The dashed line is that obtained inserting these numbers into Equation 21. The second considers a measurement of $x = 100\pm10$, of which the logarithm has been taken, to give a value $4.605^{+0.095}_{-0.105}$. Again, the solid line is the true curve and the dashed line the parametrisation. In both cases the agreement is excellent over the range $\approx \pm1\sigma$ and reasonable over the range $\approx \pm3\sigma$.

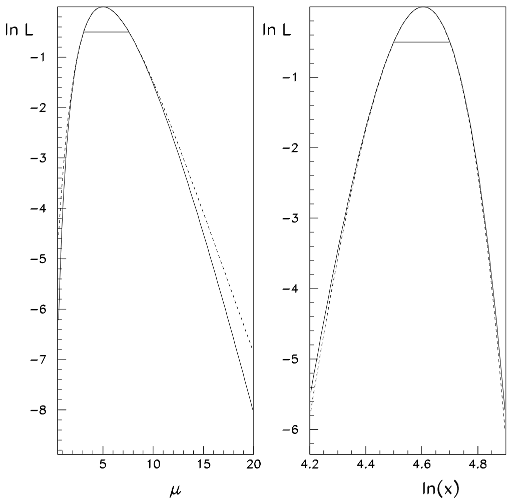


Figure 5: Approximations using Equation 21

To check the correctness of the method we can use the combination of two Poisson numbers, for which the result is known. First indications are that the errors obtained from the parametrisation are indeed closer to the true Poisson errors than those obtained from the usual technique.

3.4. Combination of Results

A related problem is to find the combined estimate $\hat{u}$ given estimates $\hat{a}$ and $\hat{b}$ (which have asymmetric errors). Here a and b could be results from different channels or different experiments. This can be regarded as a special case, constrained to $a = b$, i.e. $v = 0$, but this is rather contrived. It is more direct just to say that one uses the log likelihood which is the sum of the two separate functions, and determines the peak and the $\Delta\ln L = -\frac{1}{2}$ points from that. If the functions are known this is unproblematic, if only the errors are given then the same parametrisation technique can be used.

4. CONCLUSIONS

If asymmetric errrors cannot be avoided they need careful handling.

A method is suggested and a program provided for combining asymmetric systematic errors. It is not 'rigorously correct' but such perfection is impossible. Unlike the usual method, it is at least open about its assumptions and mathematically consistent.

Formulæ for χ^2 and weighted sums are given.

A method is proposed for combining asymmetric statistical errors if the likelihood functions are known. Work is in progress to enable it to be used given only the results and their errors.

Acknowledgments

The author a gratefully acknowledges the support of the Fulbright Foundation.

References

[1] D.E. Groom *et al.*, Eur. Phys. J. **C15** 1 (2000).
[2] W. T. Eadie et al, "Statistical Methods in Experimental Physics", North Holland, 1971.
[3] A.G. Frodesen *et al.* "Probablity and Statistics in Particle Physics", Universitetsforlaget Bergen-Oslo-Tromso (1979), pp 236-239.
[4] R. J. Barlow "Systematic Errors: Facts and Fictions" in Proc. Durham conference on Advanced Statistical Techniques in Particle Physics, M. R. Whalley and L. Lyons (Eds). IPPP/02/39. 2002.
[5] R. J. Barlow, "Asymmetric Systematic Errors" preprint MAN/HEP/03/02, ArXiv:physics/030613.
[6] G. D'Agostini "Bayesian Reasoning in Data Analysis: a Critical Guide", World Scientific (2003).
[7] The Shorter Oxford English Dictionary, Vol I (A-M) p 190 and p 551 of the 3rd edition (1977).

Constructing Ensembles of Pseudo-Experiments

Luc Demortier
The Rockefeller University, New York, NY 10021, USA

The frequentist interpretation of measurement results requires the specification of an ensemble of independent replications of the same experiment. For complex calculations of bias, coverage, significance, etc., this ensemble is often simulated by running Monte Carlo pseudo-experiments. In order to be valid, the latter must obey the Frequentist Principle and the Anticipation Criterion. We formulate these two principles and describe some of their consequences in relation to stopping rules, conditioning, and nuisance parameters. The discussion is illustrated with examples taken from high-energy physics.

1. INTRODUCTION

Many statistical analyses in physics are based on a frequency interpretation of probability. For example, the result of measuring a physical constant θ can be reported in the form of a $1-\alpha$ confidence interval $[X_1, X_2]$, with the understanding that if the measurement is replicated a large number of times, one will have $X_1 \leq \theta \leq X_2$ in a fraction $1-\alpha$ of the replications. This type of interpretation therefore requires the definition of a *reference set* of similar measurements:

> *The reference set of a measurement is the ensemble of experiments in which the actually performed experiment is considered to be embedded for the purpose of interpreting its results in a frequentist framework.*

A major appeal of frequentism among physicists is its empirical definition of probability. By the strong law of large numbers, probabilities can be approximated in finite ensembles, and such approximations converge to the true value as the ensemble size increases. In other words, frequentist confidence statements are experimentally verifiable.

Physicists use Monte Carlo generated ensembles in various applications: to check a fitting algorithm for the presence of bias, non-Gaussian pulls, or other pathologies; to calculate the coverage of confidence intervals or upper limits; to average out statistical fluctuations in order to isolate systematic effects; to calculate goodness-of-fit measures and significances; to design experiments; etc. When constructing ensembles to address these questions, one needs to pay attention to a number of subtle issues that arise in a frequentist framework: what is the correct stopping rule?; is it appropriate to condition, and if so, on what statistic?; how should nuisance parameters be handled?

The aim of this paper is to draw attention to these issues and to propose some recommendations where possible. We start by discussing basic frequentist principles in section 2 and illustrate them with an example of conditioning in section 3. The importance of stopping rules is argued in section 4. Finally, some purely frequentist methods to handle nuisance parameters are described in section 5.

2. FREQUENTIST PRINCIPLES

In order to deserve the label frequentist, a statistical procedure and its associated ensemble must satisfy two core principles, which we examine in the next two subsections.

2.1. The Frequentist Guarantee

The first principle states the aims of frequentism:

> **Frequentist Guarantee [1]:**
> *In repeated use of a statistical procedure, the long-run average actual accuracy should not be less than (and ideally should equal) the long-run average reported accuracy.*

To clarify this principle, we return to the $1-\alpha$ confidence interval procedure mentioned in the Introduction. Let $\mathcal{E}$ be an ensemble of intervals obtained by applying this procedure many times on different, independent data. The *actual accuracy* of an interval in $\mathcal{E}$ is 1 or 0: either the interval covers the true value of the parameter of interest, or it does not. The average actual accuracy is therefore simply the fraction of intervals in $\mathcal{E}$ that cover. On the other hand, the average *reported accuracy* is $1-\alpha$. The reported accuracy is often the same for all intervals in $\mathcal{E}$, but in some settings it is possible to report a different, data-dependent accuracy for each interval. Thus, averaging the reported accuracy is not necessarily a trivial operation. A procedure that satisfies the Frequentist Guarantee is said to have coverage.

In a sense, the Frequentist Guarantee is only weakly constraining, because it does not require a procedure to have coverage when applied to repeated measurements of the *same* quantity. To see how this is relevant, consider the construction of a 68% confidence interval for the mean μ of a Poisson distribution. One procedure is to take all μ values satisfying

$(n-\mu)^2/\mu \leq 1$, where n is the observed number of events. The resulting interval actually undercovers for many values of μ and overcovers for other values, so that the Frequentist Guarantee appears to be satisfied *on average*. To make this statement more precise we need a weighting function with which to carry out the average over μ. A simple proposal is to perform local smoothing of the coverage function, resulting in local average coverage [1].

Physicists may object to this notion of local average coverage on the grounds that they sometimes repeatedly measure a given constant of nature and are then interested in the coverage obtained for that particular constant, not in an average coverage over "nearby" constants. A possible answer is that one rarely measures the quantity of interest directly. Rather, one measures a combination of the quantity of interest with calibration constants, efficiencies, sample sizes, etc., all of which vary from one measurement to the next, so that an effective averaging does take place.

Finally, it could be argued that even Bayesians should subscribe to some form of the Frequentist Guarantee. If, over repeated use, a 95% credible Bayesian interval fails to cover the true value more than 30% of the time (say), then there must be something seriously wrong with that interval.

2.2. The Anticipation Criterion

Although the Frequentist Guarantee specifies how a statistical procedure should behave under many repetitions of a measurement, it does not indicate what constitutes a valid repetition, and hence a valid ensemble. To the extent that this question involves the notion of randomness, it is well beyond the scope of this paper. From a practical standpoint however, one would like to stipulate that all effects susceptible to interfere with that randomness must be recognized as such and included in the construction of the ensemble, i.e. "anticipated"[2]. Hence the second principle:

> **Anticipation Criterion:**
> *Ensembles must anticipate all elements of chance and all elements of choice of the actual experiments they serve to interpret.*

To clarify, "elements of chance" refers to statistical fluctuations of course, but also to systematic uncertainties when the latter come from nuisance parameters that are determined by auxiliary measurements. On the other hand, "elements of choice" refers to actions by experimenters, in particular how they decide to stop the experiment, and what decisions they make after stopping.

One can identify several levels of anticipation. At the highest level, the data collection and analysis methods, as well as the reference ensemble used to interpret results, are fully specified at the outset. They do not change once the data is observed. The reference ensemble is called "unconditional".

At the second highest level, the data collection and analysis methods are fully specified at the outset, but the reference ensemble is not. The latter will be fully determined once the data is observed, and is therefore "conditional". Although a conditional ensemble is not known before observing the data, it is a subset in a known partition of a known unconditional ensemble.

The lowest level of anticipation is occupied by Bayesian methods, which fully condition on the observed data. The reference ensemble collapses to a point and can therefore no longer be used as a reference.

As the level of anticipation decreases, the reference ensemble becomes smaller. A remarkable result is that within the second level of anticipation one can refine the conditioning partition to the point where it is possible to give a Bayesian interpretation to frequentist conclusions, and vice-versa [3].

3. CONDITIONING

To illustrate the interplay between anticipation and conditioning, we present here a famous example originally due to Cox [4]. Suppose we make one observation of a rare particle and wish to estimate its mass μ from the momenta of its decay products. For the sake of simplicity, assume that the estimator X of μ is normal with mean μ and variance σ^2. There is a 50% chance that the particle decays hadronically, in which case $\sigma = 10$; otherwise the particle decays leptonically and $\sigma = 1$. Consider the following 68% confidence interval procedures:

1. Unconditional
   If the particle decayed hadronically, report $X \pm \delta_h$, otherwise report $X \pm \delta_\ell$, where δ_h and δ_ℓ are chosen so as to minimize the expected length $\langle\delta\rangle = \delta_h + \delta_\ell$ subject to the constraint of 68% coverage. This yields $\delta_\ell = 2.20$ and $\delta_h = 5.06$. The expected length is 7.26.

2. Conditional
   If we condition on the decay mode, then the best interval is $X \pm 10$ if the particle decayed hadronically, and $X \pm 1$ otherwise. So the expected length is 11.0 in this case.

Note that in both cases we used all the information available: the measurement X as well as the decay mode. Both procedures are valid; the only difference between them is the reference frame. The unconditional ensemble includes both decay modes, whereas the conditional one only includes the observed decay mode.

The expected length is shorter for unconditional intervals than for conditional ones. Does this mean we should quote the former? If our aim is to report what we learned from the data we observed, then clearly we should report the conditional interval. Suppose indeed that we observed a hadronic decay. The unconditional interval width is then 10.12, compared to 20.0 for the conditional one. The reason the unconditional interval is shorter is that, if we could repeat the experiment, we might observe the particle decaying into the leptonic mode. However, this is irrelevant to the interpretation of the observation we actually made. This example illustrates a general feature of conditioning, that it usually increases expected length, and reduces power in test settings.

Another aspect of the previous example is that the conditioning statistic (the decay mode) is ancillary: its distribution does not depend on the parameter of interest (the particle mass). This is not always the case. Suppose for example that we are given a sample from a normal distribution with unit variance and unknown mean θ, and that we wish to test $H_0 : \theta = -1$ versus $H_1 : \theta = +1$. The standard symmetric Neyman-Pearson test based on the sample mean $\bar{X}$ as test statistic rejects H_0 if $\bar{X} > 0$. It makes no distinction between $\bar{X} = 0.5$ and $\bar{X} = 5$, even though in the latter case we certainly feel more confident in our rejection of H_0. Although $\bar{X}$ is not ancillary, it is possible to use it to calculate a conditional "measure of confidence" to help characterize one's decision regarding H_0 [5]. Unfortunately, a general theory for choosing such conditioning statistics does not exist.

4. STOPPING RULES

Stopping rules specify how an experiment is to be terminated. High-energy physics experiments are often sequential, so it is important to properly incorporate stopping rules in the construction of ensembles.

As a first example, consider the measurement of the branching fraction θ for the decay of a rare particle A into a particle B. Suppose we observe a total of $n = 12$ decays, $x = 9$ of which are $A \to B$ transitions, and the rest, $r = 3$, are $A \not\to B$ transitions. We wish to test $H_0 : \theta = 1/2$ versus $H_1 : \theta > 1/2$.

A possible stopping rule is to stop the experiment after observing a total number of decays n. The probability mass function (pmf) is then binomial:

$$f(x\,;\theta) \;=\; \binom{n}{x}\,\theta^x\,(1-\theta)^{n-x}, \qquad (1)$$

and the p value for testing H_0 is:

$$p_b \;=\; \sum_{i=9}^{12}\binom{12}{i}\,\theta^i\,(1-\theta)^{12-i} \;=\; 0.075. \qquad (2)$$

An equally valid stopping rule is to stop the experiment after observing a number r of $A \not\to B$ decays. Now the pmf is negative binomial:

$$f(x\,;\theta) \;=\; \binom{r+x-1}{x}\,\theta^x\,(1-\theta)^r, \qquad (3)$$

and the p value is:

$$p_{nb} \;=\; \sum_{i=9}^{\infty}\binom{2+i}{i}\,\theta^i\,(1-\theta)^3 \;=\; 0.0325. \qquad (4)$$

If we adopt a 5% threshold for accepting or rejecting H_0, we see that the binomial model leads to acceptance, whereas the negative binomial model leads to rejection.

Here is a more intriguing example [6]. Imagine a physicist working at some famous particle accelerator and developping a procedure to select collision events that contain a Higgs boson. Assume that the expected rate of background events accepted by this procedure is known very accurately. Applying his technique to a given dataset, the physicist observes 68 events and expects a background of 50. The (Poisson) probability for 50 to fluctuate up to 68 or more is 0.89%, and the physicist concludes that there is significant evidence against H_0, the background-only hypothesis, at the 1% level.

Is this conclusion correct? Perhaps the physicist just decided to take a single sample. But what would he have done if this sample had not yielded a significant result? Perhaps he would have taken another sample! So the real procedure the physicist was considering is actually of the form:

- Take a data sample, count the number n_1 of Higgs candidates, and calculate the expected background b;
- If $\mathbb{P}(N \geq n_1 \,|\, b) \leq \alpha$ then stop and reject H_0;
- Otherwise, take a second sample with the same expected background, count the number n_2 of Higgs candidates and reject H_0 if $\mathbb{P}(N \geq n_1 + n_2 \,|\, 2b) \leq \alpha$.

For this test procedure to have a level of 1%, α must be set at 0.67%. Since the *actual* data had a p value of 0.89%, the physicist should not have rejected H_0.

So now the physicist finds himself forced to take another sample. There are two interesting cases:

1. The second sample yields 57 candidate events, for a total of 125. The probability for the expected background (100 events now) to fluctuate up to 125 or more is 0.88% > 0.67%, so the result is not significant. However, the result would have been significant if the physicist had not stopped halfway through data taking to calculate the p value!

2. The second sample yields 59 candidate events, for a total of 127. The p value is now 0.52% and significance has been obtained, unless of course the physicist was planning to take a third sample in the event of no significance.

Bayesian methods are generally independent of the stopping rule. It is therefore somewhat ironic that frequentists, who start from an objective definition of probability, should end up with results that depend on the thought processes of the experimenter.

5. NUISANCE PARAMETERS

Most problems of inference involve nuisance parameters, i.e. uninteresting parameters that are incompletely known and therefore add to the overall uncertainty on the parameters of interest. To fix ideas, assume that we have a sample $\{x_1, \ldots, x_n\}$ whose probability density function (pdf) $f(\vec{x}\,;\,\mu, \nu)$ depends on a parameter of interest μ and a nuisance parameter ν, and that the latter can be determined from a separate sample $\{y_1, \ldots, y_m\}$ with pdf $g(\vec{y}\,;\,\nu)$. Correct inference about μ must then be derived from the joint pdf

$$h(\vec{x}, \vec{y}\,;\,\mu, \nu) \;\equiv\; f(\vec{x}\,;\,\mu, \nu)\; g(\vec{y}\,;\,\nu). \tag{5}$$

What is often done in practive however, is to first obtain a distribution $\pi(\nu)$ for ν, usually by combining measurement results with a sensible guess for the form of $\pi(\nu)$. Inference about μ is then based on:

$$h'(\vec{x}\,;\,\mu) \;\equiv\; \int f(\vec{x}\,;\,\mu, \nu)\, \pi(\nu)\, d\nu. \tag{6}$$

Although this technique borrows elements from both Bayesian and frequentist methodologies, it really belongs to neither and is more properly referred to as a hybrid non-frequentist/non-Bayesian approach.

We illustrate the handling of nuisance parameters with a simple p value calculation. Suppose that a search for a new particle ends with a sample of $n_0 = 12$ candidates over a separately measured background of $\nu_0 = 5.7{\pm}0.47$, where we ignore the uncertainty on the standard error 0.47. Let μ be the unknown expected number of new particles among the 12 candidates. We wish to test $H_0 : \mu = 0$ versus $H_1 : \mu > 0$.

A typical model for this problem consists of a Poisson density for the number of observed candidates and a Gaussian for the background measurement. Using equation (6) with a simple Monte Carlo integration routine, one obtains a p value of $\sim 1.6\%$. For reference, when there is no uncertainty on ν_0 the p value is $\sim 1.4\%$.

While there are many purely frequentist approaches to the elimination of nuisance parameters, few of these have general applicability. Concentrating on the latter, we discuss the likelihood ratio and confidence interval methods in the next two subsections.

5.1. Likelihood Ratio Method

The likelihood ratio statistic λ is defined by:

$$\lambda \;=\; \frac{\sup\limits_{\substack{\mu=0\\ \nu\geq 0}} \mathcal{L}(\mu, \nu \,|\, n_0, \nu_0)}{\sup\limits_{\substack{\mu\geq 0\\ \nu\geq 0}} \mathcal{L}(\mu, \nu \,|\, n_0, \nu_0)}, \tag{7}$$

where, for $\nu_0 \gg \Delta\nu$:

$$\mathcal{L}(\mu, \nu \,|\, n_0, \nu_0) \;\propto\; \frac{(\mu+\nu)^{n_0}}{n_0!}\; e^{-\mu-\nu}\; e^{-\frac{1}{2}\left(\frac{\nu-\nu_0}{\Delta\nu}\right)^2}.$$

Simple calculus leads to:

$$\begin{aligned} -2\ln\lambda \;&=\; 2\left(n_0 \ln\tfrac{n_0}{\hat{\nu}} + \hat{\nu} - n_0\right) + \left(\tfrac{\hat{\nu}-\nu_0}{\Delta\nu}\right)^2 \quad \text{if } n_0 > \nu_0,\\ &=\; 0 \qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad\;\, \text{if } n_0 \leq \nu_0, \end{aligned}$$

$$\text{with: } \hat{\nu} \;=\; \tfrac{\nu_0-\Delta\nu^2}{2} + \sqrt{\left(\tfrac{\nu_0-\Delta\nu^2}{2}\right)^2 + n_0\,\Delta\nu^2}.$$

Since λ depends on n_0 and ν_0, its distribution under H_0 depends on the true expected background ν_t. A natural simplification is to examine the limit $\nu_t \to \infty$. Application of theorems describing the asymptotic behavior of $-2\ln\lambda$ must take into account that for $n_0 < \nu_0$ the analytical maximum of the likelihood lies outside the physical region $\mu \geq 0$. The correct asymptotic result is that, under H_0, half a unit of probability is carried by the singleton $\{-2\ln\lambda = 0\}$, and the other half is distributed as a chisquared with one degree of freedom over $0 < -2\ln\lambda < +\infty$.

For our example the expected background is only 5.7 particles however, so one may wonder how close this is to the asymptotic limit. Here is an algorithm to check this. Choose a true number of background events ν_t and repeat the following three steps a large number of times:

1. Generate a Gaussian variate ν_0 with mean ν_t and width $\Delta\nu$;
2. Generate a Poisson variate n_0 with mean ν_t;
3. Calculate λ from the generated ν_0 and n_0.

The p value is then equal to the fraction of pseudo-experiments that yield a likelihood ratio λ smaller than the λ_0 obtained from the observed data.

Note that this algorithm does not "smear" the true value of any parameter, in contrast with equation (6). The price for this is that the result depends on the choice of ν_t. For ν_t varying from 0.5 to 50, the p value ranges from ~ 0.48 to $\sim 1.2\%$. A general prescription for dealing with a p value dependence on nuisance parameters is to use the so-called supremum p value:

$$p_{\text{sup}} \;=\; \sup_{\nu}\; \mathbb{P}(-2\ln\lambda \geq -2\ln\lambda_0 \,|\, \mu, \nu)\big|_{\mu=0}$$

From a frequentist point of view, the supremum p value is *valid*, in the sense that:

$$\mathbb{P}(p_{\text{sup}} \leq \alpha) \ \leq \ \alpha, \quad \text{for each } \alpha \in [0, 1], \qquad (8)$$

regardless of the true value of the nuisance parameter. Although it is often difficult to calculate a supremum, in this case it turns out to equal the asymptotic limit to a good approximation. In our example $-2 \ln \lambda_0 = 5.02$ and corresponds to $p_{\text{sup}} \approx p_\infty = 1.25\%$.

As the attentive reader will have noticed, the p value is smaller for $\Delta\nu = 0.47$ than for $\Delta\nu = 0$. This is a consequence of the discreteness of Poisson statistics; it does not violate inequality (8) because p_{sup} actually overcovers a little when $\Delta\nu = 0$. To avoid the bias resulting from this overcoverage, the use of mid-p values is sometimes advocated for the purpose of comparing or combining p values [7].

5.2. Confidence Interval Method

The supremum p value introduced in the previous section can be defined for any test statistic, although it will not always give useful results. If for example in our new particle search we take the total number n_0 of observed candidates as test statistic, the p value will be 100% since the background ν is unbounded from above. A more satisfactory method proceeds as follows [8, 9]. First, construct a $1 - \beta$ confidence interval C_β for the nuisance parameter ν, then maximize the p value over that interval, and finally correct the result for the fact that $\beta \neq 0$:

$$p_\beta \ = \ \sup_{\nu \in C_\beta} \ \mathbb{P}(N \geq n_0 \,|\, \mu, \nu)\big|_{\mu=0} \ + \ \beta.$$

It can be shown that this is also a *valid* p value.

For the sake of illustration with our example, we consider three choices of β and construct the corresponding $1 - \beta$ confidence intervals for ν_t:

$$\begin{array}{lll} 1-\beta = \ 99.5\% : & C_{0.005} & = [4.38\,,\, 7.02] \\ 1-\beta = \ 99.9\% : & C_{0.001} & = [4.15\,,\, 7.25] \\ 1-\beta = \ 99.99\% : & C_{0.0001} & = [3.87\,,\, 7.53] \end{array}$$

To calculate the p value, a good choice of statistic is the maximum likelihood estimator of the signal, i.e. $\hat{s} \equiv n_0 - \nu_0$. Under H_0, the survivor function of $\hat{s}$ is given by:

$$\mathbb{P}(S \geq \hat{s}) \ = \ \sum_{k=0}^{\infty} \frac{1 + \text{erf}\left(\frac{k - \nu_t - \hat{s}}{\sqrt{2}\,\Delta\nu}\right)}{1 + \text{erf}\left(\frac{\nu_t}{\sqrt{2}\,\Delta\nu}\right)} \ \frac{\nu_t^k}{k!} \ e^{-\nu_t}$$

We then find:

$$\begin{array}{llll} 1-\beta = \ 99.5\% : & p_\beta \ = \ 1.6\% \ +0.5\% & = \ 2.1\% \\ 1-\beta = \ 99.9\% : & p_\beta \ = \ 1.7\% \ +0.1\% & = \ 1.8\% \\ 1-\beta = \ 99.99\% : & p_\beta \ = \ 1.88\% + 0.01\% & = \ 1.89\% \end{array}$$

An important point about the confidence interval method is that, in order to satisfy the Anticipation Criterion, the value of β and the confidence set C_β must be specified before looking at the data. Since p_β is never smaller than β, the latter should be small. In particular, if p_β is used in a level-α test, then β must be smaller than α for the test to be useful.

6. SUMMARY

From the practical point of view of someone analyzing data, the most critical property of frequentist ensembles is their "anticipatoriness." This requires that all the structural elements of an analysis (i.e. test sizes, interval procedures, bin boundaries, stopping rules, etc.) be in place before looking at the data. The only exception to this requirement occurs in situations where conditioning is both possible and appropriate. Even in that case, the conditioning partition itself must be specified beforehand.

References

[1] M. J. Bayarri and J. O. Berger, "The interplay of Bayesian and frequentist analysis," `http://www.stat.duke.edu/~berger/papers/interplay.html`, April 2003.

[2] L. D. Brown, "Comment on "Conditional confidence statements and confidence estimators"," *J. Amer. Statist. Assoc.* **72**, 810–812 (1977).

[3] J. Berger, B. Boukai, and Y. Wang, "Unified frequentist and Bayesian testing of a precise hypothesis," *Statist. Sci.* **12**, 133–160 (1997).

[4] D. R. Cox, "Some problems connected with statistical inference," *Ann. Math. Statist.* **29**, 357–372 (1958).

[5] J. Kiefer, "Conditional confidence statements and confidence estimators," *J. Amer. Statist. Assoc.* **72**, 789–808 (1977).

[6] J. O. Berger and D. A. Berry, "The relevance of stopping rules in statistical inference," in *Statistical Decision Theory and Related Topics IV* (S. S. Gupta and J. O. Berger, eds.), vol. 1, pp. 29–72, Springer-Verlag, 1988.

[7] G. Berry and P. Armitage, "Mid-P confidence intervals: a brief review," *Statistician* **44**, 417–423 (1995).

[8] R. L. Berger and D. D. Boos, "P values maximized over a confidence set for the nuisance parameter," *J. Amer. Statist. Assoc.* **89**, 1012–1016 (1994).

[9] M. J. Silvapulle, "A test in the presence of nuisance parameters," *J. Amer. Statist. Assoc.* **91**, 1690–1693, (1996); Correction: *ibid.* **92**, 801 (1997).

Frequentist Hypothesis Testing with Background Uncertainty

K.S. Cranmer
University of Wisconsin-Madison, Madison, WI 53706, USA

We consider the standard Neyman-Pearson hypothesis test of a signal-plus-background hypothesis and background-only hypothesis in the presence of uncertainty on the background-only prediction. Surprisingly, this problem has not been addressed in the recent conferences on statistical techniques in high-energy physics – although the its confidence-interval equivalent has been. We discuss the issues of power, similar tests, coverage, and ordering rules. The method presented is compared to the Cousins-Highland technique, the ratio of Poisson means, and "profile" method.

1. INTRODUCTION

In the last five years there have been several conferences on statistics for particle physics. Much of the emphasis of these conferences were on limit setting and the Feldman-Cousins "unified approach", the quintessential frequentist method based on the Neyman construction. As particle physicists prepare for the Large Hadron Collider (LHC) at CERN, we will need to reexamine our list of statistical tools in the context of discovery. In fact, there has been no presentation at these statistical conferences on frequentist hypothesis testing in the presence of uncertainty on the background.

In Section 2 we will review the Neyman-Pearson theory for testing between two simple hypotheses, and examine the impact of background uncertainty in Section 3. In Sections 4- 5 we will present a fully frequentist method for hypothesis testing with background uncertainty based on the Neyman Construction. In the remainder of the text we will present an example and compare this method to other existing methods.

2. SIMPLE HYPOTHESIS TESTING

In the case of Simple Hypothesis testing, the Neyman-Pearson theory (which we review briefly for completeness) begins with two Hypotheses: the null hypothesis H_0 and the alternate hypothesis H_1 [1]. These hypotheses are called *simple* because they have no free parameters. Predictions of some physical observable x can be made with these hypotheses and described by the likelihood functions $L(x|H_0)$ and $L(x|H_1)$ (for simplicity, think of x as the number of events observed).

Next, one defines a region $W \in I$ such that if the data fall in W we accept the H_0 (and reject H_1). Conversely, if the data fall in $I - W$ we reject H_0 and accept the H_1. The probability to commit a Type I error is called the *size* of the test and is given by

$$\alpha = \int_{I-W} L(x|H_0)dx. \qquad (1)$$

The probability to commit a Type II error is given by

$$\beta = \int_{W} L(x|H_1)dx. \qquad (2)$$

Finally, the Neyman-Pearson lemma tells us that the region W of size α which minimizes the rate of Type II error (maximizes the power) is given by

$$W = \left\{ x \;\middle|\; \frac{L(x|H_1)}{L(x|H_0)} > k_\alpha \right\}. \qquad (3)$$

3. NUISANCE PARAMETERS

Within physics, the majority of the emphasis on statistics has been on limit setting – which can be translated to hypothesis testing through a well known dictionary [1]. When one includes nuisance parameters θ_s (parameters that are not of interest or not observable to the experimenter) into the calculation of a confidence interval, one must ensure coverage for every value of the nuisance parameter. When one is interested in hypothesis testing, there is no longer a physics parameter θ_r to cover, instead one must ensure the rate of Type I error is bounded by some predefined value. Analogously, when one includes a nuisance parameters in the null hypothesis, one must ensure that the rate of Type I error is bounded for every value of the nuisance parameter. Ideally one can find an acceptance region W which has the same size for all values of the nuisance parameter (*i.e.* a similar test). Furthermore, the power of a region W also depends on the nuisance parameter; ideally, we would like to maximize the power for all values of the nuisance parameter (*i.e.* Uniformly Most Powerful). Such tests do not exist in general.

In this note, we wish to address how the standard hypothesis test is modified by uncertainty on the background prediction. The uncertainty in the background prediction represents the presence of a nuisance parameter: for example, let us assume it is the expected background b. Typically, an auxiliary, or side-band, measurement is made to provide a handle on the nuisance parameter. Let us generically call that measurement M and $L(M|H_0, b)$ the prediction of that measurement given the null hypothesis with nuisance parameter b. In Section 8 we address the special case that $L(M|H_0, b)$ is a Poisson distribution.

4. THE NEYMAN-CONSTRUCTION

Usually one does not consider an explicit Neyman construction when performing hypothesis testing between two simple hypotheses; though one exists implicitly. Because of the presence of the nuisance parameter, the implicit Neyman construction must be made explicit and the dimensionality increased. The basic idea is that for each value of the nuisance parameters θ_s, one must construct an acceptance interval (for H_0) in a space which includes their corresponding auxiliary measurements M, and the original test statistic x which was being used to test H_0 against H_1.

For the simple case introduced in the previous section, this requires a three-dimensional construction with b, M, and x. For each value of b, one must construct a two-dimensional acceptance region W_b of size α (under H_0). If an experiment's data (x_0, M_0) fall into an acceptance region W_b, then one cannot exclude the null hypothesis with $100(1-\alpha)\%$ confidence. Conversely, to reject the null hypothesis (*i.e.* claim a discovery) the data must not lie in any acceptance region W_b. Said yet another way, to claim a discovery, the confidence interval for the nuisance parameter(s) must be empty (when the construction is made assuming the null hypothesis).

5. THE ORDERING RULE

The basic criterion for discovery was discussed abstractly in the previous section. In order to provide an actual calculation, one must provide an ordering rule: an algorithm which decides how to chose the region W_b. Recall, that there the constraint on Type I error does not uniquely specify an acceptance region for H_0. In the Neyman-Pearson lemma, it is the alternate hypothesis H_1 that breaks the symmetry between possible acceptance regions. Also in the unified approach, it is the likelihood ratio that is used as an ordering rule [2].

At the Workshop on conference limits at FermiLab, Feldman showed that Unified Method with Nuisance Parameters is in Kendall's Theory (the chapter on likelihood ratio tests & test efficiency) [3]. The notation used by Kendall is given in Table I. Also, Kendall identifies H_0 with $\theta_r = \theta_{r0}$ and H_1 with $\theta_r \neq \theta_{r0}$.

Let us briefly quote from Kendall:

> "Now consider the Likelihood Ratio
>
> $$l = \frac{L(x|\theta_{r0}, \hat{\hat{\theta}}_s)}{L(x|\hat{\theta}_r, \hat{\theta}_s)} \tag{4}$$

Variable	Meaning
θ_r	physics parameters
θ_s	nuisance parameters
$\hat{\theta}_r, \hat{\theta}_s$	unconditionally maximize $L(x|\hat{\theta}_r, \hat{\theta}_s)$
$\hat{\hat{\theta}}_s$	conditionally maximize $L(x|\theta_{r0}, \hat{\hat{\theta}}_s)$

Table I The notation used by Kendall for likelihood tests with nuisance parameters

> Intuitively l is a reasonable test statistic for H_0: it is the maximum likelihood under H_0 as a fraction of its largest possible value, and large values of l signify that H_0 is reasonably acceptable."

Feldman uses this chapter as motivation for the profile method (see Section 9), though in Kendall's book the same likelihood ratio is used as an ordering rule *for each value of the nuisance parameter.*

The author tried simple variations on this ordering rule before rediscovering it as written. It is worth pointing out that Eq. 4 is independent of the nuisance parameter b; however, the contour of l_α which provides an acceptance region of size α is not necessarily independent of b. It is also worth pointing out that $\hat{\theta}_r$ and $\hat{\theta}_s$ do not consider the null hypothesis – if they did, the region in which $l = 1$ may be larger than $(1-\alpha)$. Finally, if one uses θ_s instead of $\hat{\theta}_s$ or $\hat{\hat{\theta}}_s$, one will not obtain tests which are approximately similar.

6. AN EXAMPLE

Let us consider the case when the nuisance parameter is the expected number of background events b and M is an auxiliary measurement of b. Furthermore, let us assume that we have a absolute prediction of the number of signal events s. For our test statistic we choose the number of events observed x which is Poisson distributed with mean $\mu = b$ for H_0 and $\mu = s + b$ for H_1. In the construction there are no assumptions about $L(M|H_0, b)$ – it could be some very complicated shape relating particle identification efficiencies, Monte Carlo extrapolation, etc. In the case where $L(M|H_0, b)$ is a Poisson distribution, other solutions exist (see Section 8). For our example, let us take $L(M|H_0, b)$ to be a Normal distribution centered on b with standard deviation Δb, where Δ is some relative systematic error. Additionally, let us assume that we can factorize $L(x, M|H, b) = L(x|H, b)L(M|b)$ (where H is either H_0 or H_1).

For our example problem, we can re-write the ordering rule in Eq. 4 as

$$l = \frac{L(x, M|H_0, \hat{\hat{b}})}{L(x, M|H_1, \hat{b})}, \tag{5}$$

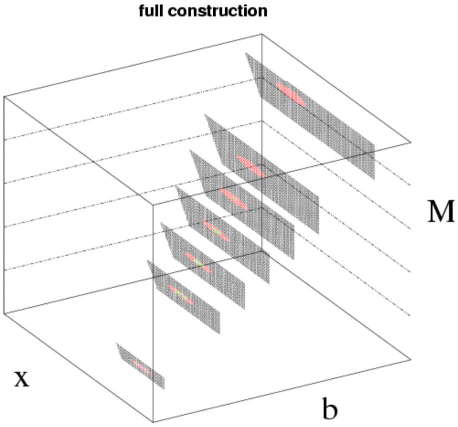


Figure 1: The Neyman construction for a test statistic x, an auxiliary measurement M, and a nuisance parameter b. Vertical planes represent acceptance regions W_b for H_0 given b. The condition for discovery corresponds to data (x_0, M_0) that do not intersect any acceptance region. The contours of $L(x, M|H_0, b)$ are in color.

where $\hat{b}$ conditionally maximizes $L(x, M|H_1, b)$ and $\hat{\hat{b}}$ conditionally maximizes $L(x, M|H_0, b)$.

Now let us take $s = 50$ and $\Delta = 5\%$, both of which could be determined from Monte Carlo. In our toy example, we collect data $M_0 = 100$. Let $\alpha = 2.85 \cdot 10^{-7}$, which corresponds to 5σ. The question now is how many events x must we observe to claim a discovery?[1] The condition for discovery is that (x_0, M_0) do not lie in any acceptance region W_b. In Fig. 1 a sample of acceptance regions are displayed. One can imagine a horizontal plane at $M_0 = 100$ slicing through the various acceptance regions. The condition for discovery is that $x_0 > x_{\max}$ where $x_{\max}$ is the maximal x in the intersection.

There is one subtlety which arises from the ordering rule in Eq. 5. The acceptance region $W_b = \{(x, M) \mid l > l_\alpha\}$ is bounded by a contour of the likelihood ratio and must satisfy the constraint of size: $\int_{W_b} L(x, M|H_0, b) = (1 - \alpha)$. While it is true that the likelihood is independent of b, the constraint on size *is* dependent upon b. Similar tests are achieved when l_α is independent of b. The contours of the likelihood ratio are shown in Fig. 2 together with contours of $L(x, M|H_0, b)$. While tests are roughly similar for $b \approx M$, similarity is violated for $M \ll b$. This violation should be irrelevant because clearly $b \ll M$ should not be accepted. This problem can be avoided by clipping the acceptance region around $M = b \pm N\Delta b$, where N is sufficiently large (≈ 10) to have negligible affect on the size of the acceptance region. Fig. 1 shows the acceptance region with this slight modification.

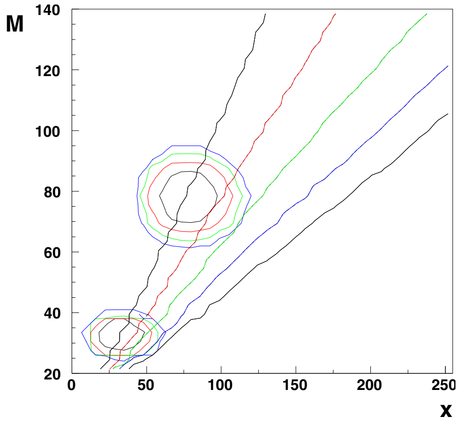


Figure 2: Contours of the likelihood $L(x, M|H_0, b)$ are shown as concentric ellipses for $b = 32$ and $b = 80$. Contours of the likelihood ratio in Eq. 5 are shown as diagonal lines. This figure schematically illustrates that if one chooses acceptance regions based solely on contours of the likelihood ratio, that similarity is badly violated. For example, data $M = 80, x = 130$ would be considered part of the acceptance region for $b = 32$, even though it should clearly be ruled out.

In the case where $s = 50$, $\Delta = 5\%$, and $M_0 = 100$, one must observe 167 events to claim a discovery. While no figure is provided, the range of b consistent with $M_0 = 100$ (and no constraint on x) is $b \in [68, 200]$. In this range, the tests are similar to a very high degree.

7. THE COUSINS-HIGHLAND TECHNIQUE

The Cousins-Highland approach to hypothesis testing is quite popular [4] because it is a simple smearing on the nuisance parameter [5]. In particular, the background-only hypothesis $L(x|H_0, b)$ is transformed from a compound hypothesis with nuisance parameter b to a simple hypothesis $L'(x|H_0)$ by

$$L'(x|H_0) = \int_b L(x|H_0, b) L(b) db, \tag{6}$$

where $L(b)$ is typically a normal distribution. The problem with this method is largely philosophical: $L(b)$ is meaningless in a frequentist formalism. In a Bayesian formalism one can obtain $L(b)$ by considering $L(M|b)$ and inverting it with the use of Bayes's theorem and the *a priori* likelihood for b. Typically, $L(M|b)$ is normal and one assumes a flat prior on b.

In the case where $s = 50$, $L(b)$ is a normal distribution with mean $\mu = M_0 = 100$ and standard deviation $\sigma = \Delta M_0 = 5$, one must observe 161 events to claim a discovery. Initially, one might think that 161 is quite

[1] In practice, one would measure x_0 and M_0 and then ask, “have we made a discovery?”. For the sake of explanation, we have broken this process into two pieces.

close to 167; however, they differ at the 4% level and the methods are only considering a $\Delta = 5\%$ effect. Still worse, if H_0 is true (say $b_t = 100$) and one can claim a discovery with the Cousins-Highland method ($x_0 > 161$), the chance that one could not claim a discovery with the fully frequentist method ($x_0 < 167$) is $\approx 95\%$. Similarly, if H_1 is true and one can claim a discovery with the Cousins-Highland method, the chance that one could not claim a discovery with the fully frequentist method is $\approx 50\%$. Even practically, there is quite a difference between these two methods.

8. THE RATIO OF POISSON MEANS

During the conference, J. Linnemann presented results on the ratio of Poisson means. In that case, one considers a background and a signal process, both with unknown means. By making "on-source" (*i.e.* x) and "off-source" (*i.e.* M) measurements one can form a confidence interval on the ratio $\lambda = s/b$. If the $100(1-\alpha)\%$ confidence interval for λ does not include 0, then one could claim discovery. This approach does take into account uncertainty on the background; however, it is restricted to the case in which $L(M|b)$ is a Poisson distribution.

There are two variations on this technique. The first technique has been known for quite some time and was first brought to physics in Ref. [6]. This approach conditions on $x+M$, which allows one to tackle the problem with the use of a binomial distribution. Later, Cousins improved on these limits by removing the conditioning and considering the full Neyman construction [7]. Cousins paper has an excellent review of the literature for those interested in this technique.

9. THE PROFILE METHOD

As was mentioned in Section 3 the likelihood ratio in Eq. 4 is independent of the nuisance parameters. If it were not for the violations in similarity between tests, one would only need to perform the construction for one value of the nuisance parameters. Clearly, $\hat{\hat{\theta}}_s$ is an appropriate choice to perform the construction. This is the logic behind the profile method. It should be pointed out that the profile method is an approximation to the full Neyman construction; though a particularly good one. In the example above with $x_0 = 167$, $M_0 = 100$, the construction would be made at $b = \hat{\hat{b}} = 117$ which gives the identical result as the fully frequentist method.

The main advantage to the profile method is that of speed and scalability. Instead of performing the construction for every value of the nuisance parameters, one must only perform the construction once. For many variables, the fully frequentist method is not scalable if one naïvely loops over on a fixed grid. However, Monte Carlo sampling the nuisance parameters does not suffer from the curse of dimensionality and serves as a more robust approximation of the full construction than the profile method.

10. CONCLUSION

We have presented a fully frequentist method for hypothesis testing. The method consists of a Neyman construction in each of the nuisance parameters, their corresponding auxiliary measurements, and the test statistic that was originally used to test H_0 against H_1. We have chosen as an ordering rule the likelihood ratio with the nuisance parameters conditionally maximized to their respective hypotheses. With a slight modification, this ordering rule produces tests that are approximately similar. We have compared this method to the most common methods in the field. This method is philosophically more sound than the Cousins-Highland technique and more general than the ratio of Poisson means. This method can be made computationally less intensive either with Monte Carlo sampling of the nuisance parameters or by the approximation known as the profile method.

Acknowledgments

This work was supported by a graduate research fellowship from the National Science Foundation and US Department of Energy Grant DE-FG0295-ER40896. The author would like to thank L. Lyons, R.D. Cousins, and G. Feldman for useful feedback.

References

[1] J.K Stuart, A. Ord and S. Arnold. *Kendall's Advanced Theory of Statistics, Vol 2A (6th Ed.).* Oxford University Press, New York, 1994.

[2] Gary J. Feldman and Robert D. Cousins. A unified approach to the classical statistical analysis of small signals. *Phys. Rev.*, D57:3873–3889, 1998.

[3] Gary J. Feldman. Multiple measurements and parameters in the unified approach, 2000. Workshop on Confidence Limits, FermiLab.

[4] Search for the standard model Higgs boson at LEP. *Phys. Lett.*, B565:61–75, 2003.

[5] R.D. Cousins and V.L. Highland. Incorporating systematic uncertainties into an upper limit. *Nucl. Instrum. Meth.*, A320:331–335, 1992.

[6] F. James and M. Roos. Errors on ratios of small numbers of events. *Nucl. Phys.*, B 172:475–480, 1980.

[7] R.D. Cousins. Improved central confidence intervals for the ratio of Poisson means. *Nucl. Instrum. and Meth. in Phys. Res.*, A 417:391–399, 1998.

Likelihood Inference in the Presence of Nuisance Parameters

N. Reid, D.A.S. Fraser
Department of Statistics, University of Toronto, Toronto Canada M5S 3G3

We describe some recent approaches to likelihood based inference in the presence of nuisance parameters. Our approach is based on plotting the likelihood function and the p-value function, using recently developed third order approximations. Orthogonal parameters and adjustments to profile likelihood are also discussed. Connections to classical approaches of conditional and marginal inference are outlined.

1. INTRODUCTION

We take the view that the most effective form of inference is provided by the observed likelihood function along with the associated p-value function. In the case of a scalar parameter the likelihood function is simply proportional to the density function. The p-value function can be obtained exactly if there is a one-dimensional statistic that measures the parameter. If not, the p-value can be obtained to a high order of approximation using recently developed methods of likelihood asymptotics. In the presence of nuisance parameters, the likelihood function for a (one-dimensional) parameter of interest is obtained via an adjustment to the profile likelihood function. The p-value function is obtained from quantities computed from the likelihood function using a canonical parametrization $\varphi = \varphi(\theta)$, which is computed locally at the data point. This generalizes the method of eliminating nuisance parameters by conditioning or marginalizing to more general contexts. In Section 2 we give some background notation and introduce the notion of orthogonal parameters. In Section 3 we illustrate the p-value function approach in a simple model with no nuisance parameters. Profile likelihood and adjustments to profile likelihood are described in Section 4. Third order p-values for problems with nuisance parameters are described in Section 5. Section 6 describes the classical conditional and marginal likelihood approach.

2. NOTATION AND ORTHOGONAL PARAMETERS

We assume our measurement(s) y can be modelled as coming from a probability distribution with density or mass function $f(y;\theta)$, where $\theta = (\psi, \lambda)$ takes values in R^d. We assume ψ is a one-dimensional parameter of interest, and λ is a vector of nuisance parameters. If there is interest in more than one component of θ, the methods described here can be applied to each component of interest in turn. The likelihood function is

$$L(\theta) = L(\theta; y) = c(y) f(y;\theta); \qquad (1)$$

it is defined only up to arbitrary multiples which may depend on y but not on θ. This ensures in particular that the likelihood function is invariant to one-to-one transformations of the measurement(s) y. In the context of independent, identically distributed sampling, where $y = (y_1, \ldots, y_n)$ and each y_i follows the model $f(y;\theta)$ the likelihood function is proportional to $\Pi f(y_i;\theta)$ and the log-likelihood function becomes a sum of independent and identically distributed components:

$$\ell(\theta) = \ell(\theta; y) = \Sigma \log f(y_i;\theta) + a(y). \qquad (2)$$

The maximum likelihood estimate $\hat{\theta}$ is the value of θ at which the likelihood takes its maximum, and in regular models is defined by the score equation

$$\ell'(\hat{\theta}; y) = 0. \qquad (3)$$

The observed Fisher information function $j(\theta)$ is the curvature of the log-likelihood:

$$j(\theta) = -\ell''(\theta) \qquad (4)$$

and the expected Fisher information is the model quantity

$$i(\theta) = E\{-\ell''(\theta)\} = \int -\ell''(\theta; y) f(y;\theta) dy. \qquad (5)$$

If y is a sample of size n then $i(\theta) = O(n)$.

In accord with the partitioning of θ we partition the observed and expected information matrices and use the notation

$$i(\theta) = \begin{pmatrix} i_{\psi\psi} & i_{\psi\lambda} \\ i_{\lambda\psi} & i_{\lambda\lambda} \end{pmatrix} \qquad (6)$$

and

$$i^{-1}(\theta) = \begin{pmatrix} i^{\psi\psi} & i^{\psi\lambda} \\ i^{\lambda\psi} & i^{\lambda\lambda} \end{pmatrix}. \qquad (7)$$

We say ψ is *orthogonal* to λ (with respect to expected Fisher information) if $i_{\psi\lambda}(\theta) = 0$. When ψ is scalar a transformation from (ψ, λ) to $(\psi, \eta(\psi, \lambda))$ such that ψ is orthogonal to η can always be found (Cox and Reid [1]). The most directly interpreted consequence of parameter orthogonality is that the maximum likelihood estimates of orthogonal components are asymptotically independent.

Example 1: ratio of Poisson means Suppose y_1 and y_2 are independent counts modelled as Poisson with mean λ and $\psi\lambda$, respectively. Then the likelihood function is

$$L(\psi, \lambda; y_1, y_2) = e^{-\lambda(1+\psi)}\psi^{y_2}\lambda^{y_1+y_2}$$

and ψ is orthogonal to $\eta(\psi, \lambda) = \lambda(\psi + 1)$. In fact in this example the likelihood function factors as $L_1(\psi)L_2(\eta)$, which is a stronger property than parameter orthogonality. The first factor is the likelihood for a binomial distribution with index $y_1 + y_2$ and probability of success $\psi/(1+\psi)$, and the second is that for a Poisson distribution with mean η.

Example 2: exponential regression Suppose $y_i, i = 1, \ldots, n$ are independent observations, each from an exponential distribution with mean $\lambda \exp(-\psi x_i)$, where x_i is known. The log-likelihood function is

$$\ell(\psi, \lambda; y) = -n \log\lambda + \psi\Sigma x_i - \lambda^{-1}\Sigma y_i \exp(\psi x_i) \quad (8)$$

and $i_{\psi\lambda}(\theta) = 0$ if and only if $\Sigma x_i = 0$. The stronger property of factorization of the likelihood does not hold.

3. LIKELIHOOD INFERENCE WITH NO NUISANCE PARAMETERS

We assume now that θ is one-dimensional. A plot of the log-likelihood function as a function of θ can quickly reveal irregularities in the model, such as a non-unique maximum, or a maximum on the boundary, and can also provide a visual guide to deviance from normality, as the log-likelihood function for a normal distribution is a parabola and hence symmetric about the maximum. In order to calibrate the log-likelihood function we can use the approximation

$$r(\theta) = \text{sign}(\hat{\theta} - \theta)[2\{\ell(\hat{\theta}) - \ell(\theta)\}]^{1/2} \mathrel{\dot\sim} N(0, 1), \quad (9)$$

which is equivalent to the result that twice the log likelihood ratio is approximately χ^2_1. This will typically provide a better approximation than the asymptotically equivalent result that

$$\hat{\theta} - \theta \mathrel{\dot\sim} N(0, i^{-1}(\theta)) \quad (10)$$

as it partially accommodates the potential asymmetry in the log-likelihood function. These two approximations are sometimes called first order approximations because in the context where the log-likelihood is $O(n)$, we have (under regularity conditions) results such as

$$\begin{aligned}\Pr\{r(\theta; y) \le r(\theta; y^0)\} &= \Pr\{Z \le r(\theta; y^0)\} \quad (11)\\ &\quad \{1 + O(n^{-1/2})\}\end{aligned}$$

where Z follows a standard normal distribution. It is relatively simple to improve the approximation to third order, i.e. with relative error $O(n^{-3/2})$, using the so-called r^* approximation

$$r^*(\theta) = r(\theta) + \{1/r(\theta)\} \log\{q(\theta)/r(\theta)\} \sim N(0, 1) \quad (12)$$

where $q(\theta)$ is a likelihood-based statistic and a generalization of the Wald statistic $(\hat{\theta}-\theta)j^{1/2}(\hat{\theta})$; see Fraser [2].

Table I The p-values for testing $\mu = 0$, i.e. that the number of observed events is consistent with the background.

upper p-value	0.0005993
lower p-value	0.0002170
mid p-value	0.0004081
$\Phi(r^*)$	0.0003779
$\Phi(r)$	0.0004416
$\Phi\{(\hat{\theta} - \theta)\hat{j}^{1/2}\}$	0.0062427

Example 3: truncated Poisson

Suppose that y follows a Poisson distribution with mean $\theta = b + \mu$, where b is a background rate that is assumed known. In this model the p-value function can be computed exactly simply by summing the Poisson probabilities. Because the Poisson distribution is discrete, the p-value could reasonably be defined as either

$$\Pr(y \le y^0; \theta) \quad (13)$$

or

$$\Pr(y < y^0; \theta), \quad (14)$$

sometimes called the upper and lower p-values, respectively.

For the values $y^0 = 17$, $b = 6.7$, Figure 1 shows the likelihood function as a function of μ and the p-value function $p(\mu)$ computed using both the upper and lower p-values. In Figure 2 we plot the *mid* p-value, which is

$$\Pr(y < y^0) + (1/2)\Pr(y = y^0). \quad (15)$$

The approximation based on r^* is nearly identical to the mid-p-value; the difference cannot be seen on Figure 2. Table 1 compares the p-values at $\mu = 0$. This example is taken from Fraser, Reid and Wong [3].

4. PROFILE AND ADJUSTED PROFILE LIKELIHOOD FUNCTIONS

We now assume $\theta = (\psi, \lambda)$ and denote by $\hat{\lambda}_\psi$ the restricted maximum likelihood estimate obtained by

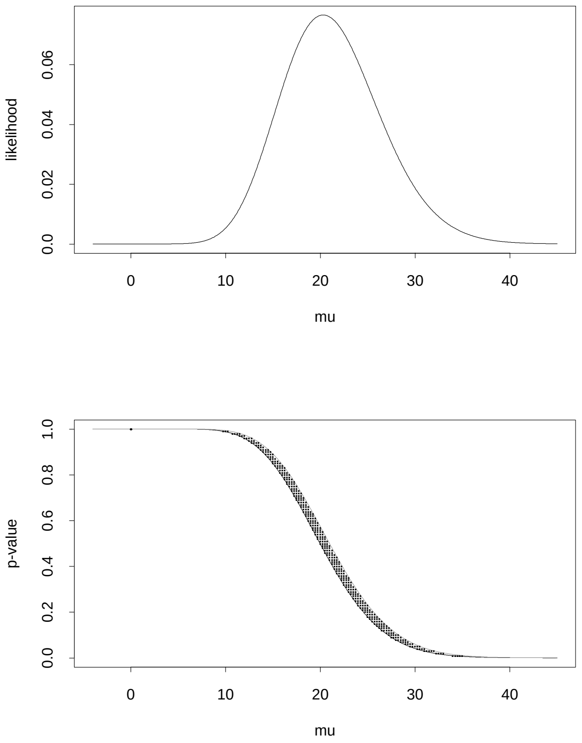


Figure 1: The likelihood function (top) and p-value function (bottom) for the Poisson model, with $b = 6.7$ and $y^0 = 17$. For $\mu = 0$ the p-value interval is $(0.99940, 0.99978)$.

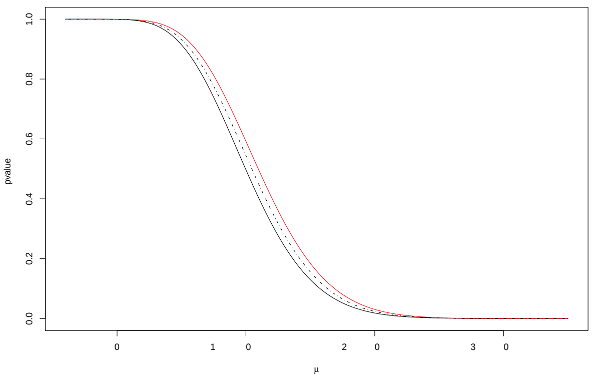


Figure 2: The upper and lower p-value functions and the mid-p-value function for the Poisson model, with $b = 6.7$ and $y^0 = 17$. The approximation based on $\Phi(r^*)$ is identical to the mid-p-value function to the drawing accuracy.

maximizing the likelihood function over the nuisance parameter λ with ψ fixed. The profile likelihood function is

$$L_p(\psi) = L(\psi, \hat\lambda_\psi); \tag{16}$$

also sometimes called the concentrated likelihood or the peak likelihood. The approximations of the previous section generalize to

$$r(\psi) = \text{sign}(\hat\psi - \psi)[2\{\ell_p(\hat\psi) - \ell_p(\psi)\}]^{1/2} \mathrel{\dot\sim} N(0,1), \tag{17}$$

and

$$\hat\psi - \psi \mathrel{\dot\sim} N(0, \{i^{\psi\psi}(\theta)\}^{-1}). \tag{18}$$

These approximations, like the ones in Section 3, are derived from asymptotic results which assume that $n \to \infty$, that we have a vector y of independent, identically distributed observations, and that the dimension of the nuisance parameter does not increase with n. Further regularity conditions are required on the model, such as are outlined in textbook treatments of the asymptotic theory of maximum likelihood. In finite samples these approximations can be misleading: profile likelihood is too concentrated, and can be maximized at the 'wrong' value.

Example 4: normal theory regression Suppose $y_i = x_i'\beta + \epsilon_i$, where $x_i = (x_{i1}, \ldots, x_{ip})$ is a vector of known covariate values, β is an unknown parameter of length p, and ϵ_i is assumed to follow a $N(0, \psi)$ distribution. The maximum likelihood estimate of ψ is

$$\hat\psi = \frac{1}{n}\Sigma(y_i - x_i'\hat\beta)^2 \tag{19}$$

which tends to be too small, as it does not allow for the fact that p unknown parameters (the components of β) have been estimated. In this example there is a simple improvement, based on the result that the likelihood function for (β, ψ) factors into

$$L_1(\beta, \psi; \bar{y})L_2\{\psi; \Sigma(y_i - x_i'\hat\beta)^2\} \tag{20}$$

where $L_2(\psi)$ is proportional to the marginal distribution of $\Sigma(y_i - x_i'\hat\beta)^2$. Figure 3 shows the profile likelihood and the marginal likelihood; it is easy to verify that the latter is maximized at

$$\hat\psi_m = \frac{1}{n-p}\Sigma(y_i - x_i'\hat\beta)^2 \tag{21}$$

which in fact is an unbiased estimate of ψ.

Example 5: product of exponential means Suppose we have independent pairs of observations y_{1i}, y_{2i}, where $y_{1i} \sim Exp(\psi\lambda_i) \quad y_{2i} \sim Exp(\psi/\lambda_i), i = 1, \ldots, n$. The limiting normal theory for profile likelihood does not apply in this context, as the dimension of the parameter is not fixed but increasing with the sample size, and it can be shown that

$$\hat\psi \to \frac{\pi}{4}\psi \tag{22}$$

as $n \to \infty$ (Cox and Reid [4]).

The theory of higher order approximations can be used to derive a general improvement to the profile

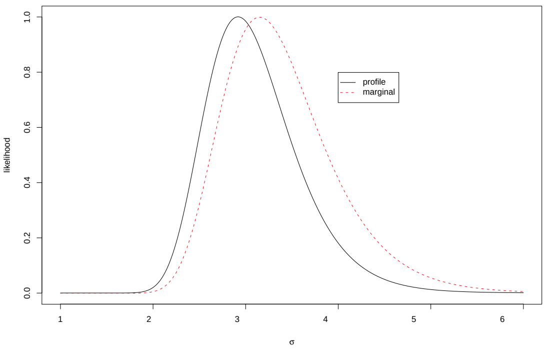


Figure 3: Profile likelihood and marginal likelihood for the variance parameter in a normal theory regression with 21 observations and three covariates (the "Stack Loss" data included in the Splus distribution). The profile likelihood is maximized at a smaller value of ψ, and is narrower; in this case both the estimate and its estimated standard error are too small.

likelihood or log-likelihood function, which takes the form

$$\ell_a(\psi) = \ell_p(\psi) + \frac{1}{2}\log|j_{\lambda\lambda}(\psi, \hat\lambda_\psi)| + B(\psi) \qquad (23)$$

where $j_{\lambda\lambda}$ is defined by the partitioning of the observed information function, and $B(\psi)$ is a further adjustment function that is $O_p(1)$. Several versions of $B(\psi)$ have been suggested in the statistical literature: we use the one defined in Fraser [5] given by

$$B(\psi) = -\frac{1}{2}\log|\varphi'_\lambda(\psi, \hat\lambda_\psi) j_{\varphi\varphi}(\hat\psi, \hat\lambda)\varphi'_\lambda(\psi, \hat\lambda_\psi)|. \quad (24)$$

This depends on a so-called canonical parametrization $\varphi = \varphi(\theta) = \ell_{;V}(\theta; y^0)$ which is discussed in Fraser, Reid and Wu [6] and Reid [7].

In the special case that ψ is orthogonal to the nuisance parameter λ a simplification of $\ell_a(\psi)$ is available as

$$\ell_{CR}(\psi) = \ell_p(\psi) - \frac{1}{2}\log|j_{\lambda\lambda}(\psi, \hat\lambda_\psi)| \qquad (25)$$

which was first introduced in Cox and Reid (1987). The change of sign on $\log|j|$ comes from the orthogonality equations. In i.i.d. sampling, $\ell_p(\psi)$ is $O_p(n)$, i.e. is the sum of n bounded random variables, whereas $\log|j|$ is $O_p(1)$. A drawback of ℓ_{CR} is that it is not invariant to one-to-one reparametrizations of λ, all of which are orthogonal to ψ. In contrast $\ell_a(\psi)$ is invariant to transformations $\theta = (\psi, \lambda)$ to $\theta' = (\psi, \eta(\psi, \lambda))$, sometimes called interest-respecting transformations.

Example 5 continued In this example ψ is orthogonal to $\lambda = (\lambda_1, \ldots, \lambda_n)$, and

$$\ell_{CR}(\psi) = -(3n/2)\log\psi - (2/\psi)\Sigma\sqrt{(y_{1i}y_{2i})}. \quad (26)$$

The value that maximizes ℓ_{CR} is 'more nearly consistent' than the maximum likelihood estimate as $\hat\psi_{CR} \longrightarrow (\pi/3)\psi$.

5. P-VALUES FROM PROFILE LIKELIHOOD

The limiting theory for profile likelihood gives first order approximations to p-values, such as

$$p(\psi) \doteq \Phi(r_p) \qquad (27)$$

and

$$p(\psi) \doteq \Phi\{(\hat\psi - \psi)j_p^{1/2}(\hat\psi)\} \qquad (28)$$

although the discussion in the previous section suggests these may not provide very accurate approximations. As in the scalar parameter case, though, a much better approximation is available using $\Phi(r^*)$ where

$$r^*(\psi) = r_p(\psi) + 1/\{r_p(\psi)\}\log\{Q(\psi)/r_p(\psi)\} \quad (29)$$

where Q can also be derived from the likelihood function and a function $\varphi(\theta, y^0)$ as

$$Q = (\hat\nu - \hat\nu_\psi)\hat\sigma_\nu^{-1/2}$$

where

$$\begin{aligned} \nu(\theta) &= e_\psi^T\varphi(\theta)\ , \\ e_\psi &= \psi_{\varphi'}(\hat\theta_\psi)/|\psi_{\varphi'}(\hat\theta_\psi)|\ , \\ \hat\sigma_\nu^2 &= |j_{(\lambda\lambda)}(\hat\theta_\psi)|/|j_{(\theta\theta)}(\hat\theta)|\ , \\ |j_{(\theta\theta)}(\hat\theta)| &= |j_{\theta\theta}(\hat\theta)||\varphi_{\theta'}(\hat\theta)|^{-2}\ , \\ |j_{(\lambda\lambda)}(\hat\theta_\psi)| &= |j_{\lambda\lambda}(\hat\theta_\psi)||\varphi_{\lambda'}(\hat\theta_\psi)|^{-2}\ . \end{aligned}$$

The derivation is described in Fraser, Reid and Wu [6] and Reid [7]. The key ingredients are the log-likelihood function $\ell(\theta)$ and a reparametrization $\varphi(\theta) = \varphi(\theta; y^0)$, which is defined by using an approximating model at the observed data point y^0; this approximation in turn is based on a conditioning argument. A closely related approach is due to Barndorff-Nielsen; see Barndorff-Nielsen and Cox [8, Ch. 7], and the two approaches are compared in [7].

Example 6: comparing two binomials Table 2 shows the employment history of men and women at the Space Telescope Science Institute, as reported in *Science* Feb 14 2003. We denote by y_1 the number of males who left and model this as a Binomial with sample size 19 and probability p_1; similarly the number of females who left, y_2, is modelled as Binomial with sample size 7 and probability p_2. We write the parameter of interest

$$\psi = \log\frac{p_1(1-p_2)}{p_2(1-p_1)}. \qquad (30)$$

The hypothesis of interest is $p_1 = p_2$, or $\psi = 0$. The p-value function for ψ is plotted in Figure 4. The p-value at $\psi = 0$ is 0.00028 using the normal approximation to r_p, and is 0.00048 using the normal approximation

Table II Employment of men and women at the Space Telescope Science Institute, 1998–2002 (from *Science* magazine, Volume 299, page 993, 14 February 2003).

	Left	Stayed	Total
Men	1	18	19
Women	5	2	7
Total	6	20	26

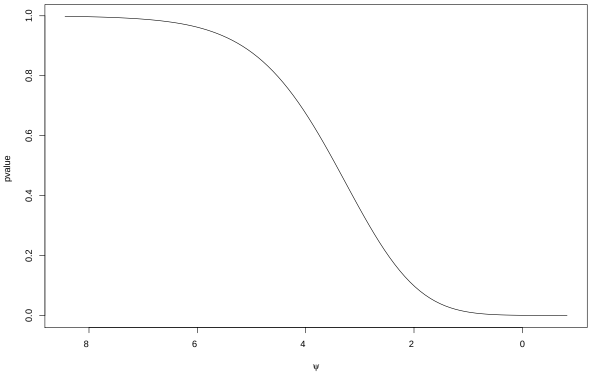


Figure 4: The p-value function for the log-odds ratio, ψ, for the data of Table II. The value $\psi = 0$ corresponds to the hypothesis that the probabilities of leaving are equal for men and women.

to r^*. Using Fisher's exact test gives a mid p-value of 0.00090, so the approximations are anticonservative in this case.

Example 7: Poisson with estimated background Suppose in the context of Example 3 that we allow for imprecision in the background, replacing b by an unknown parameter β with estimated value $\hat{\beta}$. We assume that the background estimate is obtained from a Poisson count x, which has mean $k\beta$, and the signal measurement is an independent Poisson count, y, with mean $\beta+\mu$. We have $\hat{\beta} = x/k$ and var$\hat{\beta} = \beta/k$, so the estimated precision of the background gives us a value for k. For example, if the background is estimated to be 6.7 ± 2.1 this implies a value for k of $6.7/(2.1)^2 \doteq 1.5$. Uncertainty in the standard error of the background is ignored here. We now outline the steps in the computation of the r^* approximation (29).

The log-likelihood function based on the two independent observations x and y is

$$\ell(\beta,\mu) = x\log(k\beta) - k\beta + y\log(\beta+\mu) - \beta - \mu \quad (31)$$

with canonical parameter $\varphi = (\log\beta, \log(\beta+\mu))'$.

Then

$$\varphi_{\theta'}(\theta) = \frac{\partial\varphi(\theta)}{\partial\theta'} = \begin{pmatrix} 0 & 1/\beta \\ 1/(\beta+\mu) & 1/(\beta+\mu) \end{pmatrix}, \quad (32)$$

$$\varphi_{\theta'}^{-1} = \begin{pmatrix} -\beta & \beta+\mu \\ -\beta & 0 \end{pmatrix} \quad (33)$$

from which

$$\psi_{\varphi'} = (-\beta, \beta+\mu). \quad (34)$$

Then we have

$$\chi(\hat{\theta}) = \frac{-\hat{\beta}_\mu\log(\hat{\beta}) + (\hat{\beta}_\mu+\mu)\log(\hat{\beta}+\hat{\mu})}{\sqrt{\{\hat{\beta}_\mu^2 + (\hat{\beta}_\mu+\mu)^2\}}} \quad (35)$$

$$\chi(\hat{\theta}_\psi) = \frac{-\hat{\beta}_\mu\log(\hat{\beta}_\mu) + (\hat{\beta}_\mu+\mu)\log(\hat{\beta}_\mu+\mu)}{\sqrt{\{\hat{\beta}_\mu^2 + (\hat{\beta}_\mu+\mu)^2\}}} \quad (36)$$

$$|j_{(\theta\theta)}(\hat{\theta})| = y_1y_2 = k/\hat{\beta}(\hat{\beta}+\hat{\mu}) \quad (37)$$

$$|j_{(\lambda\lambda)}(\hat{\theta}_\psi)| = \frac{y_1(\hat{\beta}_\mu+\mu)^2 + y_2\hat{\beta}_\mu^2}{(\hat{\beta}_\mu+\mu)^2 + \hat{\beta}_\mu^2} \quad (38)$$

and finally

$$Q = \left\{(\hat{\beta}_\mu+\mu)\log\left(\frac{\hat{\beta}+\hat{\mu}}{\hat{\beta}_\mu+\mu}\right) - \hat{\beta}_\mu\log\frac{\hat{\beta}}{\hat{\beta}_\mu}\right\}$$
$$\frac{\{k\hat{\beta}(\hat{\beta}+\hat{\mu})\}^{1/2}}{\{k\hat{\beta}(\hat{\beta}_\mu+\mu)^2 + (\hat{\beta}+\hat{\mu})\hat{\beta}_\mu^2\}^{1/2}}. \quad (39)$$

The likelihood root is

$$\begin{aligned} r &= \text{sign}(Q)\sqrt{}[2\{\ell(\hat{\beta},\hat{\mu}) - \ell(\hat{\beta}_\mu,\mu)\}] && (40)\\ &= \text{sign}(Q)\sqrt{}(2[k\hat{\beta}\log\{\hat{\beta}/\hat{\beta}_\mu\}) + (\hat{\beta}+\hat{\mu}) \\ &\qquad \log\{(\hat{\beta}+\hat{\mu})/(\hat{\beta}_\mu+\mu)\} \\ &\quad -k(\hat{\beta}-\hat{\beta}_\mu) - \{\hat{\beta}+\hat{\mu}-(\hat{\beta}_\mu+\mu)\}]). && (41)\end{aligned}$$

The third order approximation to the p-value function is $1-\Phi(r^*)$, where

$$r^* = r + (1/r)\log(Q/r). \quad (42)$$

Figure 5 shows the p-value function for μ using the mid-p-value function from the Poisson with no adjustment for the error in the background, and the p-value function from $1-\Phi(r^*)$. The p-value for testing $\mu = 0$ is 0.00464, allowing for the uncertainty in the background, whereas it is 0.000408 ignoring this uncertainty.

The hypothesis $Ey = \beta$ could also be tested by modelling the mean of y as $\nu\beta$, say, and testing the value $\nu = 1$. In this formulation we can eliminate the nuisance parameter exactly by using the binomial distribution of y conditioned on the total $x+y$, as described in example 1. This gives a mid-p-value of 0.00521. The computation is much easier than that outlined above, and seems quite appropriate for testing the equality of the two means. However if inference about the mean of the signal is needed, in the form of a point estimate or confidence bounds, then the formulation as a ratio seems less natural at least in the context of HEP experiments. A more complete comparison of methods for this problem is given in Linnemann [8].

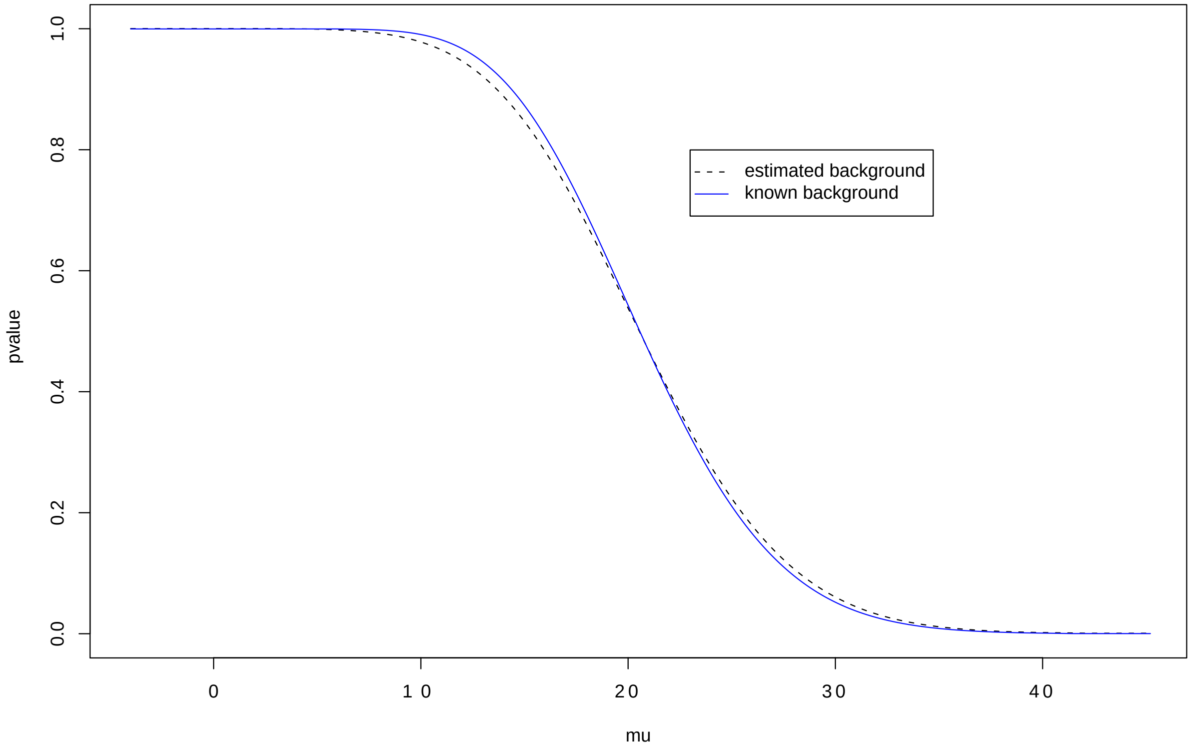


Figure 5: Comparison of the p-value functions computed assuming the background is known and using the mid-p-value with the third order approximation allowing a background error of ±1.75.

6. CONDITIONAL AND MARGINAL LIKELIHOOD

In special model classes, it is possible to eliminate nuisance parameters by either *conditioning* or *marginalizing*. The conditional or marginal likelihood then gives essentially exact inference for the parameter of interest, if this likelihood can itself be computed exactly. In Example 1 above, L_1 is the density for y_2 conditional on $y_1 + y_2$, so is a conditional likelihood for ψ. This is an example of the more general class of linear exponential families:

$$f(\underline{y};\psi,\lambda) = \exp\{\psi s(\underline{y}) + \lambda' t(\underline{y}) - c(\psi,\lambda) - d(\underline{y})\}; \quad (43)$$

in which

$$f_{cond}(s \mid t;\psi) = \exp\{\psi s - C_t(\psi) - D_t(s)\} \qquad (44)$$

defines the conditional likelihood. The comparison of two binomials in Example 6 is in this class, with ψ as defined at (30) and $\lambda = \log\{p_2/(1-p_2)\}$. The difference of two Poisson means, in Example 7, cannot be formulated this way, however, even though the Poisson distribution is an exponential family, because the parameter of interest ψ is not a component of the canonical parameter.

It can be shown that in models of the form (43) the log-likelihood $\ell_a(\psi) = \ell_p(\psi) + (1/2)\log|j_{\lambda\lambda}|$ approximates the conditional log-likelihood $\ell_{cond}(\psi) = \log f_{cond}(s \mid t;\psi)$, and that

$$p(\psi) = \Phi(r^*) \qquad (45)$$

where

$$\begin{aligned} r^* &= r_a + \frac{1}{r_a}\log(\frac{Q}{r_a}) \\ r_a &= \pm[2\{\ell_a(\hat\psi_a) - \ell_a(\psi)\}]^{1/2} \\ Q &= (\hat\psi_a - \psi)\{j_a(\hat\psi)\}^{1/2} \end{aligned}$$

approximates the p-value function with relative error $O(n^{-3/2})$ in i.i.d. sampling. An asymptotically equivalent approximation based on the profile log-likelihood is

$$p(\psi) = \Phi(r^*) \qquad (46)$$

where

$$\begin{aligned} r^* &= r_p + \frac{1}{r_p}\log(\frac{Q}{r_p}) \\ r_p &= \pm[2\{\ell_p(\hat\psi) - \ell_p(\psi)\}]^{1/2} \\ Q &= (\hat\psi - \psi)\{j_p(\hat\psi)\}^{1/2}\frac{|j_{\lambda\lambda}(\psi,\hat\lambda_\psi)|^{1/2}}{|j_{\lambda\lambda}(\hat\psi,\hat\lambda)|^{1/2}}. \end{aligned}$$

In the latter approximation an adjustment for nuisance parameters is made to Q, whereas in the former the adjustment is built into the likelihood function. Approximation (46) was used in Figure 3.

A similar discussion applies to the class of transformation models, using marginal approximations. Both classes are reviewed in Reid [9].

Acknowledgments

The authors wish to thank Anthony Davison and Augustine Wong for helpful discussion. This research was partially supported by the Natural Sciences and Engineering Research Council.

References

[1] D.R. Cox and N. Reid, "Parameter Orthogonality and Approximate Conditional Inference", *J. R. Statist. Soc.* B, **47**, 1, 1987.

[2] D.A.S. Fraser, "Statistical Inference: Likelihood to Significance", *J. Am. Statist. Assoc.* **86** 258, 1991.

[3] D.A.S. Fraser, N. Reid and A. Wong, "On Inference for Bounded Parameters", `arXiv: physics`/0303111, v1, 27 Mar 2003. to appear in *Phys. Rev.* D.

[4] D.R. Cox and N. Reid, "A Note on the Difference between Profile and Modified Profile Likelihood", *Biometrika* **79**, 408, 1992.

[5] D.A.S. Fraser, "Likelihood for Component Parameters", *Biometrika* **90**, 327, (2003).

[6] D.A.S. Fraser, N. Reid and J. Wu, "A Simple General Formula for Tail Probabilities for Frequentist and Bayesian Inference", *Biometrika* **86**, 246, 1999.

[7] N. Reid, "Asymptotics and the Theory of Inference", *Ann. Statist.*, to appear, 2004.

[8] J. T. Linnemann, "Measures of Significance in HEP and Astrophysics", available in these Proceedings on page 35.

[9] N. Reid, "Likelihood and Higher-Order Approximations to Tail Areas: a Review and Annotated Bibliography", *Canad. J. Statist.* **24**, 141, 1996.

A Unified Approach to Understanding Statistics

F. James
CERN, Geneva, Switzerland

Developing the Frequentist and Bayesian approaches in parallel offers a remarkable opportunity to compare and better understand the two.

1. TEACHING FREQUENTIST AND BAYESIAN STATISTICS IN PARALLEL

At the risk of hopelessly confusing my students, I have recently been teaching statistics with a "unified" approach, giving the Frequentist and Bayesian points of view in parallel. The goal of course is to enhance our understanding of both approaches, and I think that at least the teacher has learned something from the exercise.

The principal advantage becomes clear when we come upon a serious limitation or weakness in one approach, at which point we immediately see how the other approach handles (and often solves) this problem. The solution may come at the expense of accepting some other limitation or weakness not present in the first approach. The student then knows the trade-offs and eventually can decide which methodology he or she prefers for a given problem.

My motivation is the realization that if I want to solve all the problems generally considered to be statistical in nature, and I want to do that in the way most people do it, then both approaches are needed. If I limit myself to one approach, I can only solve a subset of problems in the appropriate way. This is justified below.

2. WHY DO WE NEED BOTH FREQUENTISM AND BAYESIANISM ?

I personally became convinced of this necessity when I realized the immense importance of two most successful statistical devices, one of which is Frequentist and the other Bayesian:

1. **Pearson's Chi-Square Test**. This test is over 100 years old and I estimate it is used $> 10^6$ times/sec on computers around the world. We know that theoretically it is one of the weaker parts of Frequentist statistics, since there can be no optimal Goodness-of-fit (Gof) test unless the alternative hypothesis is specified, but still the fundamental principle of converting a distance measure to a p-value has been so successful for so long that it cannot be dismissed out of hand. Physicists use it not only to accept and reject hypotheses, but we also use the fact that we know the error of the first kind to correct counting rates, and we compare the distribution of p-values with that (uniform) expected under the null hypothesis to obtain further information used to calibrate the apparatus and to estimate background (errors of the second kind). It is hard to imagine doing physics without the Chi-square test. Statistics without the Chi-square test is like California without the automobile: some may consider it an improvement, but it's not going to happen, so we had better learn to live with it.
2. **Bayesian Decision-Making**. If there is any statistical method used more often than Pearson's Chi-square test, it is Bayesian decision theory. It is used not only by research workers and scientists, it is used (implicitly) by everyone every day. That is how we all update our knowledge about the world around us and that is how we make the hundreds of big and small decisions we have to make to live our everyday lives. Whether we realize it or not, most of our thinking goes along Bayesian lines, so this must also occupy an important place in our statistical toolbox.

But Gof testing is not allowed in the Bayesian paradigm because it violates the Likelihood Principle. And Bayesian reasoning is not allowed in the Frequentist paradigm because it requires subjective input. Since I wish to have both in my statistical toolkit, I cannot adopt either of the fundamentalist exclusive approaches, but must somehow allow elements of both.

3. DUALITY IN PHYSICS

Duality of this kind is well known in Physics. For centuries physicists argued about whether light was particles or waves. The argument was about the "nature of light". It was assumed that light had to be of one nature or the other. Now it is known that in order to get the right answer to all problems, we need to use:

- wave formalism when the light is propagating, and

- particle formalism when it is interacting with matter.

The lesson: If your principles restrict you to only one formalism, you won't get the right answers to all the problems.

4. BASIC CONCEPTS

I organize statistics in the traditional way:

1. Basic Concepts
2. Point Estimation
3. Interval Estimation
4. Hypothesis Testing
5. Goodness-of-Fit Testing
6. Decision Theory

This structure is important for classical statistics. Bayesian methods are more unified through their common use of Bayes' Theorem, so this separation of topics is not so important for the Bayesian side, but it should still be valid. The direct confrontation between the two methodologies for each of these topics is both interesting and revealing.

4.1. Probability

Nowhere is the confrontation more interesting than in the definition of probability, which is of course at the root of all methods. I distinguish three different kinds of probability:

1. **Mathematical Probability** is an abstract concept which satisfies the axioms of Kolmogorov. There is no operational definition (no defined way to measure it).

2. **Frequentist Probability** is defined as the limiting ratio of frequencies, which restricts its application to repeatable phenomena (see footnote below). It satisfies Kolmogorov's axioms, so it is a probability in the mathematical sense.

3. **Bayesian Probability** is a little harder to define because of the many different (and often vague) definitions found in the Bayesian literature, but I call it "degree of belief" and I use the operational definition based on the *coherent bet* of de Finetti, which also satisfies Kolmogorov's axioms.

Here the trade-off is pretty clear. Methods based on Frequentist probability will be limited to repeatable[1] experiments. On the other hand, Frequentist probability is in principle independent of the observer, whereas Bayesian probability is as much a property of the observer as it is of the system being observed, so methods based on Bayesian probability are necessarily subjective (see 10.2 below).

4.2. Random Variable

This concept is very important for Frequentist statistics where one must know what is random and what is fixed. Random variables can take on different values when the identical experiment is repeated. Fixed variables always have the same value, even if the value is unknown.

In Bayesian statistics, on the other hand, experiments are not repeated and the concept of random variable is not needed. Instead, the important concept is "known" or "unknown". Probabilities can be assigned to the values of a quantity if and only if those values are unknown.

The result of the above is that the way of treating data and hypotheses is just the opposite in the two methodologies. Since the data is known, but random, Frequentists assign probabilities to data, but Bayesians do not [see below, $P(\text{data}|\text{hyp})$]. On the other hand, since the truth of a hypothesis is not random, even if it is unknown, Frequentists cannot assign probabilities to hypotheses, but Bayesians do.

Some Bayesian authors refer to unknown values as "random", which is misleading, since even a physicist who uses Bayesian methods and assigns probabilities to different ranges of possible values of an unknown parameter would not consider the true value as random.

4.3. $P(\text{data}|\text{hyp})$

This probability is meaningful and important in both approaches, but is used in different ways depending on the approach. First we note that in the case where the data are the measured values of continuous variables, this is not actually a probability, but rather a probability density function (pdf). Then in the usual case where the hypothesis is the value of a continuously variable physical parameter (such as a particle mass or lifetime), this pdf is a function of both the data and the variable hypothesis, but it is only a pdf with respect to the data.

[1] In fact, it is not strictly necessary to be able to repeat the identical experiment, but only to have an ensemble of experiments analyzed with the same procedure

In Frequentist statistics, $P(\text{data}|\text{hyp})$ is sometimes used as a probability (for discrete data) or as a pdf (for continuous data). That is when it is considered *as a function of the data,* and all possible data are considered, including the data not observed.

If the data are considered fixed (at the measured values), then this function is no longer a probability or a pdf, it becomes a *likelihood function.* It is sometimes used this way in Frequentist statistics, and *always* in Bayesian statistics. Fisher, who assigned the name *likelihood* to this function, was quick to point out that it was not a probability, since it does not behave mathematically according to the laws of probability.

Fisher's perspicacity was commendable, but unfortunately his advice has been lost in some quarters. Perhaps because of the way Bayes' Theorem (see 4.6 below) is normally written with all functions denoted by a P including the likelihood function, there is confusion in some writings between probability and likelihood.

The book of O'Hagan [1] is particularly sloppy on this point, as it confuses probability, probability density, and likelihood throughout. This results among other things in a large part of the book being devoted to the analysis of properties of the posterior pdf which are in fact arbitrary because they are not invariant under transformation of variables in the parameter.

4.4. $P(\text{hyp}|\text{data})$

This probability is meaningful only in Bayesian statistics, where it is called the *posterior probability*, calculated using Bayes' Theorem. When the hypothesis involves a continuous variable parameter (the usual case), this is in fact not a probability but a pdf.

4.5. $P(\text{hyp})$

This probability is again meaningful only in Bayesian statistics, where it is called the *prior probability*, a problematic input to Bayes' Theorem. As in the previous paragraph, this is usually a pdf, not a probability.

4.6. Bayes' Theorem

This important theorem is a part of the mathematics of probability, so it holds for any kind of probability. Thus the use of Bayes' Theorem does not necessarily make a method Bayesian. It is rather the kind of probability appearing in the use of Bayes' Theorem that determines which approach is being used.

In practice, Bayes' Theorem is not used extensively in Frequentist statistics. On the other hand, it is the cornerstone of Bayesian statistics, the theorem which expresses $P(\text{hyp}|\text{data})$ as the product of $P(\text{data}|\text{hyp})$ and $P(\text{hyp})$. Used in this way, it is of course meaningful only in the Bayesian framework.

5. POINT ESTIMATION

The Fisher criteria for point estimates are properties of the sampling distribution of the estimates: consistency, bias, variance. This leads to the selection of the maximum likelihood (ML) estimator because it has optimal properties under rather general assumptions. Surprisingly, although the Fisher criteria are not invariant under transformation of variables in the hypothesis (the bias of the square of an estimator is not equal to the square of the bias), the maximum likelihood estimate **is** invariant.

For physicists, invariance is a very important property and we feel uneasy about methods that would give different answers depending on whether we estimate mass or mass squared. In particular, I would like to know if invariant equivalents of the Fisher criteria can be found. For example, if the bias were defined in terms of the expected median of the estimates instead of the expected mean, it would be invariant. Of course the median is more computationally intensive and it is not a linear operator, which was certainly an overriding consideration in Fisher's time, but that should no longer be an obstacle. Since the ML method is invariant, everything is fine, but it would be nice to understand better how an invariant method can arise from criteria that are not invariant.

On the Bayesian side, the criterion for point estimation is usually taken as maximizing the posterior belief, which if you use a uniform prior leads to exactly the same maximum likelihood solution as in the Frequentist case. However, in the Bayesian case the invariance is a fortuitous consequence of two particular choices, both of which are hard to justify and neither of which is invariant:

- The uniform or "flat prior", which is of course only uniform in some metric, so it is not invariant, and there is not always any obvious preferred metric.

- Taking the maximum of the posterior pdf as the point estimate, which is also metric dependent. In fact, in the "natural metric", the one in which the posterior density is uniform, there is no maximum. If we follow the recommendation that the metric to be used for the posterior should be the one in which the prior is uniform, that would seem to remove the arbitrariness from the procedure, but I don't know how to justify this recommendation except that it yields the ML result.

6. INTERVAL ESTIMATION WITH EXACT COVERAGE

The most commonly used methods for interval estimation are in fact only approximate, in the sense that they do not necessarily give exact coverage. The approximation tends to be good in the limit of large data samples, which is also the limit in which Bayesian and Frequentist methods give the same numerical results. Such methods, important as they are in practice, are not of much interest in this course, so I skip directly to exact methods, valid for small samples.

6.1. Obtaining Exact Coverage

The concept of coverage is of course very different for the two approaches:

- Bayesian coverage is the observer's degree of belief that the true value of the parameter lies inside the interval quoted. It is always exact in the sense that there is no need to make any approximations, but of course belief is not easily measured with high accuracy. The degree of approximation hidden in the prior density is seldom considered.

- Frequentist coverage on the other hand, is a property of the ensemble of confidence intervals that would result from repeating the experiment many times with the same analysis. Thus when we quote a 90 % Frequentist confidence interval, that doesn't mean that the resulting interval has 90 % coverage, but rather that the method produces confidence intervals 90 % of which will contain the true value of the parameter. Sometimes it may even be seen that a particular confidence interval is one of the 10 % that is wrong, an embarrassing phenomenon for the experimentalist who wants to publish that. On the other hand, the coverage can in principle be calculated to any desired accuracy and does not depend on the observer's beliefs.

Bayesian intervals come straight out of Bayes' Theorem as soon as the prior belief is specified, and Frequentist intervals with exact coverage can always be calculated using the Neyman procedure, but a few important problems may arise:

- **discrete data.** When the data are discrete (e.g., Poisson, binomial) the Neyman construction cannot be made to give exact coverage for all possible values of the unknown parameter. Then it is necessary to overcover for some values (usually all but a set of measure zero, in fact). This can be fixed, but only at the expense of making the interval depend on an extraneous measurement or a random number. Bayesian intervals do not have this problem.

- **multidimensional data or hypotheses**. There is not much experience with the Neyman construction in more than two dimensions of data (normally the data are reduced to a statistic). The Feldman-Cousins variant [3] of the Neyman construction has been applied successfully to two-dimensional hypotheses, but we don't know much about higher dimensions. For Bayesian interval estimation, it is known that high-dimensional prior densities pose a very serious problem.

- **nuisance parameters**. Bayesians need only a prior for the nuisance parameters, then they integrate. This can be expensive computationally but is conceptually clean. For Frequentist intervals, various approximate treatments are possible, but it does not seem to be known theoretically how to obtain exact coverage for a large number of nuisance parameters. In practice it is tempting to add a pinch of Bayes and integrate over some density for nuisance parameters.

6.2. Choosing between Intervals with Exact Coverage

Coverage is not a sufficient criterion to determine confidence intervals unambiguously, so an additional criterion is needed. This is true for both Frequentist and Bayesian intervals.

- **Central intervals.** The most obvious solution is to take central intervals, with the same probability under each tail of the pdf. Frequentist intervals would be central in the data, but Bayesian intervals would be central in the hypothesis (the parameter being estimated). This means that a Frequentist central interval is not necessarily central in the parameter, and in fact may turn into an upper limit, whereas a Bayesian central interval must be two-sided, even when the data clearly indicate an upper limit only. On the other hand, Frequentist central intervals can be non-physical, whereas Bayesian central intervals must always lie in the allowed region for the parameter.

- **Most powerful intervals.** As central intervals have problems in both approaches, it is natural to look for a better criterion which will lead to "best intervals" in some sense. The sense is of course different for the two approaches, but the idea is essentially the same.

  For the Bayesian approach, the obvious criterion is to accept the interval containing the values with the highest probability density. This is in fact what is usually recommended, but we should note that this is not the same thing as choosing values of highest probability. Given a posterior pdf for a parameter, there is no way to define which values have highest probability, We know only the probability density, which is metric-dependent.

For the Frequentist approach, the situation is better. Even though there is no Uniformly Most Powerful range for the two-sided case, one can still apply the criterion suggested by optimal hypothesis testing, and choose the interval with the highest maximum likelihood ratio. It seems to be hard to find this construction described in the statistics literature, but it has appeared in the Physical Review [3].

- **Invariance of intervals under transformation of parameter**. It should be noted that the Neyman construction in general, and the Feldman-Cousins construction in particular are both invariant under transformations of variables in both the data and the parameters. Physicists are very fond of this property, since we know that any "true" theory has to obey some invariance principles including these. We are naturally uneasy about the Bayesian method of *highest posterior density* intervals. Of course it is always possible to argue that the posterior pdf contains all you could want to know about the parameter, and the interval concept is really a Frequentist invention not needed in Bayesian analysis, but still if Bayesian theory is "philosophically superior", as O'Hagan claims [2], it ought to be possible to find an interval with good mathematical properties.

- **Non-physical (empty) intervals.** This topic provides a good opportunity to contrast the basic principles of Bayesian and Frequentist inference. Some Frequentist methods can yield measurements and even confidence intervals which lie entirely in the non-physical region. This is of course embarrassing since nobody really wants to publish such a value. One of the nice properties of Feldman-Cousins intervals [3] is that they cannot be empty (non-physical). On the other hand, this means they can be *biased* in the sense of hypothesis testing, whereas central intervals are not. So there is a trade-off between *bias* and *unphysicalness*, the old problem of measurements near a physical boundary. If a set of measurements near a physical boundary is unbiased, then some of them should lie on the wrong side of the boundary in order that the average be unbiased. In order to do that, some people have to publish values they know are wrong. For Frequentists, who think in terms of getting the ensemble right, this poses no problem, but in the Bayesian approach, based on getting values you would want to bet on, it is stupid to propose a value that is wrong. In this case the duality is easy to understand: If you want to bet, use the Bayesian approach; if you want to get the ensemble right, use a Frequentist approach. Neither method is always right or always wrong. In practice people may seek a compromise like Feldman-Cousins, which has correct Frequentist coverage but avoids empty intervals at a minimum cost in bias.

7. HYPOTHESIS TESTING

Here we consider the testing of two simple hypotheses. If there is only one hypothesis, that is Goodness-of-fit testing, which comes in the following section.

The Bayesian approach produces the ratio of probabilities of the two hypotheses being correct. It can only give the ratio, because the individual factors in the ratio are not normalizable as true individual probabilities, but since the unknown normalization factor is the same, the ratio can be calculated. This ratio is exactly what is wanted in the Bayesian framework, since it is the *betting odds* which allows one to make an optimal bet for each data sample to belong to one or the other of the hypotheses.

The Frequentist method is based on determining the optimal way of classifying data in order to minimize the number of wrong classifications. Unfortunately, there are two ways of being wrong (the errors of the first and second kind), and the best one can do is, for a fixed probability of the error of one kind, to minimize the error of the other kind. Physicists refer to the error of the first kind as *loss* or *inefficiency*, and the error of the second kind is *contamination* or *background*. Traditionally one sets the level of loss that is acceptable, and the Neyman-Pearson lemma shows how to make the test with the lowest possible contamination.

It should be noted that there is no way to infer the error of the first or second kind knowing only the Bayesian betting odds, since the errors of the first and second kind require using the full sampling space and therefore violate the Likelihood Principle. Similarly, knowing only the errors of the first and second kind does not allow us to calculate betting odds, because that would require prior probabilities.

In this particular area, it seems to be especially difficult for people brought up on one approach to understand how the other one works. I know from my own experience, I had a lot of trouble to understand the meaning of the ratio of probabilities, since this combination does not normally appear in Frequentist methods. It does not allow you to calculate what physicists often want, which is acceptance and background, but of course it is just what you want for betting. Similarly, the Bayesian literature contains statements (I am thinking in particular of Howson and Urbach[5]) that show confusion between acceptance and rejection of hypotheses on the one hand and decision theory on the other.

8. GOODNESS-OF-FIT TESTING

This is theoretically the weakest part of Frequentist statistics. There can be no optimal Gof test, because there is no alternative hypothesis, so it is impossible to define the power of the test. Gof is nevertheless the most successful part of Frequentist statistics, as pointed out earlier. The absence of an optimal test has been good for making work for those who like to invent new tests. I suppose it is this plethora of empirical methods which has given rise to the accusation by Bayesians that Frequentist statistics is *ad hockery.*

But if Gof is hard for Frequentists, it is even harder for Bayesians. The Likelihood Principle makes it impossible to do traditional Gof testing in the Bayesian framework, and attempts to do it in a proper Bayesian way are relatively recent and extraordinarily complex.

When I teach this, I simply say there is no Bayesian Gof test, which is in practice probably true, but of course the experts know this is simplifying things quite a bit. Since Bayesian Goffing requires an alternative hypothesis which should encompass all possible alternatives, the result is a very big hypothesis. This “hypothesis” is in fact a giant family of hypotheses containing additional unknown parameters, which requires a multidimensional prior in addition to the prior probability that this funny hypothesis is true.

I predict that James Berger’s virtuoso attempt to make a Bayesian Gof test (the Bayesian expression is *point null hypothesis*) will not be used by any physicist more than once. The proposed method has been published in two forms: The more accessible one in *American Scientist* [6] does not actually define the method, and the technical one is in *JASA* [7].

9. DECISION THEORY

This is the domain of Bayesian methodology. Bayesian decision rules are always best in the sense that for any non-Bayesian decision rule, there is always a Bayesian rule that performs at least as well. In addition, decisions are anyway subjective because loss functions are subjective. So it is really natural to use Bayesian methods for decision-making.

A simple example makes it clear why decisions should be Bayesian: Physicists are planning the next experiment or the next accelerator. The main goal is to find a new particle that is expected on theoretical grounds, but of course all the properties of this particle (for example the mass) are not known. The detector can be optimized only for a certain range of masses, so a decision has to be made about the design of the detector. It is reasonable to optimize the detector for the properties we believe the new particle is going to have. That is a real prior belief. It may or may not be based on solid evidence or theoretical ideas. It is a mixture of knowledge and belief and intelligent guessing. It involves a real decision, namely there is money to build only one detector, and we have to decide how we are going to build it.

Once the experiment is performed, and let us assume that the new particle is found, the data also produce a measurement of the mass of the new particle. This step is no longer a decision, it is scientific inference. According to the scientific method, the measurement should be objective, not depending on the prior beliefs of the scientist doing the experiment. This is a good example to show the difference between a decision (which must depend on the subjective prior belief about the mass) and inference (which should **not** depend on prior beliefs).

10. SOME RELATED QUESTIONS

Now we are hopefully better prepared to address some more general questions.

10.1. Question: When is a calculation Bayesian?

> Example: The Birthday Problem.
> There are N people in a room. What is the probability that at least two of them have the same birthday. Assume all birthdays are equally probable and independent.

This is a standard problem in probability. Both Frequentists and Bayesians would get the same answer.

> Variant of the Birthday Problem. In practice, there are small differences in the frequencies of occurrences of actual birthdays, so a variation consists in giving a (non-uniform) probability distribution for birthdays, or even correlations between them.

Again, Frequentists and Bayesians would make the same calculations, so there is still nothing particularly Bayesian here, although people with a Bayesian upbringing will view the input probabilities as *priors*, whereas for others they are simply the given conditions of the problem.

> Final variant. You are put into a real room with 20 real people and asked to bet on

whether at least two have the same birthday.

NOW the problem becomes Bayesian, because now you must guess the (prior) probabilities and correlations between birthdays. It also becomes necessarily subjective, as real betting problems tend to be.

10.2. Question: Are Bayesian results necessarily subjective?

First of all, does it matter? In some cases, the answer is surely **yes**. Scientists want to be objective in reporting results. Bayesian fundamentalists may argue that complete objectivity is anyway impossible, so you better bring your subjectivity up front with a Bayesian analysis, still the typical physicist hesitates to introduce prior beliefs explicitly, and is especially unhappy to have to define probability as a degree of belief. In a recent course on Bayesian statistics at CERN, a young physicist in the audience commented: "But I can't publish my *beliefs*, no journal would accept that!"

As a result, many physicists' approach to Bayesian ideas is different from that of the statisticians. Many physicists think there is a "correct" prior, representing some kind of physical reality. Some even think of the prior as the distribution from which God randomly chose the values of the physical constants. Jaynes and Jeffreys tried to make Bayesian methods objective, arguing for example that two physicists with the same data and the same (lack of) prior knowledge should reach the same conclusions. I have the feeling that statisticians do not generally consider these efforts successful, but physicists tend to be less critical, probably because they would like it to be objective.

For me, the great advantage of Bayesian methods is to make the subjectivity explicit. He who tries to hide it (for example, with a flat prior), should ask whether he would not after all be happier with a method which doesn't need a prior at all.

11. WISH LIST FOR THE FUTURE

1. Scientists should learn both Frequentist and Bayesian statistics. The current situation is that most Bayesians seem to learn Frequentist statistics from other Bayesians (which is a disaster), and most people don't learn Bayesian statistics at all (which is equally bad).

2. I would like to see the Neyman interval construction and a Dinosaur plot (coverage as a function of the true value) in every new book or course on statistics. This topic is admittedly difficult for beginners, but even if the students don't understand it, they will still learn (hopefully) that there is a unique Frequentist method for exact interval construction and they may get some grasp of what coverage means.

3. Give us a book on Bayesian Statistics which is mathematically rigorous (not pretentious, just correct) and does not make incorrect or unjustified statements about Frequentism. The best book I know is de Finetti [4], but he does indulge in some polemics, and the book has to be read with a pinch[2]of salt.

4. Investigate whether the Fisher criteria for point estimation can be made invariant under transformations of the parameter being estimated.

5. We should learn how to introduce nuisance parameters into the Neyman interval construction preserving coverage.

6. In some sense the ultimate problem is what to do when we wish to get Frequentist confidence limits on a counting rate, a typical classical problem except that one of the nuisance parameters is clearly Bayesian, for example its value is calculated by a theory in which we have a certain degree of belief. It appears that we must use both Frequentist and Bayesian probabilities in the same problem. Is there a way?

12. MY VIEW OF THE BAYESIAN-FREQUENTIST DUALITY

Example: Medical Research vs. Medical Practice

1. A research team investigates the effectiveness of different drugs in treating influenza. They are analyzing frequencies: how many people get better, how many do not, etc. They must be objective, and they must get the ensemble right. They use frequentist statistics and publish P-values, errors of the first and second kind, confidence intervals with exact coverage, etc.

2. A doctor has just finished reading the report of the above research team when a patient enters his office and complains of influenza-like symptoms. The doctor now uses Bayesian decision theory to decide how to treat the patient. He should find the best treatment for **this** patient, not for the ensemble of patients he might see. This is necessarily subjective, but based on the

[2] The book entitled *Theory of Probability* begins by stating that *probability does not exist.*

objective research of the research team. It is the doctor who wants to introduce any prior beliefs he has into his decisions; he does not want priors introduced already by the research team.

References

[1] Anthony O'Hagan, Kendall's Advanced Theory of Statistics, Vol. 2B, Bayesian Inference, Arnold (1994)

[2] Anthony O'Hagan, *op.cit.*, p. 17

[3] Feldman & Cousins, PRD 57 (1998) 3873-3889

[4] B. de Finetti, Theory of Probability, J. Wiley (1974)

[5] C. Howson and P. Urbach, Bayesian Reasoning in Science, Nature, 350, no. 6317, 371-374 (1991)

[6] J. Berger and D. Berry, Statistical Analysis and the Illusion of Objectivity, American Scientist Vol. 76 p. 159 (1988)

[7] J. Berger and T. Sellke, J. Am. Statist. Soc. Vol. 81, p. 112 (1987)

Statistical Challenges of Cosmic Microwave Background Analysis

Benjamin D. Wandelt*
University of Illinois at Urbana-Champaign, IL 61801, USA

The Cosmic Microwave Background (CMB) is an abundant source of cosmological information. However, this information is encoded in non-trivial ways in a signal that is difficult to observe. The resulting challenges in extracting this information from CMB data sets have created a new frontier. In this talk I will discuss the challenges of CMB data analysis. I review what cosmological information is contained in the CMB data and the problem of extracting it. CMB analyses can be divided into two types: "canonical" parameter extraction which seeks to obtain the best possible estimates of cosmological parameters within a pre-defined theory space and "hypothesis testing" which seeks to test the assumption on which the canonical tests rest. Both of these activities are fundamentally important. In addition to mining the CMB for cosmological information cosmologists would like to strengthen the analysis with data from other cosmologically interesting observations as well as physical constraints. This gives an opportunity 1) to test the results from these separate probes for concordance and 2) if concordance is established to sharpen the constraints on theory space by combining the information from these separate sources.

1. OVERVIEW

What is cosmic microwave background (CMB) statistics and what is challenging about it?[1] It involves estimating the covariance structure of of a spatial random field with 10^6–10^8 pixels, given only ONE realization of this field. The covariance matrix of these pixels is a complicated non-linear function of the physical parameters of interest. Of these physical parameters there are between 10 and 20, so even finding the maximum likelihood point is hard—determining and summarizing confidence intervals around the maximum likelihood point is very non-trivial. Cosmologists want to do all this *and* have the option of building in exact or approximate physical constraints on relationships between parameters. In addition, since collecting cosmological data is so difficult and expensive we want to combine all available data sets—both to test them for mutual disagreement which might signal new physics, and to improve the parameter inferences. In all of this the quantification of the uncertainties in the results is extremely important—after all the stated significance of our results will either drive or stop theoretical investigations and the design of new observational campaigns.

Before I get on to CMB specifics in section 2 let me give you the short version of (most) of this talk for statisticians: "The CMB is an isotropic (homoschedastic) Gaussian random field s on the sphere. The desired set of cosmological parameters $\Theta = \{\theta_i \, i = 1, ..., n\}$, are related in a non-linear way to the spatial covariance structure $S_{ij} \equiv \langle s_i s_j \rangle$ of the field. Observers present us with a sampled, noisy, filtered and censored/polluted measurement of this field in several 'colors'. The analysis task is two-fold: infer the covariance structure of the field s. Infer the parameters Θ." This is what could be termed "canonical" CMB analysis.

In this talk I will mainly describe challenges presented by this canonical CMB analysis. After a brief review of the scientific motivation for studying the CMB in section 2 I will describe the form of CMB data as well as the current status and prospects of obtaining it in section 3. Section 4 then outlines a framework for extracting cosmologically useful information from the data and section 5 illuminates some examples of challenges that arise when implementing this framework. I will touch on statistical questions concerning "non-canonical" CMB analysis in section 6 and then conclude in section 7.

So why are we interested in facing the statistical challenges of CMB analysis?

2. WHAT CAN WE LEARN FROM THE CMB?

Cosmologists are interested in studying the origins of the physical Universe. In order to do so they have to rely on data. For cosmologists, one of the great practical advantages of Einstein's relativity over Newtonian physics is the fact that we *cannot help but look into the past.* Therefore, by observing light that reaches us from farther and farther away, we can study the Universe directly at earlier and earlier times, at least to the extent to which the Universe is transparent to light. Since the early Universe was a hot and opaque plasma we can only see back to the time when the plasma cooled sufficiently (due to the Hubble expansion) to combine into neutral atoms and the mean free time between photons collisions became of order of the present age of the Universe. Photons that we

*NCSA Faculty Fellow

[1] For online material relating to this talk please refer to `http://www-conf.slac.stanford.edu/physstat2003/talks/wandelt/invited/`

observe today which scattered for the last time in the primordial plasma *are* the CMB.

The change from plasma to gas, happened when the Universe was approximately 380,000 years old. CMB photons emitted at this time are therefore the most direct messengers that we can detect today of the conditions present in the Universe shortly after the Big Bang.[2] They constitute a pristine snapshot of the infant Universe which provide us with direct cosmological information uncluttered by the complex non-linear physics which led to the formation of stars and galaxies. One of the main attractions of the CMB is the conceptual simplicity with which it can be linked to the global properties of the Universe and the physics which shaped it at or near the Planck scale.

As a simple example, the serendipitous discovery of the CMB by Penzias and Wilson in 1965 showed that the CMB is isotropic to a very high degree. This was one of the key motivations for the development of the inflationary paradigm [4–6]. Inflation describes generically the emergence (from the era of quantum gravity) of a large homogeneous and isotropic Universe. Inflation also predicted the spectrum of small metric perturbations from which later structure developed through the gravitational instability (though it was not the only mechanism to do so). The corresponding anisotropies in the CMB were first convincingly detected by the DMR instrument on the COBE satellite in 1992. The rejection of alternative mechanisms for the generation of the primordial spectrum of metric perturbations in favor of inflation was a major advance driven by the measurement of the large angle CMB anisotropy. These developments have led to inflation becoming part of the current cosmological standard model. A very robust prediction of generic model implementations of inflation is the Gaussianity and homogeneity of the resulting perturbations.

Quantitatively, the properties of our Universe are encoded in a set of $n \sim 10-20$ *cosmological parameters* Θ where n depends on the level of detail of the modelling or, commonly, on the specification of theoretical priors which fix some of these parameters to "reasonable" values. These parameters specify the geometry and average energy density of the Universe, as well as the relative amounts of energy density contributed by the ingredients of the primordial soup (dark matter, ordinary baryons, neutrinos and dark energy and photons). In addition, the anisotropy carries information about the spectrum of primordial (inflationary) perturbations as well as their type (adiabatic or isocurvature). By combining observations of the anisotropy of both the effective temperature and the polarization of the CMB photons we can infer how transparent the Universe really was for the CMB photons on their way from last scattering to hitting our detectors. This in turn can tell us about the history of star formation.

A very exciting prospect is that by studying the details of CMB polarization we can infer the presence or absence of gravitational waves at the time of last scattering. A detection would offer an indirect view of one of the elusive messengers that started their journey at an even earlier epoch, adding a nearly independent constraint on the properties of the Universe at the Planck scale.

All this information is not encoded in actual features in the CMB map of (temperature or polarization) anisotropies. In fact in a globally isotropic universe the absolute placement of individual hot and cold spots is devoid of useful information. Information can, however, be stored in the invariants of the photon brightness fluctuations under the group of rotations SO(3). These are the properties of the field that only depend on the *relative* angular distance between two points of the field. For a Gaussian field, where 2-point statistics specify all higher order moments, this means that the angular power spectrum coefficients of the anisotropies contain all of the information.

The challenge for theoreticians was then to develop a detailed theory of the angular power spectrum C_ℓ, as a function of angular wavenumber ℓ, given the cosmological parameters Θ. While conceptually simple, it required a decade-long intellectual effort to model the relevant physical processes at the required level of precision. As a result, there now exist several Boltzmann codes (e.g. CMBFAST [2] or CAMB [3]), which numerically compute $C_\ell(\Theta)$ to 1% precision or better. The power spectra $C_\ell(\Theta)$ are sensitive functions of certain combinations of the parameters and weak functions of others (degeneracies). These weakly constrained parameter combinations are referred to as *degeneracies*. Within the context of the standard cosmological model, the theory of this dependence is well-understood.

It is clear from the preceding discussion that the CMB is an extremely valuable source of what amounts to "cosmological gold": information about the physics and the global properties of the early Universe. So what are the observational prospects?

3. DATA FROM CMB OBSERVATIONS

A major international effort is underway to make high quality observations of the microwave sky using ground-based, balloon borne and space missions [19]. Space missions have the advantage of being able to scan the whole microwave sky. NASA's "Wilkinson Microwave Anisotropy Probe" (WMAP) was launched

[2] There are other messengers, namely neutrinos and gravitational waves, reaching us from even earlier times but we do not (yet) have the technology to detect them in relevant quantities.

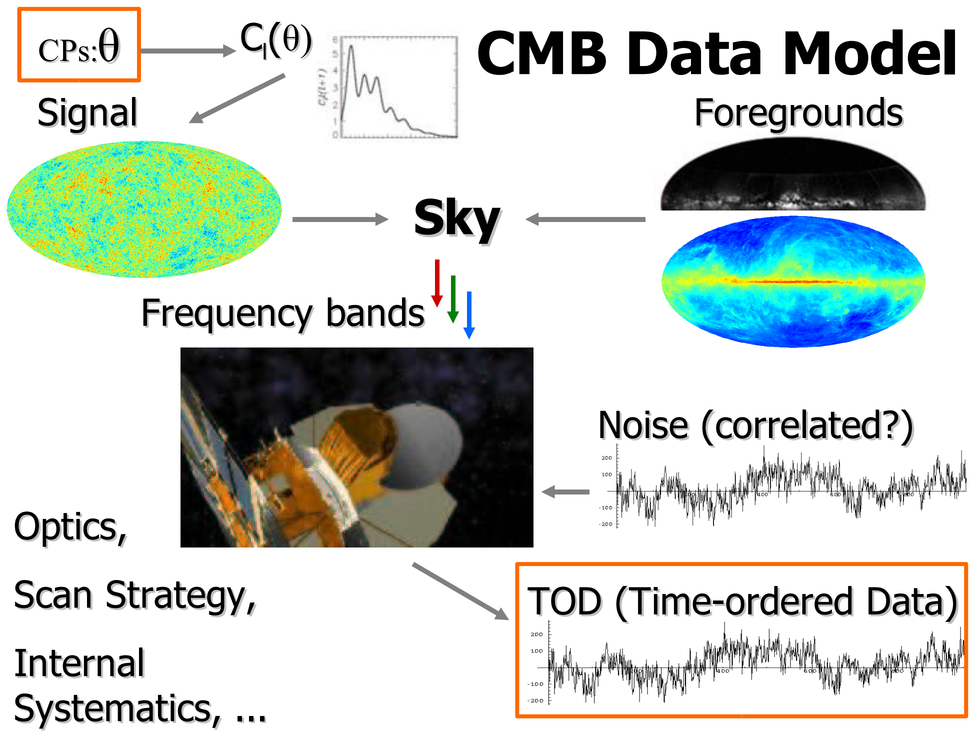


Figure 1: A schematic of how the cosmological parameters Θ (top left) are linked to the time ordered data CMB experiments actually observe (bottom right). Please see the discussion in the text.

in 2001 with great success and is currently in operation. It reported its first year of data earlier this year [16]. WMAP will continue to collect data for at least another three years. In the medium term (ie. in late 2007) we anticipate the launch of "Planck," a joint ESA/NASA space mission[3] will focus on measuring the polarization anisotropies in the CMB. In the meantime ground and balloon-based missions are jostling to accelerate our learning curve by providing maps at high angular resolution on small patches of the sky (up to a few degrees large). As a result, by the end of this decade we will have a mountain of CMB data.

From this mountain (and of course from the part of it which we have already available today) we would like to extract the cosmological gold. In order to do so we need to understand how the data and the information are related. In Figure 1 I show a simplified schematic of this relationship and I will now go through the various steps.

Starting with a set of cosmological parameters Θ we can use a Boltzmann code to compute the power spectrum C_ℓ. Given this power spectrum and the assumption of Gaussianity and isotropy we have all the information we need to create a statistical realization of a CMB anisotropy map. This is most simply done working in the Fourier representation: we draw the spherical harmonic coefficients $a_{\ell m}$ from uncorrelated Gaussians with zero mean and variance given by C_ℓ. Then we compute the spherical harmonic transform to obtain

$$T(\hat{n}) = \sum_{\ell=2}^{\ell=\infty} \sum_{m=-\ell}^{m=\ell} T_{\ell m} Y_{\ell m}(\hat{n}), \qquad (1)$$

where $Y_{\ell m}(\hat{n})$ are the spherical harmonics, an orthonormal and complete basis for functions on the sphere. The unit vector $\hat{n}$ points in a direction on the sky.

Unfortunately, the $T_{\ell m}$ or $T(\hat{n})$ cannot be observed directly. We are tied to our location in the Galaxy. There are various sources of copious amounts of microwave radiation in the Galaxy, such as dust and electrons from winds which are accelerated in the Galactic magnetic field. These foregrounds add to the CMB signal to make our sky.

This sky is then observed in various frequency bands (indicated by the red, green and blue arrows) by a CMB instrument (the figure shows an artist's concep-

[3] http://www.rssd.esa.int/index.php?project=PLANCK

tion of the WMAP satellite [33]). This instrument itself has a complicated transfer function: since weight and size constraints force satellite optics to be built compactly the optics are not free from distortion. Microwaves have macroscopic wavelengths and therefore diffract around the edges of the instrument. This leads to sidelobes in the beam maps. The instrument scans the sky in a certain pattern (the "scan strategy") and internal instrument systematics are added to the scanned signal. The microwave detectors (either radiometers or bolometers) add noise to generate the time ordered data (TOD).

There are different levels of detail of the CMB data. To give an idea for the orders of magnitude involved, let us go through the sizes of the various objects in Figure 1 for the Planck experiment.

- The complete TOD for Planck will take up of order 1 Terabyte (=10^{12} bytes) of storage (without counting house-keeping and pointing data).
- Each of the 100 detectors (channels) results in a map which is of order 10-100 Megabytes.
- The channels are grouped into of order 10 frequency bands.
- The combined maps at these different frequency bands will be combined into maps of the physical components (such as dust, synchrotron, CMB).
- The CMB power spectrum has a few thousand coefficients C_ℓ.
- These power spectrum coefficients are a function of $10-20$ cosmological parameters.

Note that there is a trade-off between the level of compression and what assumptions are implemented in the data analysis. Note also, that except for the raw data each of these data products by themselves mean little unless some means of assessing their statistical uncertainty is provided.

We immediately run into practical problems. For example, if we would like to specify a noise covariance matrix for each combined map at each frequency we would have to specify 10 times $\sim (10^6)^2/2$ elements of a matrix. The necessary storage space of order 10,000 Gigabyte basically precludes practical public distribution of the data. Other ways of specifying the uncertainty must be found.

Before we discuss in detail how cosmological information is encoded in the data, let us comment about an aspect of CMB statistics that is challenging even before there is any data in hand. This aspect is *experimental design.* While our experimental colleagues are putting a great deal of valuable thought into their instruments and observational strategies, it is currently still done in an informal way, by ingenuity rather than by formal method. This is partially so because it is hard to define optimality criteria that everyone would agree with and partially because of the immense effort involved to actually carry out the optimization. Evaluating any one proposed design requires many simulations of the full process from observation to analysis—the very process that presents the challenges I am discussing in this talk. Doing this repeatedly to search the space of design parameters for an optimal solution would be an immense task. Asymptotic techniques have been implemented on the basis of the Fisher matrix minimum variance predictions, but it is important to keep in mind that these are lower bounds on the expected variance, assuming a unimodal, Gaussian likelihood shape.

4. INFERENCE FROM THE DATA

A stochastic model of the data and the information contained in it can be summarized in the following equations. For simplicity we will limit the discussion to a single channel. The TOD, d say, is modelled as the result of the action of a linear operator A which encodes the optics and scanning strategy of the instrument, on the sky, made up of signal s and foregrounds f:

$$d = A(s+f) + n. \tag{2}$$

Our assumptions is that s is an isotropic Gaussian random field with zero mean and power spectrum C_ℓ. We will take the noise correlations $N = \langle nn^T \rangle$ and A as given—though one of the statistical challenges of CMB analysis is to relax this assumption to some degree. We will not discuss this further here.

The task is then to extract as much information as possible about the cosmological parameters Θ from the TOD. To set up this inverse problem we write down Bayes' theorem

$$P(s, C_\ell, f, \Theta|d)P(d) = P(d|f, s, C_\ell, \Theta)P(f, s, C_\ell, \Theta). \tag{3}$$

In the Bayesian context, solving the inverse problem means exploring and summarizing the posterior density $P(s, C_\ell, f, \Theta|d)$. On the right hand side we can use the conditional independence of the C_ℓ and the data given s, and the plausible independence of f and s to simplify

$$P(f, s, C_\ell, \Theta) = P(f)P(s|C_\ell)P(C_\ell|\Theta)P(\Theta). \tag{4}$$

Traditionally, inference is performed in a linear sequence of steps which are concatenated into a pipeline. At each individual step a likelihood is written for the data in its current representation (e.g. TOD, maps, C_ℓ) in terms of the parameters describing the next stage of compression. Due to the complexity of evaluating the likelihoods, often the likelihood approach is abandoned and approximate, suboptimal but unbiased estimators are constructed.

Within the above framework we can understand each of the steps as a limit of Eq. 3. For example the "map-making" step is signal estimation with

$$P(f,s,C_\ell,\Theta)=P(f)P(s|C_\ell)P(C_\ell|\Theta)P(\Theta)=const. \tag{5}$$

Similarly, "Power spectrum estimation" summarizes the marginal posterior $P(C_\ell|d)$, where the d is taken in compressed form as the estimate of the CMB component. The summary proceeds through a maximum likelihood estimate (MLE) and the evaluation of the curvature of $P(C_\ell|d)$ at the MLE. In fact in most practical cases a non-optimal estimator is used due to the computational complexity of evaluating $P(C_\ell|d)$ or its derivatives and hence of computing the MLE.

Finally, "parameter estimation" signifies the step from the MLE of C_ℓ to a density $P(\Theta|d)$. The posterior $P(\Theta|d)$ is usually summarized in terms of the marginalized means and variances of the parameters or in terms of two-dimensional projections or marginalizations through the n-dimensional parameter space.

Recently, we developed MAGIC, a method that shows promise for global inference from the joint posterior [15]. The method, based on iterative sampling from the joint posterior, exploits the full dependence structure of the different science products from a CMB mission. For example it is useful for the separation of signal and foregrounds if the covariance of the signal is known. The information obtained on C_ℓ can be fed back into the CMB map estimate, which would in turn lead to a better estimate of C_ℓ, etc. But since I discussed this technique in my contributed talk [14] I will not go into details here.

I will now discuss a selection of the challenges we face in the "map-making," "power spectrum estimation" and "parameter estimation" steps on the way from the TOD to Θ.

5. CHALLENGES IN CMB ANALYSIS

5.1. Map-making Challenges

Pixelizations of the sphere. A very basic challenge in CMB analysis is the fact that the CMB is a random field on the *sphere*. Convenient numerical techniques for storing and manipulating functions on the sphere needed to be developed. For CMB analysis in particular, fast methods for spherical harmonic transforms had to be developed and implemented. Point sets were needed that had good generalized quadrature properties such that discrete numerical approximations to Eq. 1 and its inverse

$$a_{\ell m}=\int d^2\hat{n}Y_{\ell m}(\hat{n})T(\hat{n}) \tag{6}$$

generate accurate results. For local operations such as nearest neighbor searches and multi-resolution work, pixelizations of the sphere that allow hierarchical refinement are useful.

Various pixelizations of the sphere have been proposed [23–25]. Of these, HEALPix features pixels with exactly equal areas throughout, approximate equidistribution of pixel centers over the sphere, simple analytical equations for the pixel boundaries, fast spherical harmonic transforms and very favorable quadrature properties. Due to these features, HEALPix has developed into the standard pixelization for astrophysical all sky maps.

Beam Deconvolution Map-making. It is easy to show that the MLE $\hat{m}$ for the map $m=s+f$ is the result of maximizing Eq. 3 using Eq. 5. This results in

$$m\equiv(A^TN^{-1}A)^{-1}A^TN^{-1}d \tag{7}$$

with the associated noise covariance matrix $C_N=<mm^T>=(A^TN^{-1}A)^{-1}$.

These equations are generally valid whatever the forms of A and N. However, the non-trivial structure in the observation matrix A which is a consequence of the unavoidable imperfections in satellite optics makes solving for the map challenging. It is important to deconvolve the beam functions, since the beam distortions and side lobes lead to spurious shadow images of bright foregrounds. These images are spread throughout the map in a complicated way that depends on the scanning strategy. In addition, not accounting for beam imperfections results in distorting the signal itself. Both of these effects can bias not just the map but also the result of the covariance estimation.

For polarization map-making this deconvolution becomes even more important, since the polarization signal is weak and beam asymmetries can introduce spurious polarization signals into the data. Further, if bright foregrounds are significantly polarized, they may induce a significant polarization through their shadow images if beam convolution effects are neglected in map-making. Making high-quality maps of the polarization of the CMB anisotropy is the next frontier in CMB map-making.

To simulate the effects of realistic beams we have to have a general convolution technique for a beam map with a sky map along the scan path. If implemented as discrete sums over two pixelized maps, rotated to all possible relative orientations, such a technique requires of order $n_p^{2.5}$ operations, where n_p is the number of pixels in the maps. By doing the convolution in spherical harmonic space it was shown in [21] that this can be reduced to n_p^2 in the general case and to $n_p^{3/2}$ in interesting limit cases. This fast convolution method was generalized to polarization in [22].

Component separation. Once a map of the sky is made from each channel, these maps can be compressed losslessly into maps at each frequency. A great

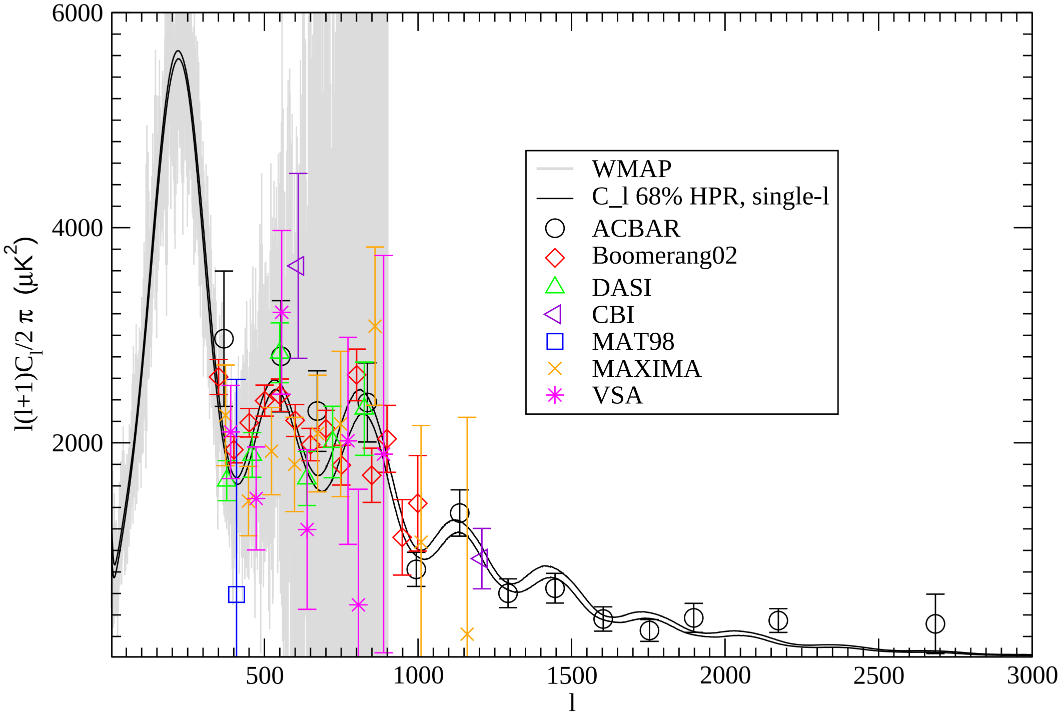


Figure 2: Our compilation of all recent CMB power spectrum (C_ℓ) data (points), including the WMAP data (gray band). Pre-WMAP data which provide redundant information about the power spectrum at low ℓ are omitted. Also shown (as two solid lines) is the 68% constraint on the power spectrum after implementing a prior which restricts the range of theories to a 10-parameter space of adiabatic inflationary theories (from [18]).

deal of work in the field has gone into devising methods for then obtaining an estimate of the CMB sky from these foreground contaminated maps at each frequency band, both for temperature and for polarization. We can either choose to model the foregrounds physically (e.g. [30, 31]) or we can attempt "blind separation," by defining an algorithm for automatic detection of different components in the maps, e.g. based on the statistical independence of these components (e.g. [29]).

Lensing. The CMB has traveled past cosmological mass concentrations which perturb the photon geodesics. This distortion contains valuable and complementary information about cosmological parameters. At the same time it mixes the polarization modes in the CMB data, contaminating the primary signal for the detection of the primordial gravitational wave background. Methods need to be devised that can measure this distortion and extract the information contained in it. It is intriguing that [32] find that exact techniques have significant advantages over approximate, quadratic estimators for the reconstruction of B polarization maps from lensed CMB.

5.2. Power Spectrum Estimation

The Computational Problem. For perfect (all-sky, pure signal, no noise) data, power spectrum estimation would be easy. Just compute the spherical harmonic transform, Eq. 6. Then the estimator

$$\hat{C}_\ell = \frac{\sum_m |a_\ell|^2}{2\ell + 1} \tag{8}$$

is the MLE for C_ℓ. It is also easy to evaluate and explore the perfect data posterior to quantify the uncertainty in the estimates. However, in the general case we would like to evaluate $P(S(C_\ell)|d, N)$ which is the result of integrating out ("marginalizing over") s in the joint posterior. We obtain[4]

$$P(S(C_\ell)|d, N) = G(m, S(C_\ell) + C_N). \tag{10}$$

Here S is the signal covariance matrix, parameterized by the C_ℓ. Since $S + C_N$ is not a sparse matrix and since the determinant in the Gaussian depends on S, evaluating $P(S(C_\ell)|d, N)$ as function of the C_ℓ costs of order n_p^3 operations.

For Planck, $n_p \sim 10^7$ so if the constant factor in the scaling law was 1 (an unrealistic underestimate)

[4] We use $G(x, X)$ as a shorthand for the multivariate Gaussian density

$$G(x, X) = \frac{1}{\sqrt{|2\pi X|}} \exp\left(-\frac{1}{2} x^T X^{-1} x\right). \tag{9}$$

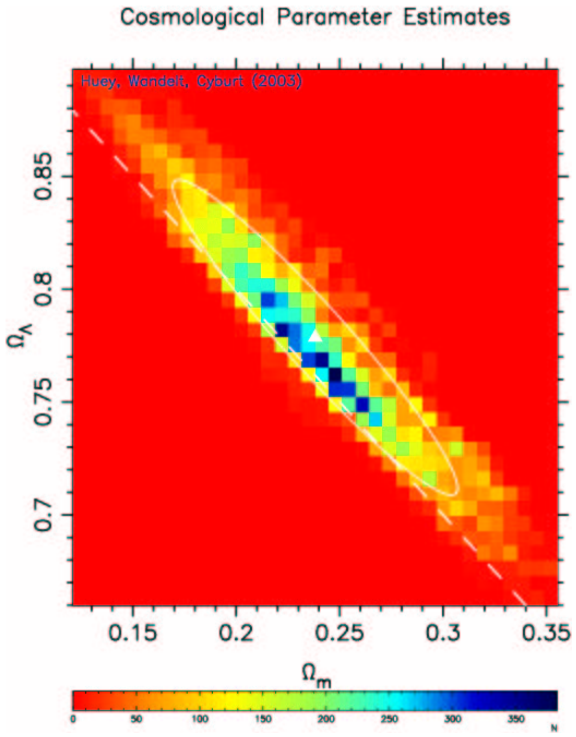


Figure 3: An example result of an online exploration of the marginal posterior of the dark matter density (Ω_m) and dark energy density (Ω_Λ) in our Universe. All recent CMB data, including the WMAP data, as well as the Hubble Space Telescope key project results and the Supernova cosmology project were included to obtain these constraints. Points in the figure are colored (from red to blue) according to how well the parameter combination at that point agrees with the data.

this would mean 10^{21} operations for one likelihood evaluation. For a 10 GFLOP CPU this means 1000s of CPU years of computation.

Various approaches have been suggested to counter this challenge. They can be broadly divided into two classes: 1) specialized exact (maximum likelihood estimation) algorithms exploit advantageous symmetries in the observational strategy of a CMB mission [9, 10] to reduce the computational scaling from order n_p^3 to order n_p^2, and 2) approximate algorithms which filter unwanted properties of the data and simply compute the power spectra on the filtered and incomplete data, and then de-bias the results using Monte Carlo simulations after the fact[12]. This second class of algorithms is known as pseudo-C_ℓ algorithms[11] and has gained a great deal of popularity in recent analyses of CMB data, including the WMAP data [13].

Cosmic Variance. Aside from computational problems there is an interesting conceptual problem with CMB power spectrum analysis: the fact that we only have one sky.

This fact induces a fundamental limitation to how well we will be able to constrain cosmological parameter estimates from the CMB. The usual way of phrasing this limitation is in terms of cosmic variance.

Essentially, power spectrum estimation is variance estimation. For perfect data the solution that maximizes Eq. 10 is

$$\hat{C}_\ell = \frac{\sum_m |a_\ell|^2}{2\ell+1}. \tag{11}$$

So we are estimating the variance of the spherical harmonic coefficients. But, for example, for $\ell = 2$ there are only 5 such coefficients and a variance estimate from 5 numbers is statistically uncertain. So "cosmic variance" is expression of the fundamental limit to the precision of any measurement of the C_ℓ caused by the fact that the sphere is a bounded space and the fact that causality will not allow us to observe independent patches of the universe.

This fundamental limit to our knowledge provides a powerful motivation to do the best possible job in analyzing cosmological data.

5.3. Parameter Estimation

Techniques. The problem of parameter estimation is challenging since we need to explore the posterior density $P(\Theta|d)$ which varies over $\sim 10-20$ dimensions. Various techniques have been used to do this. For smaller number of dimensions (up to about 5) gridding techniques have worked well. However, the current state of the art is to use Markov Chain Monte Carlo methods [20] such as the Metropolis Hastings algorithm to sample from $P(\Theta|d)$ and to then base inferences on summaries of the posterior density computed from the sampled representation.

The question how to implement physical priors and constraints was one theme that was discussed at this conference. Within the Bayesian framework there exists a unique prescription for applying physical constraints through the specification of informative priors. The first example of a CMB parameter estimation which can be explored interactively online is the Cosmic Concordance Project [17]. This compiles data from several recent CMB observations and combines them with the user's choice of other, non-CMB experiments and physical priors. The result is displayed as the 2D marginal posterior density for any 2 parameters chosen by the user. An example is displayed in figure 3. I invite you to have a look at our prototype implementation at http://galadriel.astro.uiuc.edu/ccp.

A first scientific result from this project was a measurement of the fraction of ^{4}He in the Universe, both from CMB data alone and in the combination of standard Big Bang Nucleosynthesis with the CMB data.

Since standard BBN links the primordial ^{4}He abundance to the baryon to photon ratio which is determined exquisitely well by CMB data, we obtain the most precise measurement of the primordial ^{4}He abundance to date [18].

Several open questions remain to be addressed. What number n of parameters is needed to fit the data? Which selection of parameters from the full set should we use in the analysis? These questions are currently handled on the basis of personal preference of the authors or computational convenience—a statistically motivated procedure for letting the data selecting certain parameters has not yet been implemented.

Operationally, it is very computationally expensive to sample from a posterior with 10 parameters or more, because each likelihood evaluation requires running a Boltzmann code which computes $C(\Theta)_\ell$. In many dimensions the Metropolis sampler produces correlated samples (regardless of whether the target density is correlated or not—the correlations come from the sampler taking small steps through the many dimensional space). New numerical techniques for evaluating and sampling from the likelihood in high-dimensional spaces are needed.

Beating cosmic variance. In Figure 2 we show another application of parameter estimation techniques which is relevant to several themes we touched on in this talk and at this conference. The figure shows our compilation of CMB data and the mean and ±68% errors on the range of power spectra which are contained in our 10 parameter fits. In other words, we have implemented the optimal non-linear filter for the C_ℓ if the Universe is really described by our 10 parameter model. The physical prior has reduced the cosmic variance error bars far below the limit set by Eq. 11. Since we are using the physical prior that the CMB power spectrum is the result of plasma oscillations in the primordial photon baryon fluid the resulting smoothness of the power spectrum is used in constructing the estimate.

6. TESTING THE ASSUMPTIONS: CHALLENGES OF NON-CANONICAL CMB ANALYSIS

In addition to the canonical analyses outlined above it is fundamentally important to test the assumptions on which these analyses rest. I will very briefly mention two of them here.

Is the CMB signal really Gaussian? Is it isotropic? These questions touch simultaneously on the issue of hypothesis testing, as well as model selection. What is the evidence in the data for the assumptions of isotropy and Gaussianity on which canonical analysis of the CMB are built? Can a goodness of fit criterion be defined which allows assessing whether the standard model of cosmology is a complete description of the CMB data?

The idea of testing the goodness of fit can be extended to cross tests of CMB data with other cosmological data. Ultimately the goal of the Cosmic Concordance Project is to allow users to select various data sets and explore interactively whether the parameter constraints from various data sets are compatible with each other. If this agreement is established the the constraints can be combined to generate yet stronger constraints. If disagreement is found this is motivation for observational groups to collect more data or for theoretical groups to work out new mechanisms that can reconcile the discrepant observations.

Testing for statistical isotropy in the CMB is a well-defined operation, since it is easy to specify alternative models. It requires checking whether a model of the correlations in the fluctuations in terms of rotationally invariant quantities (such as the power spectrum) is better or worse than a model that contains quantities that do not transform as scalars under SO(3). A frequentist test statistic has been suggested in [34], but a Bayesian treatment has not yet been attempted. The detection of a significant deviation from statistical isotropy would be a very important result, since isotropy is a fundamental prediction of inflation.

Testing for non-Gaussianity (NG) is similarly important, but much less well-defined. In the absence of physical models for NG a Bayesian treatment is not possible. The standard approach is to define some NG statistic, e.g. the skewness of the one-point function of the CMB fluctuations. Then this statistic is applied to the data and to a sample of Gaussian Monte Carlo samples (which are usually constrained to match the two-point statistics of the observed data). This Monte Carlo sample represents the null-hypothesis. A way is defined to assess discrepancy of the data and the Monte Carlo sample and if the discrepancy is statistically significant a detection is claimed.

Even though this is a straightforward frequentist procedure there is a great deal of arbitrariness in choosing the test statistic. Usually the choice is made based on a vague notion of genericity, in the sense that, for example, the skewness of the data is probably a more generic NG statistic than the temperature in pixel number 2,437,549, say. However, it is clearly not trivial to define a proper measure on the space of all possible statistics. The arbitrariness in the selection of the statistic (topological features of the map, n-point functions in pixel and in spherical harmonic space, wavelets, excursion sets, etc.) makes the test results difficult to interpret. For example, is clearly true that there will always be *some* statistic that gives an n-sigma result for any n!

To circumvent this arbitrariness, we could test for the NG predicted in certain variants of inflationary models ([26–28] and references therein). These physical models allow a Bayesian analysis in principle, and it would provide the best possible constraints on the presence of NG, but implementing Bayesian NG inference these models is computationally tedious.

7. CONCLUSION

In this talk I reviewed some of the more severe statistical and computational challenges of CMB analysis. CMB data are fundamentally important to cosmology, the problems I outlined are intellectually rich, and the knowledge that we need to make the most of the one CMB sky we can observe and analyze, present a powerful motivation to solve the CMB analysis problem.

Acknowledgments

I thank the organizers of PHYSTAT2003 for making this conference so enjoyable and stimulating. This work has been partially supported by the National Computational Science Alliance under grant number AST020003N and by the University of Illinois at Urbana-Champaign through an NCSA Faculty Fellowship.

References

[1] Mason, B.S., *et al.*, Ap. J. 591, 540 (2003)
[2] Seljak, U., and Zaldarriaga, M., Ap. J. 469, 437 (1996)
[3] Lewis, A., Ap. J. 538, 473 (2000)
[4] Guth, A., Phys. Rev. D. 23, 347 (1981)
[5] Linde, A., Phys. Lett. 108B, 389 (1982)
[6] Phys. Rev. Lett 48, 1220 (1982)
[7] C. L. Bennett *et al.*, Ap. J. 464, L1 (1996)
[8] J. R. Bond, A. H. Jaffe, and L. Knox, Physical Review D 57, 2117 (1998)
[9] S. P. Oh, D. N. Spergel, G. Hinshaw, Ap. J. 510, 551 (1999)
[10] Wandelt, B. D. and Hansen, F., Phys. Rev. D67, 023001 (2003)
[11] B. D. Wandelt, E. Hivon, and K. M. Górski, Phys. Rev. D 64, 083003 (2001)
[12] E. Hivon et al., Ap. J. 567, 2 (2002)
[13] G. Hinshaw, et. al., ApJS 148, 135 (2003)
[14] Wandelt, B. D., "MAGIC: Exact Bayesian Covariance Estimation and Signal Reconstruction for Gaussian Random Fields", available in these Proceedings on page 229.
[15] Wandelt, B.D., Larson, D., and Lakshminarayanan, A., astro-ph/0310080, PRD in press.
[16] C. L. Bennett *et al.*, Ap.J.Suppl. 148, 1 (2003)
[17] Huey, G., Wandelt, B. D., and Cyburt, R., in preparation.
[18] Huey, G., Cyburt, R., and Wandelt, B. D., astro-ph/0307080, PRD in press.
[19] A. D. Milleret al. 1999 Astrophys.J. 524 (1999) L1-L4; N. W. Halverson et al. 2001, astro-ph/0104489; S. Hanany et al. Astrophys.J. 545 (2000) L5; Keith Grainge et al. Mon. Not. R. Astron. Soc. 000, 15 (2002); Kuo, C. L. et al. 2002, Ap. J., astro-ph/0212289; J. E. Ruhl et al. 2002, astro-ph/0212229; S., Padin et. al., ApJ 549, L1, (2001); for a comprehensive list see `http://lambda.gsfc.nasa.gov`
[20] Christensen, N., *et al.*, Class. Quantum Grav. 18, 2677 (2001)
[21] Wandelt, B.D. and Gorski, K., Physical Review D63, 123002 (2001).
[22] Challinor, A. *et al* Physical Review D62, 123002 (2000).
[23] Górski, K.M., Hivon, E., and Wandelt, B.D., in "Evolution of Large Scale Structure from Recombination to Garching," Banday, A. *et al* (eds.), Garching, Germany (2000). See also the HEALPix homepage (`http://www.eso.org/science/healpix`).
[24] Crittenden, R.G. and Turok, N.G., astro-ph/9806374 (1998)
[25] Doroshkevich, A., *et al.* astro-ph/0305537 (2003)
[26] E. Komatsu, et.al., ApJS 148, 119 (2003)
[27] E. Komatsu and D. Spergel, Physical Review D63, 063002 (2001)
[28] E. Komatsu, D. Spergel, and B. D. Wandelt, astro-ph/0305189 submitted to Ap. J. (2003)
[29] D. Maino, et al., MNRAS 334, 53 (2002)
[30] Stolyarov, V, et al., MNRAS 336, 97 (2002)
[31] Bennett, C. L., et al., ApJS 148, 97 (2003)
[32] Hirata, C. M. and Seljak, U., PRD 68, 083002 (2003)
[33] `http://map.gsfc.nasa.gov/`
[34] A. Hajian and T. Souradeep, Ap. J. 597, L5 (2003)

Uncertainties of Parton Distribution Functions

Daniel R. Stump
Department of Physics and Astronomy, Michigan State University, East Lansing, MI, 48824

Issues related to the analysis of uncertainties of parton distribution functions are discussed, focusing on the methods used in the CTEQ global analysis.

1. INTRODUCTION

High-energy particles interact through their quark and gluon constituents—the partons. By the asymptotic freedom of QCD, the parton cross sections may be approximated by perturbation theory. Then by the factorization theorem of QCD, the parton distribution functions of hadrons are the link between perturbative QCD and experimental measurements. This theory applies to a variety of high-energy processes, including deep-inelastic lepton scattering from nucleons (DIS), Drell-Yan production of $\mu\overline{\mu}$ pairs in nucleon-nucleon collisions (DY), and production of high-p_T jets at $p\overline{p}$ or pp colliders.

The parton distribution functions (PDFs) are important in high-energy physics. Any scattering experiment with nucleons in the initial state will require PDFs for the analysis and interpretation of the experiment. Currently, the HERA accelerator experiments measure ep and $\overline{e}p$ scattering; the Tevatron collider creates $p\overline{p}$ collisions. Tests of the standard model and the search for new physics rely on PDF phenomenology.

QCD global analysis has several goals: to construct an accurate set of PDFs; to know the uncertainties of the PDFs, which come from experimental measurement errors and from theoretical approximations; and to enable predictions, with realistic uncertainties, for future experiments.

The full, systematic study of PDF uncertainties developed slowly [1, 2]. The first practical parton distributions with full error bands were produced by Botje [3] using DIS data. Today many groups and individuals are involved in this active field of research. Both the CTEQ group [4] and the MRST group [5] have provided complete uncertainty analyses of their PDFs; the PDFs from these groups are widely used in high-energy physics. The HERA collaborations, ZEUS [6] and H1 [7], have constructed PDF models based on their own data, including error bands. Other individuals have made important contributions [8, 9]; and the Fermilab group [10] has developed a new methodology for PDFs by Monte Carlo sampling in the parton function space.

This paper will discuss several issues, focusing on the CTEQ methods and results. Section 2 describes the CTEQ input. Section 3 concerns the treatment of systematic errors. Section 4 discusses the compatibility of different data sets within a global fit to data. Section 5 explains the CTEQ uncertainty analysis.

2. GLOBAL ANALYSIS OF QCD

The aim of global analysis of short-distance processes using perturbative QCD is to construct a set of PDFs that yield good agreement with data from many disparate experiments. The program of global analysis is not a routine statistical calculation, because of systematic errors—both experimental and theoretical. Therefore we must sometimes use physics judgement in producing the PDF model, as an aid to the objective fitting procedures.

A parton distribution function $f_i(x, Q)$ (where i labels the parton species) depends on two variables: the momentum fraction x carried by the parton and the momentum scale Q at which the nucleon is observed. Heuristically, $f_i(x, Q)$ is the density of parton species i per unit of momentum fraction. The PDFs are parametrized at a low momentum scale Q_0, of order 1 GeV, by a standard functional form with adjustable parameters $(a_0, a_1, a_2, \ldots)$

$$f(x, Q_0) = a_0 x^{a_1} (1-x)^{a_2} P(x); \qquad (1)$$

here $P(x)$ is a smooth function with a few additional free parameters. Separate functions exist for each parton species, so that the total number d of adjustable parameters is of order 20. The Q dependence of $f(x, Q)$ is determined by the QCD evolution equations [11], depending on the renormalization and factorization schemes for the perturbative calculation of parton cross sections. The most widely used CTEQ parton distributions are based on next-leading order (NLO) perturbation theory in the modified minimal subtraction renormalization scheme.

Table 1 lists the experiments used in the CTEQ6 global analysis—the most recent generation of CTEQ parton distributions [12]. Thirteen data sets are employed, and the number N of data points in a set ranges from scores to hundreds. The total number of data points is ~ 1800. These data come from major experiments of the past ten years. The CTEQ6 PDFs agree satisfactorily with all the experiments, from the fact that each χ^2/N is near 1 and from more detailed comparisons.

Table I Experimental data sets used in the CTEQ6M global analysis [12]. CorrMat: availability of information on correlations of systematic errors. N is the number of data points. References for the experiments may be found in Ref. [12].

CTEQ6						
	process	data set	CorrMat	N	χ^2	χ^2/N
1	μ DIS	BCDMS F2p	Y	339	378	1.11
2	μ DIS	BCDMS F2d	Y	251	280	1.11
3	$\bar{e}$ DIS	H1 (a)	Y	104	98.6	0.95
4	e DIS	H1 (b)	Y	126	129	1.02
5	$\bar{e}$ DIS	ZEUS	Y	229	263	1.15
6	μ DIS	NMC F2p	Y	201	305	1.52
7	μ DIS	NMC d/p	Y	123	112	0.91
8	$p\bar{p} \to$ jet	D0	Y	90	69	0.77
9	$p\bar{p} \to$ jet	CDF	Y	33	49	1.47
10	$\nu(\bar{\nu})$ DIS	CCFR F2 + F3	Y/N	156	150	0.96
11	Drell-Yan	E605	N	119	95	0.80
12	Drell-Yan	E866 d/p	N	15	6	0.40
13	$p\bar{p} \to W$	CDF (Lasy)	N	11	10	0.91

3. TREATMENT OF EXPERIMENTAL SYSTEMATIC ERRORS

What is a systematic error? Suppose two experimental groups measure the same quantity, but one has a positive systematic error and the other negative. The measurements will not agree within the statistical errors, and no theory could agree with both experiments. If the groups are aware of the possible errors then the measurements will in fact be consistent within the published uncertainties including systematics. But if the errors are not accurately characterized, an incompatibility will appear, recognized as a systematic difference of the results.

The situation described in the previous paragraph is analogous to what often happens in global analysis of PDFs. But in the case of PDFs, the systematic differences are only visible through the process of global analysis. The disparate experiments may measure different energy domains, or altogether different scattering processes; the results cannot be compared directly. Only through the combined fitting of PDFs are the systematic differences revealed.

For a global analysis, a crucial feature of the systematic errors is that they are highly correlated. Therefore we construct the PDFs using a procedure of χ^2 minimization with fitting of systematic errors. First consider an experiment with statistical errors alone; we would define

$$\chi^2 = \sum_{i=1}^{N} \frac{(D_i - T_i)^2}{\sigma_i^2} \begin{cases} D_i \,:\, \text{data value}(i = 1 \ldots N) \\ T_i \,:\, \text{theoretical value} \\ \sigma_i \,:\, \text{statistical error (S.D. of } D_i) \end{cases} \tag{2}$$

The theory value T_i is a function of d adjustable parameters $\{a_1, \ldots, a_d\} \equiv \mathbf{a}$. Then minimization of χ^2 with respect to the $\{a_\mu\}$ would give the optimal PDF model $\{a_{0\mu}\}$. This procedure would be ideal if the data value D_i is $T_{0i} + \sigma_i r_i$ where r_i is a random Gaussian fluctuation with $\langle r_i^2 \rangle = 1$.

A. Normalization error.

In any scattering experiment there is an overall normalization uncertainty from the luminosity. To take this possible error into account, we introduce a normalization factor $f_{\rm N}$ and define a new χ^2 function,

$$\chi^2(\mathbf{a}, f_{\rm N}) = \left(\frac{1 - f_{\rm N}}{\sigma_{\rm N}}\right)^2 + \sum_{i=1}^{N} \frac{(f_{\rm N} D_i - T_i)^2}{\sigma_i^2}, \tag{3}$$

where $\sigma_{\rm N}$ is the published normalization uncertainty. Minimization of χ^2 with respect to both the model parameters $\{a_\mu\}$ and the normalization factor $f_{\rm N}$ yields an optimized model, within the normalization uncertainty. Rather than fitting T_i to D_i, we fit it to $f_{\rm N} D_i$. The choice of $f_{\rm N}$ is also optimized by the minimization of χ^2. The penalty term in (3) ensures that $f_{\rm N}$ will not deviate too much from 1.

An overall normalization error of theoretical origin, such as a K-factor from high-order QCD corrections, would be indistinguishable from a luminosity error. Thus the correction for systematic errors may involve both experimental and theoretical errors, and these would be entwined.

B. General systematic errors.

High-precision experiments publish *many* systematic errors. In general, the difference between data and theory will be

$$D_i = T_{0i} + \alpha_i r_i + \sum_{j=1}^{K} \beta_{ij} \widehat{r}_j, \tag{4}$$

where α_i is the uncorrelated error of D_i and $\{\beta_{ij}, j = 1 \ldots K\}$ is a set of K systematic (and 100% correlated) errors. The $\{r_i\}$ and $\{\widehat{r}_j\}$ denote random fluctuations. To account for the systematic errors, we introduce a systematic shift s_j for each source of error and define a new χ^2 function,

$$\chi'^2(\mathbf{a}, \mathbf{s}) = \sum_{i=1}^{N} \frac{\left(D_i - \sum_j \beta_{ij} s_j - T_i\right)^2}{\alpha_i^2} + \sum_{j=1}^{K} s_j^2. \tag{5}$$

Minimization of χ'^2 with respect to both the functional parameters $\{a_\mu\}$ and the systematic shifts $\{s_j\}$ determines the optimal model. The theory is not fit to the central values of the published data, but to values that are shifted by amounts consistent with the published systematic errors $(\beta_{ij}s_j)$. The fitting procedure produces a compatible model within the systematic errors.

Because χ'^2 depends quadratically on the $\{s_j\}$ we can solve for the optimized shifts analytically, $\mathbf{s} \to \mathbf{s}_0(\mathbf{a})$. Thus the systematic shifts are continually optimized as we vary the functional parameters $\{a_\mu\}$ in seeking the optimal PDFs.

The procedure outlined above accounts for the statistical errors (by weighting with α_i^{-2}), the overall normalization uncertainty (by numerical fitting of $f_{\rm N}$), and the other systematic errors (analytically).

Finally, the *global* χ^2 *function* is the sum of the $\chi_e'^2$ over all the experiments e. We might also apply weighting factors $\{w_e\}$ when combining the experiments, with default values $w_e = 1$. The spirit of global analysis is compromise—the PDF model should fit all data sets satisfactorily. If a model has poor agreement with some data, we may construct another model with better agreement by giving that data an enhanced weight in the global χ^2. The second fit may be judged to be preferable as a standard fit to the global data, even though it is not the best fit to other experiments. In making a subjective decision like this, physics judgement enters into the calculation.

The quality of the final CTEQ6 PDFs can to some extent be gauged by the χ^2 values in Table 1. But just looking at a single number for a data set does not do justice to the theory. More detailed comparisons, e.g., by plotting data and theory superimposed or by the "pull" distributions, reveal the beautiful success of QCD. [12]

4. A STUDY OF COMPATIBILITY

Because we seek to construct PDFs that agree with many disparate experiments, so that the functions contain the best available information on all aspects of the parton structure simultaneously, an important issue is whether the data from different experiments are in agreement with each other. If two data sets do not agree with one another, then no theoretical model can agree with both. The philosophy of the CTEQ program is to use high-precision data from all relevant experiments. We expect to observe minor incompatibilities between experiments because of systematic errors. Then practical PDF models will require compromises among the different experiments. However, we should not encounter true inconsistencies, i.e., which cannot be reconciled by a reasonable model.

Several methods have been used to judge the compatibility of different data sets. One method is to study alternate fits that result from changing the weights of the data sets in the global χ^2 function. One example is shown in Table 2. The data sets and PDFs in this study are not identical to CTEQ6 but are quite similar. The PDF model A is the standard set—for which all data sets have the default weight $w_e = 1$ in the fitting. The PDF model B is an alternate in which the three H1 data sets and two BCDMS data sets are given a large extra weight.[15] The final column $\Delta\chi^2$ gives the difference in χ^2 for each experiment, between models B and A. By heavily weighting H1 and BCDMS, their χ^2's have decreased significantly, by -38.8 units (out of ~ 970). But on the other hand the χ^2's of other experiments have increased significantly, by $+149.7$ (out of ~ 1390). For example, while the agreement with H1 data improved, the quality of the fit to ZEUS data got worse. The result is that the model in best agreement with H1 and BCDMS data will not agree satisfactorily with other experiments. The change in χ^2 can be rather large. There is a minor incompatibility.

Giving extra weight to one or a few experiments is equivalent to using that data alone to construct the PDFs. An experimental collaboration that uses only its own data can expect to model that data better than a general set of PDFs from a global analysis. However, their resulting PDFs will not describe the many other data as well.

The inverse of giving extra weight to an experiment is to give it less weight, of which the extreme case is to give it weight 0, i.e., simply to remove it from the global analysis. When such exercises are performed, we find for most data sets that the χ^2's of other experiments may decrease significantly, by amounts of order 10 units in some cases, while the χ^2 for the dropped data increases by a similar amount. Again, the implication is that the best fit to the remaining data may not give a satisfactory fit to the dropped data.

These minor incompatibilities suggest that there are systematic errors, whether of experimental or theoretical origin, that prevent a purely statistical analysis of the uncertainties. Furthermore, the results show that reasonable PDFs may differ in global χ^2 by amounts much larger than 1.

Other clever ways to test the compatibility of different data sets have been devised. One method is to plot χ^2 of the global data versus χ^2 of an individual data set, for alternate PDF models generated by the Lagrange Multiplier method [13]. Another approach

[15] In generating the alternate PDFs, the overall normalization factors $f_{{\rm N},e}$ were kept fixed at their values for the standard set. All other systematic shifts were allowed to readjust to the new PDFs.

Table II Experimental data sets used in compatibility studies. The data sets and PDFs are not identical to CTEQ6, but quite similar. PDF set A: the standard set. PDF set B: PDFs obtained by giving large extra weight to the H1 and BCDMS data sets in the fitting process.

	data set	N	$\chi^2[A]$	$\chi^2[B]$	$\Delta\chi^2$
1	BCDMS p	339	366.1	362.7	−3.4
2	BCDMS d	251	273.6	258.5	−15.1
3	H1 (a)	104	97.8	85.4	−12.4
4	H1 (b)	126	127.3	123.0	−4.3
5	H1 (c)	129	108.9	105.3	−3.6
6	ZEUS	229	261.1	288.5	+27.5
7	CDHSW F2	85	65.6	84.8	+19.2
8	NMC p	201	295.5	303.5	+8.0
9	NMC d/p	123	115.4	111.6	−3.8
10	CCFR F2	69	84.9	139.4	+54.5
11	E605	119	94.7	96.7	+2.0
12	E866 pp	184	239.2	242.9	+3.7
13	E866 d/p	15	5.00	5.6	+0.6
14	D0 jet	90	62.6	84.6	+22.0
15	CDF jet	33	56.1	55.1	−1.0
16	CDHSW F3	96	76.4	87.5	+11.0
17	CCFR F3	87	27.0	32.7	+5.9
18	CDF W	11	8.7	8.9	+0.2

is to apply the Bootstrap Method, i.e., to generate alternative PDF models from resampled data points. In the latter method, it has been found that to obtain significant changes in the PDFs, one must apply the resampling procedure to entire data sets rather than just to individual uncorrelated data points.

5. UNCERTAINTY ANALYSIS

As part of the CTEQ global analysis, a group at Michigan State University has developed several computational methods for analyzing fully the uncertainties of parton distributions [4]. In these calculations, we continue to use $\chi^2_{\rm global}$ as a figure of merit for the quality of model PDFs. The methods explore $\chi^2_{\rm global}$ in the neighborhood of its minimum, as illustrated in Fig. 1(a). The point represents the standard fit, denoted S_0, which corresponds to the position in parameter space where χ^2 is minimum; and the ellipse indicates nearby points that are also deemed to be acceptable fits to the global data set. The computational problem is to explore the full d-dimensional neighborhood of S_0. When this problem has been solved, we can assess the implications for PDF uncertainties.

C. The Hessian Method.

The Hessian is the matrix of second derivatives of χ^2,

$$H_{\mu\nu} = \frac{1}{2}\frac{\partial^2\chi^2}{\partial a_\mu \partial a_\nu}\bigg|_0 \qquad (6)$$

where μ and ν range from 1 to d. The classical error formula for the deviation of an observable $X(\mathbf{a})$ from its predicted value is

$$(\Delta X)^2 = \left(\Delta\chi^2\right)\sum_{\mu,\nu}\frac{\partial X}{\partial a_\mu}\left(H^{-1}\right)_{\mu\nu}\frac{\partial X}{\partial a_\nu} \qquad (7)$$

where $\Delta\chi^2$ is any specified increase in the value of χ^2 from the minimum. We obtain better numerical convergence by using the eigenvectors of the Hessian as the basis for the parameter space. To compute the variations of $X(\mathbf{a})$ along the eigenvector directions, we generate, for each eigenvector κ, a pair of points—along the + and − directions of the eigenvector—denoted $S_\kappa^{(\pm)}$. These are displaced from the standard point S_0 by a distance $T = \sqrt{\Delta\chi^2}$. Then the finite-difference analog of (7) is

$$(\Delta X)^2 = \frac{1}{4}\sum_{\kappa=1}^{d}\left[X\left(S_\kappa^{(+)}\right) - X\left(S_\kappa^{(-)}\right)\right]^2 \qquad (8)$$

which we call the “master formula” for calculating PDF uncertainties. The eigenvector basis sets $\{S_\kappa^{(\pm)}\}$ have been published for general use by high-energy physics.

D. The Lagrange Multiplier method.

The second method is used to analyze the uncertainties of PDF-dependent predictions, in a way that does not require the linearized error analysis. The idea is to perform a constrained fit, by minimization of the function $\chi^2(\mathbf{a}) + \lambda X(\mathbf{a})$ with respect to variations of the functional parameters $\{a_\mu\}$, for a fixed value of the Lagrange multiplier λ. The result is the best fit to data for which the observable X has the value given by that fit. As we vary the Lagrange multiplier λ, we thus trace out the curve of $\chi^2_{\rm constrained}$ versus X. The corresponding points in parameter space are treated as alternative models for the parton structure.

5.1. The question of tolerance

How should the uncertainty of PDFs be assessed? One simple, and too naive, idea is to use the increase of $\chi^2_{\rm global}$ to define the uncertainty. Let $X(a)$ be some quantity of interest that depends on the PDFs. We have, by either the Hessian or Lagrange multiplier method, the curve of $\chi^2_{\rm global}$ versus X, as illustrated in

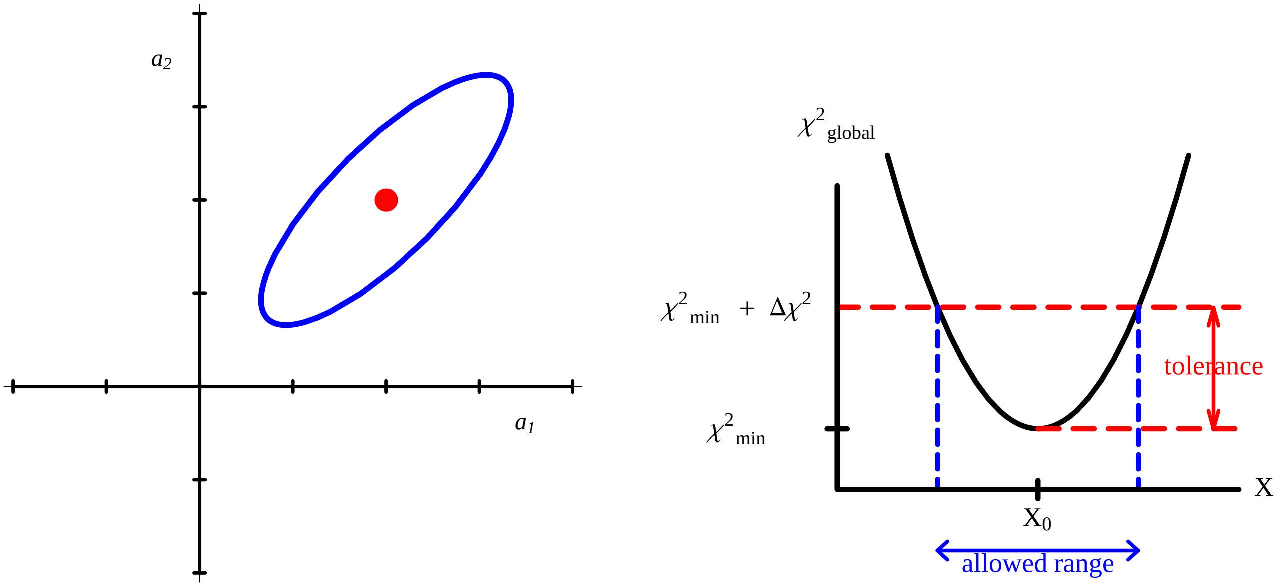


Figure 1: (a) The neighborhood of the minimum of χ^2 in parameter space. (b) Tolerance and allowed range.

Fig. 1(b), for constrained fits over a range of values of X. The predicted value of X is X_0, the value for the standard set S_0, for which χ^2 is minimum. Then for any tolerated increase in χ^2, denoted "tolerance" $\Delta\chi^2$ in Fig. 1(b), there is a corresponding allowed range for the quantity X. So, if we choose a canonical increase in χ^2 for the tolerance, the error on X is directly computed as $\pm\Delta X$.

For normal, i.e., Gaussian, errors, the tolerance would be $\Delta\chi^2 = 1$. However, it is well-known from experience that the changes in PDFs for which the global χ^2 increases by 1 unit are quite trivial. For example, we have seen in Section 4 that omitting or including a particular data set in the global analysis can easily induce changes in the global χ^2 much larger than 1. Nevertheless it is important to understand that the $\Delta\chi^2 = 1$ criterion is appropriate for normal statistics, even with systematic errors. Consider an experiment to measure an observable θ, with N measurements $\{\theta_i\}$ that are assumed to have statistical errors $\{\sigma_i\}$ and a common systematic error β; then we write $\theta_i = T + \sigma_i r_i + \beta\widetilde{r}$ where T is the true value and r_i and $\widetilde{r}$ are random fluctuations. The combined measurement is determined by minimization of $\chi'^2(\theta, s)$ in the manner described earlier,

$$\chi'^2(\theta,s) = \sum_{i=1}^{N} \frac{(\theta_i - \beta s - \theta)^2}{\sigma_i^2} + s^2 \left\{ \begin{array}{l} \theta : \text{ observable} \\ s : \text{ systematic shift} \end{array} \right. \tag{9}$$

The combined measurement θ_c and its standard deviation $\Delta\theta_c$ are then

$$\theta_c = \frac{\sum_i \theta_i/\sigma_i^2}{\sum_i 1/\sigma_i^2} \qquad \text{and} \qquad (\Delta\theta_c)^2 = \frac{1}{\sum_i 1/\sigma_i^2} + \beta^2. \tag{10}$$

It can be shown that the increase in χ'^2 for one SD of θ_c is

$$\chi'^2\left[\theta_c \pm \Delta\theta_c, s_0(\theta_c \pm \Delta\theta_c\right] - \chi'^2\left[\theta_c, s_0(\theta_c)\right] = 1; \tag{11}$$

here $s_0(\theta)$ is the optimal systematic shift as a function of θ.

The criterion $\Delta\chi^2 = 1$ depends on the assumption that the statistical and systematic errors are known. Suppose, however, that the true statistical and systematic errors are rather $\widehat{\sigma}$ and $\widehat{\beta}$ (independent of i). Then the SD of θ_c becomes $(\Delta\theta_c)^2 = \widehat{\sigma}^2/N + \widehat{\beta}^2$; and the increase of χ'^2 is $(\widehat{\sigma}^2 + N\widehat{\beta}^2)/(\sigma^2 + N\beta^2)$. For example, if the systematic error were *omitted* from the fitting procedure of χ'^2 (i.e., $\beta = 0$) then the increase in χ'^2 would be large for large N.

Rather than base the uncertainty estimate on one single number—the global χ^2 value, which we already suspect has non-ideal behavior from the minor systematic differences or incompatibilities between different data sets—we go back to inspect the χ^2's of the individual data sets. By the Hessian or Lagrange Multiplier method we generate a series of alternate fits, which differ by the value of the observable X of interest. Then each data set defines a "prediction" (from the fit with lowest χ^2) and a "range" (from a specified increase of its χ^2). The estimated uncertainty of X is the intersection of the allowed ranges from the separate experiments. This procedure, while it does not produce a precise confidence level, does give a reasonable estimate of the uncertainty of the prediction attributable to PDFs, in the spirit of a 90% confidence level. So long as no data set can rule out a value of X, a reasonable set of PDFs will exist that describes the global data satisfactorily with that value of X.

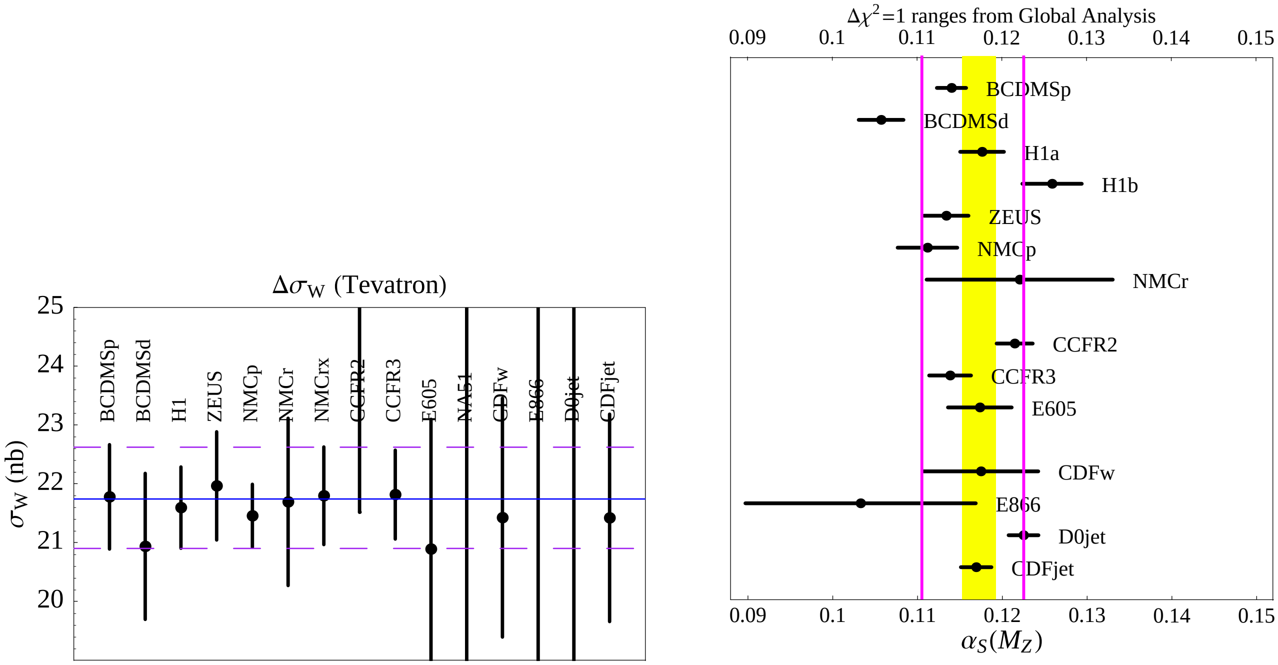


Figure 2: (a) Allowed ranges for σ_W from separate data sets in a global analysis; these are 90% confidence ranges. (b) Allowed ranges for $\alpha_S(m_Z)$ from separate data sets; these are $\Delta\chi^2 = 1$ ranges.

Figure 2 illustrates our uncertainty estimates for two cases. Figure 2(a) shows the 90% confidence ranges, for the separate data sets, of the cross section σ_W for production of a W boson at the Tevatron [4]; the dashed lines are our estimated uncertainty range on σ_W. Figure 2(b) shows the $\Delta\chi^2 = 1$ ranges for the values of the strong coupling $\alpha_S(M_Z)$. The shaded band is the PDG world-averaged 1-sigma range. The vertical lines are our estimated uncertainty range from global analysis. The global fitting is nicely consistent with the PDG value—another success of the standard model—but we find that the uncertainties in global analysis are larger than those of the experimental techniques used in the PDG determination of $\alpha_S(M_Z)$.

Returning to the question of the increase of the global χ^2, we find that extreme alternate PDF sets, i.e., that would be judged unacceptable by our studies, tend to differ in χ^2 from the minimum value by an amount of order 100 (for $N_{\rm tot} \sim 1800$ for CTEQ6). Therefore we have adopted $\Delta\chi^2 = 100$ as a typical tolerance for constructing the eigenvector basis sets $\{S_\kappa^{(\pm)}\}$ published for general use in PDF applications. Obviously this increase in χ^2 is very large compared to the normal 1.[16]

All PDF research groups must confront the issue of “tolerance,” and different groups have made different judgements. The CTEQ and MRST groups, which employ data from many processes, examine the agreement with separate data sets to estimate the uncertainty. Other studies, based on fewer processes, apply the Gaussian criterion $\Delta\chi^2 = 1$ [7, 9]. Further understanding of this issue will emerge as the groups compare their physical predictions.

6. CONCLUSION

Global analysis of QCD requires multivariate fitting to varied data with both statistical and systematic errors. Naturally any such problem is intricate. A routine statistical analysis will not solve the problem.

This review of CTEQ methods has emphasized the experimental errors—how measurement errors propagate to predictions that depend on PDFs. Other sources of PDF uncertainty exist, from theoretical assumptions. The MRST group has recently published an extensive study of theoretical uncertainties (the second paper in [5]). The paper identifies and analyzes four categories of theoretical uncertainty, and concludes that the theoretical uncertainty is as large

[16] When varying the PDFs to generate alternate fits, we keep the normalization factors $f_{{\rm N},e}$ fixed. The change of χ^2 would be smaller if the $\{f_{{\rm N},e}\}$ were allowed to readjust. For normal errors the optimized systematic shifts must be allowed to float when the criterion is $\Delta\chi^2 = 1$.

as, or perhaps larger than, the experimental uncertainty. In any case theoretical and experimental errors are entwined by the fitting procedure; a theory error may be compensated by correlated data shifts within the published range of systematic errors.

Because PDFs are important for the interpretation of current and future collider experiments [14], improved theoretical calculations and methods of uncertainty analysis will no doubt be developed, especially as data with ever higher precision become available.

Acknowledgments

I would like to thank Louis Lyons and the other conference organizers; also my MSU colleagues Jon Pumplin, Wu-Ki Tung and Joey Huston; and Sergei Alekhin for useful conversations about PDF uncertainties during a visit to MSU.

References

[1] J. C. Collins and D. E. Soper, arXiv:hep-ph/9411214 (1994).

[2] C. Pascaud and F. Zomer, Technical Note LAL-95-05 (1995).

[3] M. Botje, Eur. Phys. J. **C14**, 285 (2000) [hep-ph/9912439].

[4] J. Pumplin, D. R. Stump and W. K. Tung, Phys. Rev. D **65**, 014011 (2002) [hep-ph/0008191]; D. R. Stump *et al*, Phys. Rev. D **65**, 014012 (2002) [hep-ph/0101051]; J. Pumplin *et al*, Phys. Rev. D **65**, 014013 (2002) [hep-ph/0101032].

[5] A. D. Martin, R. G. Roberts, W. J. Stirling and R. S. Thorne, Eur.Phys. J. C **28**, 455-473 (2003) [hep-ph/0211080]; and arXiv:hep-ph/0308087 (2003).

[6] S. Chekanov *et al* [ZEUS Collaboration], Phys. Rev. D **67**, 012007 (2003).

[7] C. Adloff *et al* [H1 Collaboration], arXiv:hep-ex/0304003 (2003).

[8] V. Barone, C. Pascaud, and F. Zomer, Eur. Phys. J. **C12**, 243 (2000) [hep-ph/9907512].

[9] S. I. Alekhin, Eur. Phys. J. **C10**, 395 (1999) [hep-ph/9611213]; Phys. Rev. D **63**, 094022 (2001) [hep-ph/0011002].

[10] W. T. Giele and S. Keller, Phys. Rev. **D58**, 094023 (1998) [hep-ph/9803393]; W. T. Giele, S. Keller and D. Kosower, arXiv:hep-ph/0104052 (2001).

[11] V.N. Gribov and L.N. Lipatov, Sov. J. Nucl. Phys. **15**, 438 (1972); L.N. Lipatov, Sov. J. Nucl. Phys. **20**, 94 (1975); G. Altarelli and G. Parisi, Nucl. Phys. **B126**, 298 (1977); Yu.L. Dokshitzer, Sov. Phys. JETP **46**, 641 (1977).

[12] J. Pumplin, D. R. Stump, J. Huston, H. L. Lai, P. Nadolsky and W. K. Tung, JHEP **0207**, 012 (2002) [hep-ph/0201195].

[13] J. C. Collins and J. Pumplin, arXiv:hep-ph/0105207 (2001).

[14] D. R. Stump, J. Huston, J. Pumplin, W. K. Tung, H. L. Lai, S. Kuhlmann and J. F. Owens, arXiv:hep-ph/0303013 (submitted to JHEP).

A View of PHYSTAT 2003

John A. Rice
University of California, Berkeley, 94720 USA

1. INTRODUCTION

I present the view of a statistician for whom the landscape of statistical methods in high energy physics (HEP) and astrophysics was initially very foggy. As the fascinating conference progressed, the fog partially cleared and some of the underlying issues began coming into clearer focus. Because of my perspective, this summary is by no means comprehensive, but rather selectively concentrates on some of the statistical issues that arose during the conference. I thus apologize to the many participants whose contributions are not adequately reflected in what follows.

2. METHODOLOGICAL OVERVIEWS

The conference featured a number of valuable overview talks, some focused on statistical methodology and some on the underlying scientific issues.

Physicists might find it helpful to view the contrasting perspectives of Diaconis, Friedman, and Stark as spanning a large portion of the space of current statistical thinking: contemporary high-dimensional Bayesian methodology, pragmatic empiricism, and a hard-line frequentist view.

Many other talks can be placed within this barycentric coordinate system. Efron's contained substantial projections on all three, illustrating how the interaction between classical maximum likelihood analysis, Stein estimation, and Bayesian modeling of high dimensional Gaussians has enriched our empirical understanding of estimation in this prototypical context. The hierarchical models and smoothing techniques presented by van Dyk illustrated the modern Bayesian approach at work on an interesting concrete problem in astronomy. Reid provided a valuable, concise presentation of classical and recent frequentist methods of dealing with nuisance parameters. Genovese illustrated the use of modern methods of smoothing and simultaneous inference for curves, power spectra in particular.

Of the scientific overviews of statistical methodology, those of Nichol and Prosper echoed the empiricism espoused by Friedman. Barlow's presentation of statistical issues in HEP was much more classical and less broadly inclusive than that of Feigelson on statistical issues in astronomy. This contrast may be in large part due to fundamental differences between the ways data are collected in the two areas, HEP relying on carefully designed experiments, and astronomy being necessarily observational. Formal inference thus plays a larger role in the former and exploratory data analysis a larger role in the latter. Similarly, formal methods of inference dominated James' presentation. This contrast persisted in the presentations of Porter and Digel, which provided for the statisticians a concrete sense of the scope of statistical methods in the two sciences.

3. SOME SPECIFICS

In addition to the general overviews, I gained some particular insights relating to the use of statistical methods, among them:

What is meant by 180.1 ± 3.6 ± 4.0? Physicists are deeply concerned with systematic bias, much more so than is usual in statistics (Sinervo gave a very instructive overview). Variance is small in these high precision measurements, whereas in more typical statistical applications, bias is often swamped by variance.

The null hypothesis is often seriously entertained in HEP: the hypothesized phenomena may not exist. In many statistical applications, the null hypothesis serves much more the role of a strawman.

HEP experiments are done with extraordinary care—witness the clever techniques of blinding (Roodman).

Small counts are ubiquitous in HEP, and are frequently as Poisson as one can imagine for actual data.

As a consequence of the previous point, classical results in theoretical statistics, such as Neyman similar tests, are really of substantial importance, whereas to many statisticians these appear as the artifacts of a distant past. It is gratifying that some scientists really care about uniformly most powerful unbiased tests!

In addition to these classical concerns, more contemporary developments in statistics and machine learning play an important role, not only in astronomy, but also in HEP. Examples included the presentations of Askew, Cardossa, Cranmer, and Knuteson among others.

Care in constructing probability models, commensurate to that used in carrying out the experiments, can really pay off (Canelli).

Physicists rely on Monte Carlo (MC) to a degree surprising even for statisticians. (In fact, many MC methods have their roots in physics). For example, goodness of fit tests are often performed relative to a distribution that can only be expressed via Monte Carlo.

Finally, despite the best efforts of several patient physicists, "cuts" remain a mystery to me, but an interesting one.

4. FORMAL METHODS OF INFERENCE

Both Bayes and frequentist methods of formal inference rely on mathematical idealizations, such as probability spaces, measurable functions, independence, and distributions which are exactly known (at least up to parameters). HEP seems as well suited to this imaginary world as any scientific field—counts as Bernoulli as one could hope, experiments done with extraordinary care, careful and detailed modeling, both parametric and via MC, and precise numerics (Quayle). Using these concepts correctly requires care in modeling and in getting the math right (Linneman, Demortier).

Formal methods being applicable, there still remains the question of how to relate them to the science. This theme was present, at least implicitly in many of the talks and explicitly in others, for example Porter, Punzi, Rolke, Shawan. It is interesting that the HEP community seems to have settled on requiring more stringent standards for discovery than for setting upper limits when using the frequentist paradigm. There is also the use of code words, "evidence for" and "discovery of," corresponding to different significance levels. A detailed explication of why the 5σ level has been apparently accepted as conclusive evidence would be quite interesting.

Relating formal Bayesian methods to the science is not straightforward either, as illustrated by Shawan. Savage [3] wrote that one cannot enjoy the Bayesian omelet without breaking the eggs, but even having agreed in principle to make an omelet, many further decisions have to be made. Whether one enjoys the omelet depends up the quality of the chef, the quality of the ingredients and the taste of the diner. HEP and astronomy certainly offer high class ingredients and we have seen several skillful chefs at work, for example Diaconis, Loredo, Scargle, and van Dyk. One must choose what style of omelet: subjective omelet (Savage), uninformative omelet (Jeffreys), hierarchical omelet (van Dyk), least favorable omelet (Stark), or the truly exotic omelets concocted for high or infinite dimensional parameters. One wonders how these decisions could be made collectively by the army of chefs of an HEP collaboration.

In the statistics community, the Bayes/Frequentist wars have waned, in part because many of the old combatants have perished, and in part for reasons I discuss in the next section. The theme that contemporary statistics is not prescriptive emerged repeatedly: "much depends on attitudes and psychology" (Efron); "Bayes and frequentist are not opposing points of view" (van Dyk); James teaches both in parallel. The interpretation of statistical results within a scientific context has a large subjective component, and it is implausible that there exists a single, formal paradigm for how one learns from data, even setting aside the key role of informal exploratory data analysis.

5. STATISTICS IN HIGH DIMENSIONS

High dimensional statistical phenomena were a feature of many talks, and effective treatment of these sorts of problems is hard whatever point of view one takes, Bayes, frequentist, or opportunistic empiricism.

Maximum likelihood estimates behave badly in high dimensions, fluctuating too wildly as Efron pointed out for the canonical multivariate normal case. The unfolding problem is another well known example. A common method of stabilizing maximum likelihood estimates is via penalization, so that rather than maximizing log-likelihood(θ), one maximizes log-likelihood(θ) + J(θ,λ).

J(θ,λ) can often be usefully interpreted as the log of a prior on θ, or may arise directly from a prior, bringing Bayes and frequentist perspectives into close alignment. (Scargle's presentation is one example). But life is not so simple as just specifying a prior and integrating—the choice of prior really matters, as again illustrated by Efron. While we may think we may understand what it means to put a uniform prior on a scalar parameter belonging to the unit interval, what it means to put priors on high dimensional parameters is very hard to understand. Our experience of the relationship of geometry and measure in three dimensions does not provide good intuition for high dimensional Euclidean space, where:

- Almost all the volume of the unit cube lies near the corners.
- Most of the mass of the cube lies close to any subspace of R^n.
- Almost all the volume of the Euclidean ball lies near the edge.
- Gaussian measure concentrates on a thin shell far from the origin.

(For these and other cheerful facts, see [1]).

Of course the prior on the unit interval can "matter" in one dimension as well, which is reasonable from the classical view of a prior as personal probability. However our lack of intuition about measure and geometry in high dimensions and the exotic priors (cf. Diaconis, van Dyk, Scargle) makes interpretation in terms of personal belief at best highly formal. A 95% posterior region resulting from such priors is not 95% "credible" in the ordinary English language sense of the word. It is thus imperative to assess such procedures in terms of their frequentist and empirical properties, which again brings the perspectives into closer alignment.

We have heard relatively little about empirical Bayes procedures in this meeting, but the times seem ripe. In astronomy, for example, one measures repeatedly the properties of many, many similar objects sampled from a universe of such objects. It would thus make sense to not treat the analysis of each measurement de novo, but to integrate over the accumulation of experience. However, this is easier to say than do, and there are real challenges to be met in developing such methodology.

6. BEYOND FORMALISMS

There is more to statistics these days than dreamt of in the philosophies of Bayes, Fisher and Neyman. As I. J. Good wrote [2], "In our theories we rightly search for unification, but real life is both complicated and short, and we make no mockery of honest ad-hockery." Computing power has and continues to produce a rich buffet of procedures, as illustrated in the presentations of Knuteson, Prosper, Friedman, Nichols, Vilalta, Lee, Cardosa, Askew, Cranmer, and Levi.

Some may be concerned that this explosion of computing power coupled with unfettered imagination is leading to the degeneration of the stern discipline of statistics established by our forebears to a highly esoteric form of performance art. Looking over the spectrum of current work in statistics, one in fact does see far too many examples consisting of a proposed algorithm, a small number of simulations over a restricted range of scenarios, and maybe one real example.

There are, thankfully, constraints on unfettered imagination. First, one can try to do the math right; asymptotic analyses and considerations of optimality can provide insight into finite sample behavior. Extensive simulations, informed by such mathematical insight as is available, are valuable in sorting the wheat from the chaff. And finally there is the crucial test of utility with real data in real scientific contexts, revealing whether proposed methods actually lead to sensible and meaningful results. There is hope that evolutionary pressures will lead to extinction of the frail and survival of the fittest.

References

[1] K. Ball, "An Elementary Introduction to Convex Geometry", Mathematical Sciences Research Institute, 1977.

[2] I. J. Good. *Estimation of Probabilities.* MIT Press, 1965.

[3] L. J. Savage. *The Foundations of Statistics.* Wiley, 1954.

Panel Discussion

Panel members:

Persi Diaconis, Jerry Friedman, Fred James and Tom Loredo

Chairman: Louis Lyons

Lyons Now for the Panel Discussion. For some time on the PHYSTAT2003 web-site, we asked participants to submit questions for this session. Some of you did just that, and that was a distinct improvement on the first meeting we had at CERN where for a similar panel discussion, the number of submitted questions was zero which was perhaps appropriate for a meeting about low statistics. I asked the members of the panel to pick out some questions from the list (see page 309) and they will say something. After that the audience will be able to contribute their comments or corrections or whatever. So Jerry, I think you had chosen multi-dimensional goodness of fit.

Friedman Yes, I did. I actually have some transparencies.
[Jerry Friedman's remarks appear as a separate paper at the end of this Panel Discussion on page 311.]

Lyons Thank you very much Jerry. Anybody else on the panel want to say anything on that topic? OK

Scargle First a comment and then a question. I guess that is the usual order. David Wolpert, who is now at Ames, wrote a paper some time ago on an essentially Bayesian approach to a similar problem. It is 'Determining whether two data sets are drawn from the same distribution in maximum entropy and Bayesian methods', K. Hanson and R. N. Silver (editors), Kluwer Academic Publishers. The question is: At the beginning you assume I.I.D. (independently and identically distributed) Does that mean correlated data cannot be treated in this way?

Friedman I.I.D means that when we draw an observation, it doesn't know about the drawings before or afterwards. But you can have correlations among the variables.

Zech Probably you could compute the distribution of the test statistic by mixing two Monte Carlo samples. What I mean is that you just generate the amount of data in addition to the original Monte Carlo and then do the permutation of the Monte Carlo alone. Then you get no correlations with the data. Is this possible?

Friedman It is an interesting idea. Then it certainly isn't a permutation distribution that you would get. I am not sure, does anyone else know? To repeat the question, I presented two ways: (i) when the permutations have been performed over the pooled data and (ii) where you just repeatedly drew from Monte Carlo and directly computed the distribution under the null hypothesis. The question is what if you just generate one Monte Carlo and do permutations only within it. I have to think about that, I am not sure.

Loredo What about Brad? I am sure Brad knows the answer.

Efron I don't know the answer, I will think about. If I come up with an answer before we adjourn, I will let you know.

Snyder In some circumstances we are particularly interested in tests that don't require toy Monte Carlo. That is because we have situations where we can't really model the physics with the toy but we need the full

Monte Carlo and it is then too expensive in CPU time to run that many many times to produce the statistic distribution. We are particularly interested in a test where we have some analytical or at least computational means of deriving these distributions, like the one you mentioned perhaps.

Friedman Well the permutation test only requires one draw of the Monte Carlo so if you have the Monte Carlo generated once, that's all you need. The other one is if you use the k nearest neighbors as your machine learning procedure. Then you can directly compute the null distribution. You still are going to need the Monte Carlo to get the test statistic but you can just compute the null distribution without repeatedly doing the permutations test because you can actually derive it analytically. But keep in mind that the k nearest neighbors procedure is a kernel procedure and if you use Euclidian distance to find the distance between neighbors, you are stuck with the fact that you will lose power against alternatives where in different regions of the space the differences are in different variables.

Diaconis I would be helped by knowing what example you have in mind. I actually as a statistician was very interested in this problem. I don't know what the fine points are, how high dimensional are these distributions, how well do or don't you know them. That is something that many people in the audience know but I don't know. Could you elaborate a little?

Snyder A particular example I have in mind is a $D^{*}l\nu$ form factor analysis in which there are four variables in four-dimensional space that we measure and we are trying to fit. To a certain extent this is not quite analytic because we have to fold the detector acceptance into this. To generate the sample is very expensive. It requires a full Monte Carlo and we would typically have them generated only for one set of the parameters of the form factor - to get the others we then will weight the generated events.

Diaconis Why is it expensive? Because you have to simulate your detector in some way?

Snyder Right, you need the full Geant4 simulation of the detector.

Diaconis And how long does that take? I just don't have a feeling for the numbers. Does it take up to 10 minutes or 10 hours or weeks?

Snyder Weeks.

Diaconis Weeks. I see. To generate one point?

Snyder No, no, not one point but to generate a reasonable sample comparable to our data sample.

Friedman I think Harrison Prosper was mentioning a typical dimensionality of 17, you said, something like that? They are not really large like in some other areas but they are not super-small either.

Lyons So maybe it is time to move on to another question. One that actually a few people have asked me about sort of independently is question 13 on the list (see page 309) and that is whether hypothesis testing and parameter determination are equivalent. People quote the famous book by Kendal and Stuart that seems to say they are, and yet often we are told that they are not quite. So I think the experimental physicists would appreciate some enlightenment on that.

Diaconis I can start and a lot of people here can help. The idea is you have a family of probability distributions which is someway specified but has a parameter (and that parameter can be a vector value parameter). Perhaps you might have a test that the parameter $\theta = \theta_0$. Then often there a test statistic t and you will reject the null hypothesis that $\theta = \theta_0$ if t is bigger than some value. So that is a test. Then you can convert that into a confidence procedure by inverting it as we say, that is by taking the confidence interval for θ as the set of all θs that you don't reject at the 0.05 level or something like that. And so in that sense you can convert a testing procedure into a procedure for making confidence intervals, and you can go backwards. Now the difference is that the desiderata for a good test are that you want to have both the right level and you want it to be powerful against alternatives of a certain flavor. But the desiderata we have for a good confidence interval, well it would be the right coverage probability but also small in volume and perhaps other properties. So the desiderata are different, and if you were trying to figure out a good confidence procedure (and were able to say what you mean by good, and that is harder or less studied than saying what we mean by a good test), you would be led to different procedures I think because what you are trying to do is different with the two procedures. There is an equivalence but you can be led to very funny regions by just inverting and if you think about if you really want regions that are small, or convex, or have some other properties e.g. smaller in volume, then that leads you to a different procedure but if you inverted those and made them into tests, that is does my confident region contain H_0, if the confidence region were chosen to be good, it might not be that the test is particularly good. Anyway that is an easy off-the-shelf answer.

Lyons Very often we are interested in whether the value of a parameter is equal to zero or not. For example, we may be estimating the production rate of a hypothesized particle, such as the Higgs. If the production rate turns out to be non-zero, maybe that is a sign that we are actually producing the new particle we are looking for. So when we obtain a confidence range, we might look to see if it includes zero or not. So what you are saying is that the optimal way of determining the confidence interval might not be equivalent to testing whether the particle exists.

Diaconis It might be that the test you get is not the one you would be happy with for general use. I think trying to separate those is a good idea.

Loredo It might be worth commenting on the relationship between estimation and testing in the Bayesian framework. There is very similar math needed for parameter estimation and for calculating a Bayes factor for testing. But there isn't a one-to-one mapping from, say, how big a credible region must be to include the null value, and whether you conclude that posterior odds against the null are strong. That will always depend on the range that you searched to fit the parameter of interest—that enters as a prior volume factor in the Bayes factor. In time series problems, we search through a range in frequencies and these ranges are things that we are very comfortable saying we determined beforehand and can write down for computing our Bayes factors. In other settings, like determining if the amplitude of a signal is zero or not, there is sometimes not an obvious upper limit to this range a priori. There is just traditionally or conventionally less comfort with that kind of a specification. That basically is the mathematics of it. You are basically measuring how many "sigmas" your best value is away from the default, the null point, but measured relative to the volume that you have searched.

Efron The way I tend to think of it is that estimation is usually conducted with your mind thinking of a smooth parameter space of fixed dimension, maybe many dimensions but you'll give the parameter points relatively equal weight within that space. In hypothesis testing you almost always have subsets of lower dimension that you are especially interested in and the problem cuts across dimensions, like between zero dimension and one dimension if you are testing $H0 : \theta = \theta_0$. So even methods like confidence intervals that use testing to get estimates really do tacitly assume that you are rather equally interested in all the points in the parameter space.

Raja Perhaps I can add a little bit to this problem by taking an example I have been working on. That is the problem of goodness of fit in unbinned likelihood fitting, for which I have proposed a solution. If you had no goodness of fit criteria and you just did maximum likelihood fitting, you will be able to determine the optimum values of parameters with no measure of the goodness of the fit. But you have no way of testing if the theory fits, because you just get the optimum fit parameters. If you use a likelihood ratio of theory likelihood versus "data likelihood" worked out using pdf estimates of the data, then a goodness of fit results. This tests how well the theoretical hypothesis fits the data. If you now do goodness of fit for two separate theories, using likelihood ratios, then you can compare each theory to data on its own merit. The ratio of the two likelihood ratios will compare two theories against each other, without telling us if either theory fits well. The key ingredient in obtaining goodness of fit is the concept of the likelihood of the data using an estimate of the "data pdf".

Lyons OK, maybe we will move on to another question then. Tom, you have something that you would like to talk about.

Loredo This is a question that is not on the list but I have been urged to raise it by a few participants, both astronomers and a statistician who will go unnamed (though he is reviewing the Conference later!). This is just to clarify for those of us who aren't particle physicists some issues involving cuts: what is done and when it is done and why it's done. So I guess there are two things I would like to say. The first would be just to explain my level of understanding of what happens and then ask for some input on refining that; and secondly to raise some points about cuts that are used in astronomy and whether there is some relationship to what is done in particle physics.

So the first issue, to kind of state it provocatively, would be: "Why ever cut, why not just always weight?" In the opening talk of the Conference, an algorithm was described that weighted, but then through the Conference, my sense of the community is that there is no obvious consensus on what should be done by different collaborations or different groups when there are overlapping distributions. Some groups try to find, by whatever criteria, some place where they will make a cut and say everything on one side will be classified one way, and everything on the other side will be classified the other way; and all events are given equal weight within those two divisions. Another subset of groups, especially those focusing on likelihood ratios or Bayesian methods, apparently seem to like to weight by the PDFs, but perhaps also including a cut that gets tuned by Monte Carlo analysis or blind analysis, and my impression is that even in those situations they would like not to have to cut. The cut reflects some impression that the underlying models that are used to calculate the likelihood ratios aren't accurately describing the experiment, and so in places where some of these ratios are extreme, one just cuts the data rather than be contaminated by misweighted events. That's my impression, and so my question is whether it is an accurate impression, and if it is, whether there are well-established reasons for this divergence in viewpoints. So for those who cut and don't weight, is there some reason why they are opposed to weighting? And for those who do weight, have they done calculations showing how much things are improved by weighting? That is kind of an open question.

Yabsley Just a comment on one part of that. I think it might have been a little bit misleading, in that some part of the descriptions of our analysis get suppressed in any of these presentations. I have never met a particle physicist who doesn't cut and I have never seen an analysis that has no cuts. If some of those cuts haven't been mentioned, it is simply that to us they are too obvious and too boring.

Loredo So it is done for the sake of bandwidth for example?

Yabsley Some of it is done for the sake of bandwidth, and some of it is done just because there are events which are absolutely utterly irrelevant to the problem that you are trying to deal with and you just want them out of there. I will give you a very specific example in my area. I work with the B-factory but I am actually doing charm studies and I want to get all of the beauty events out of the sample before I work on it, partly for bandwidth but partly also just to reduce the confusion. It creates irrelevant backgrounds but there are other backgrounds as well that we would have to model if they were included in the data set. But they get suppressed by the same cut that gets rid of the first one. So almost every analysis in my group contains one of a family of cuts that essentially eliminates a class of events, reduces the data size and simplifies the fitting, by simplifying the kind of hypotheses you need to make about your backgrounds. That is a specific example but it has generic features, and all of these data sets have that. But it is not just data volumes, it has to do with the fact there are quite distinct processes going on, different classes of things and you just want to get rid of some of them.

Loredo So is it that you know without calculation that if you weighted, those events would have negligible weights so you just don't bother using up the computer time to do that? Is that basically what is going on with that?

Yabsley That is part of it. There are some things you know without calculation would have zero weight. There are others that you know would introduce complexities into the problem that are just not worthwhile.

Lyons I think the cuts that Bruce is talking about are where you get essentially complete separation, between things you are interested in and things you are not. I think Tom's question is for overlapping distributions, where the cut is really doing something.

Barlow Can I just say you are absolutely right. There are a lot of people making cuts and saying these are just inside so we will treat them as absolute pukka data events. These are just outside so that we are not going to look at them any further. But this is wasteful of data. It is true as Bruce said if you start off with 80 million events and you are going to reduce that to 80 events, you don't want to have to carry over 79 million very very small weights around through your histograms. But at a last stage of the analysis when you are going from a few thousands to a final few hundred, then weighting will actually buy you statistical precision and people ought to be doing it more. It is true that if you don't really understand the background PDFs properly, you can get weights which are slightly wrong and that will introduce errors, but if you cut instead of weight you have the same problem because you are not quite sure how many events made it through your cut, or what your backgrounds are, or what your efficiency is. So cutting is simpler but doesn't get you out of the problem of ignorance of PDFs.

Snyder I actually have something of a contrary point of view to Roger's and I think in effect in some ways BaBar is starting to go overboard in doing likelihood fits and rejecting simple cut-based analysis, which are perfectly adequate for the problem at hand, simply because they are not using weights. Sometimes the problem is so simple and you get a good enough answer. So why go through all the work to crank up this apparatus? The other one is systematic errors. Weighting usually means a likelihood fit. It has a statistical regime where it is the best thing to do. At very low statistics, it's bad because you get into these problems with the likelihood that isn't really working yet, and you don't have enough statistics for the central limits theorem to be behaving properly, so you have to do a whole lot of toy Monte Carlos to interpret the errors.

Loredo So just to clarify, you mean that you don't actually know the likelihood function in that region because there are small statistics?

Snyder It's small statistics and we can get around that to an extent. If we can do toy Monte Carlo we can interpret the error, but the error we get out of Minuit in these situations is unreliable, so the regime of very low statistics is where cut and count is much simpler. Take the example from Frank Porter where you have a prediction of five background events and you see nothing in the signal box. Well there is a cut and count that is very simple you know exactly what it means. I predicted five, I see zero, end of story. Everybody knows what it means. If you do the likelihood you get all this funny stuff with the thing going negative or even unmathematical and you put in some ad hoc fix for that. I would say in that kind of situation, with low statistics, you are better off sticking with cut and count because it is easily interpreted and you know what it means. Another place is intermediate statistics. There you start to really gain from using the likelihood method, you make a difference in statistical claims. But if you go to very high statistics then the fact that you don't understand the PDF of the signal in particular starts to come in and in some of our analysis, already the largest systematic error is the uncertainty on what we call the ΔE distribution. It is the difference between the B particle energy and the energy of the machine, and so if you go to high enough statistics, you will be better off cutting well outside the peak in that signal, in which case it doesn't make very much difference if you get it wrong because it doesn't change the prediction of how many fall out very much because it is almost zero anyhow and the backgrounds are typically flatish. So you are less sensitive to misunderstanding the background than you are to misunderstanding the peak if you cut broad. So there comes a regime with high statistics where again cut and count maybe a more robust a method with a smaller systematic error. There is always a trade off between the systematic error and statistical one at some level.

Loredo Have there been studies showing different regimes or different instruments where you know cut and count didn't hardly lose anything over a likelihood or actually was better? Have people actually compared the two to see how much you get from one versus the other in any particular analysis?

Snyder There have been cases where people converted and they see the improvement, though probably not as much as you might hope.

Loredo I have no idea what to hope and that is why I am asking.

Prosper In fact the issue of whether to cut or to weight has been studied at the last Snowmass meeting. Walter Giele and other people have been looking at whether there is some gain in actually weighting. Let me give you the context. We are interested to know whether one could weight distributions so as to obtain, for example, the spectrum of transverse momentum for jets rather than arrive at the distribution through a series of cuts. Typically, what we do right now is that we make clusters in space on the sphere. These clusters of energies are called jets and are actually cuts. So if something is within the circle you add it to the cluster and if it is out of the circle, you ignore it. What Giele and others showed was that if you weighted these energy deposits rather than actually doing cuts as you form these jets, you increase the precision quite considerably in the estimate of the jet cross section. In fact he has a paper which has been published. So certainly there are people who are making the case even if they are doing physics, in actually comparing theory with data. Weighting in fact is something that we really look at very seriously.

Loredo So is that a case where we just happen to know the distribution very accurately? Do people worry about the systematic error part?

Prosper The problem with cuts as Roger noted is that at a boundary of course small changes in the resolution function of your experiment or what you think the resolution function is can have a rather large effect. It seems to be the case that even if your weights are not quite correct that the error that you get making a cut is slightly larger than the error you get with slightly incorrect weights. I cannot prove that, but it seems to be what one finds.

Snyder As a matter of fact, the answer depends a lot on the problem. A rapidly falling p_T spectrum in fact is very sensitive in a cut analysis to how well you understand the energy scale and weighting might be just as good or better, and gain you in statistics whereas the one I was describing of a signal which is a sort of a peaky thing can be very insensitive to where you put the cut. So the cross-over to where weights become advantageous is going to vary from physics channel to physics channel. I think it just has to be looked at for each case.

Raja A quite different example entirely. In $D0$ Run I, we had discussions on identifying electrons with greater efficiency and we had a cut method using several calorimeter layers transversely and longitudinally. When you make several cuts, you lose signal efficiency very fast. One of the first multivariate algorithms used in $D0$ was the Hessian matrix and we had a 40 x 40 matrix at each of the rapidity bins trained using Monte Carlo. It would look at all the correlations in the cluster. This gives you a single χ^2 number, so the only cut is on that χ^2 and the gain is much better efficiency that way. So there are cases where pattern recognition benefits by using multivariate techniques. This gain in efficiency was crucial to $D0$ in all its analysis involving electrons, especially the top quark discovery, which was made without the use of a magnetic field, which would have allowed matching of momentum and energy.

Diaconis I have a comment and then a reaction one of which is from before hearing this discussion. Statisticians also have a way we talk about this and it's hard and soft thresholding. If you want to estimate a lot of parameters, you can either set some to zero and then estimate the rest; or do soft thresholding. It has been sort of proved in various circumstances that soft thresholding is more accurate. That certainly is all built into wavelets, the modern way we send pictures and it's has been very successful. But what I hear is this tremendous amount of knowledge that in this circumstance we do this, and we all know that. One wonders always is that written down so that anybody else knows, or do we just have to talk to you and talk to you and talk to you, and that is how your students learn. That would be one reason for maybe trying to write it down more carefully. For example, when I was listening I heard you say one thing was that the cost in the complexity of keeping track of all of the other things just isn't worth it. Well, you could sort of try to say "If it cost you this much to keep track of things, you know, what is the cost against accuracy trade off?" Somebody else was saying that the systematics come in and you could try to write that down too and maybe trying to compile knowledge and trying to say “How would you write down as a math problem or as an applied math problem this question of whether to cut or weight?” I think that one good thing that would come out of that is just recording some of the things you guys know and think about and it would be valuable. I think you were saying by nodding that it is not all written down any place, it is all seat of the pants.

Friedman Is there a general agreement on the seat of the pants procedure? I mean some people might write different documents?

Roe I just wanted to comment that usually you don’t know your distributions terribly well way out on the wings. So away from the wings usually you want to do weighting. But I think way out on the wings almost everybody is going to cut these things just because the lack of knowledge there is going to kill you.

Loredo Yeh, I think in one of the contributed sessions, there was a paper by Canelli from $D0$ and that was my impression of exactly what they did. They tried to weight as much as they could and they use blind analysis and knowledge of the experiment to try to find where they couldn’t trust things, and that is where they cut. Is that an accurate description what $D0$ do?

Someone Is this is just an intellectual question, or is there something behind there?

Loredo Well, just for the reason that Persi alluded to, both from the Bayesian perspective and even the frequentist likelihood perspective. Conditional on knowing the distributions, you always want to weight. So I wanted to know how much of the use of cuts was just a distrust of statistics, how much of it was the fact that you don’t really know these distributions, how much of it is tradition? And is there room for other people to do work in this area?

Lyons Jerry do you want to comment on the first part?

Friedman Yeh, this sort of sounds to me like déjà vu all over again. Thirty years ago when I was a student in High Energy Physics, we had exactly these same discussions and I am just amazed that it hasn’t been resolved. I mean it has been 30 years and whether it is better to cut or to weight, and how should you cut, and how should you weight, ... So I don’t know, maybe we are coming to a closure here but . . . .

Someone No, at the next Conference.

Friedman It just reminds me of my youth.

Loredo I mentioned there might be a second part of my question on cuts. There are related issues in astronomy that Jeff Scargle reminded me of. In astronomy we deal with cuts, not so much in the setting of trying to separate overlapping distributions, but when we are sampling a single distribution. For example, we look at the distribution of stars as a function of distance, and there is a cut in our survey at the lowest observable brightness, and we want to measure the distribution function. The complication that arises often—and it is pretty well understood how to deal with it from work dating back to Eddington and Jeffreys—is that we cut on an observable that’s related to the actual thing we are going to infer, like we would cut on brightness and we would like to infer distance, and that observable has uncertainty. Also, the distribution we are trying to infer, the distance distribution, is changing in space. Near the cut in brightness, due to uncertainty, there are objects scattering across that cut in both directions, but because their underlying distribution is inhomogenous there are more scattering in one direction than the other, and if you ignore that there will be biases in the survey. I was just curious if there is a similar situation here, whether that’s dealt with in this separation of two overlapping distributions somehow, and how? In astronomy this is known as, well, there isn’t really a single accepted general name but it arises under the names of Malmquist bias, Lutz-Kelker bias, edge bias, or threshold bias.

Lyons In particle physics, if you have a cut you would estimate what background you have inside your cut and would also allow for the amount of signal that was outside the cut.

Loredo Well, in astronomy again, it is just a single distribution that you are trying to infer. First of all for each one of your objects you need to know the uncertainty distribution for the observable. Then you also need to calculate—and we do this by Monte Carlo simulation of our surveys or other methods—as a function of the unknown true parameter, what is the efficiency of coming into the cut. I hadn't seen both of these functions appear in the physics analysis so I was just curious if this stuff is all known and you know how to deal with it.

[Nods and general assent from several Particle Physicists in the audience]

Diaconis Let me read one of the questions to all of us, myself included. "What is the most effective way for physicists to get collaboration started with local statisticians? Joint graduate students support, one-to-one talking, . . . ? " Well, I think at least for the people that I know here who are statisticians, we would be thrilled to get involved in serious collaborations. It is not that we can do so much in "Hey, what is the reference for this?" I mean of course we are happy to help for that, but what we actually want is to sit down and say "Look, this is the problem, let me take several cups of coffee, several weeks and explain how we do things, and what we do, and what is the state of the art." I think that any of us would really be happy to get involved in long-term collaborations but that means you have to want to do it too. I spent several years at SLAC, Jerry hired me when I first started here and it was very very hard to get a physicist then to let you actually interact with the problems. They were happy to take a reference but that is probably not the right way to go. That is probably trying to say "OK, here is serious problem. We want to get you seriously involved and let's talk about it and try to do some work together." Now one of the things I have been doing, I find I have done it a lot, is to say there are good and not so good statisticians. Now, that is not a Bayesian and non-Bayesian dichotomy. What I mean by that, well first of all there are people who are good at doing calculations in the back-room, but aren't so good at talking to you and interacting and figuring out what the question is in English and stuff like that. When somebody asks me a question, instead of my trying to give them an off-the-shelf answer, I say "What university are you at? Ah, X is in your university, talk to X." OK, I am happy, I think any of us, you know, John Rice, Brad, Jerry, we are happy to answer that question. You know, I am diaconis@math.stanford.edu. If you want to write and say "Look, I am interested in this", you have to sort of say what you are interested in. "Who's local, who I can talk to?" I will try to find you somebody both young enough that they have the time and interest and not just spend 5 minutes talking to you, and who is good. I think that is a very important service that could be provided, that is a statistical exchange. Every university has at least one, some have more than others, so I urge you to cultivate a local friendly statistician.

I have one other thing I want to say. When I think about what I heard, and about what would help your practice most, I come up with a plea. It's the fact that everybody is using χ^2. Well bless you, you are used to it, but none of us uses it, and there is a reason. So please take a look at some treatment of modern goodness of fit. There is a very nice book ('Goodness of Fit Techniques') by Ralph D'Agostino and Michael Stephens. It's 15 years old, but it will say what we knew back then. A very very nice survey paper by Fann is in an article in JASA. I mentioned one Bayesian test in my talk. There is another using Bayesian testing in a paper by Isabella Verdinelli and Larry Wasserman, and all the references to these papers—there were three of them—are in the article that I talked about with Mark Coram, which is in Journal of Physics A36 (2003) 2883. It is also on the web, and that will at least talk about the state of the art. I mean, using χ^2 the way you people do is really like throwing away a fair chunk of your data. If there is one thought to get across, at least take a look at it.

Lyons Thanks very much Persi. Certainly, we all welcome this remark that there are statisticians out there who would really like to help us. Just from listening to the talks at this Conference, there are clearly problems that non-statisticians struggle with, for instance, the treatment of nuisance parameters in a frequentist-type confidence interval or upper limits. And so willingly, more than willingly, we very much welcome help on that one and a whole host of other problems too. So I think your spare time is going to get used up quickly in the future. Anybody else have anything they want to say on that one?

James I'd like to talk a few minutes about the likelihood principle because somebody has said "Is the likelihood principle sensible, if so why, if not why not?" I suppose that if somebody is as worried as I was a few years ago, maybe I can calm this person down a little bit. I was very worried because the likelihood principle didn't seem sensible to me and yet is something that can be proven in principle. Now, I know Brad said once that the likelihood principle is an amazing thing which, although it can be proven rigorously, seems not to be true. And that seems in fact to me to be the case, I agree with that. A lot of very reasonable people believe it, so I can't say that my opinion is the only one. There are a lot of people who believe it but the way you prove it is based on sufficiency and ancillarity. I don't think you have to worry about this proof. The proof is all right if you believe the proof, if you don't believe the proof it is not all right and a lot reasonable people don't believe the proof. I have read Birnbaum's article and I have read Berger and Wolpert, and all of that. I tried very hard because if it is a real principle that you can prove, then you ought to believe it. I would like to believe it but frankly I don't see it. Now it's been pointed out that it contradicts practically all of frequentist statistics, so that is another good reason why maybe a lot of people would not believe it. So I think the argument about whether it is true or not will go on for a long time, but if you don't believe it don't worry about that. I don't think it's serious.

Lyons Does anybody have a different opinion?

Diaconis I certainly have puzzled over the likelihood principle too because I basically believe that people are sensible and that if there are frequentist procedures that are widely used, there must be some good Bayesian justification for them and that is very often true and when it's not, often people agree that we want to do something about it. For me an essence of the likelihood principle says that you shouldn't use data that isn't the data that you actually saw in the experiment. Thinking after you collect your data about data that might have been is going to lead you into difficulties roughly. And therefore, p-values for me are very very difficult. But that is not from higher principles, it is actually visceral. That is, you know a lot of other things might have happened, but I have the data that I have, I have to make a decision now and what should I do when thinking about what might have been really does seem taking me into some world of fantasy that doesn't seem relevant to the problems that I am being faced with. So that is something where I think I relate to the likelihood principle in a direct way and the fact that that is what the likelihood principle leads me to makes it a little bit more believable. But it is difficult for me too, other than that it follows from Birnbaum's argument and if you are a Bayesian. But it's still a hard thing to think about.

James Birnbaum doesn't believe it any more by the way.

Efron I just want to point out that what Persi said contradicted what Jerry said.

Friedman That has been true for the whole Conference.

Diaconis But we like each other anyway.

Loredo I'd just like to make a little comment about that. Some people probably don't really know what the likelihood principle is or where it comes from so I'll try to briefly explain it. It says that in two experiments where you have likelihood functions that are proportional you should come to the same inferences. That is not generally true of frequentist procedures because they don't just use the functional dependence of the likelihood on the parameter, they also use its dependence on the data. And two likelihood functions that are proportional to each other as a function of a parameter could have very different data dependences. So let's turn to the proof that Fred referred to. The likelihood principle does not seem obvious to many people, but there are two simpler principles that seem compelling to more people that this proof shows together lead to the likelihood principle. One of those is the sufficiency principle which says that your inferences should only depend on values of sufficient statistics, if they exist. This is generally true of common frequentist and Bayesian procedures, so it is not controversial. The other principle is the one that is the focus of controversy, and that is the conditionality principle. And that principle says that if I have a choice between two experiments and I choose one by making a flip of a coin, my inferences should depend only on the actual experiment I chose to do and not on the fact that I might have chosen the other one. That seemingly weak statement and the sufficiency principle together imply the likelihood principle. In fact a slightly stronger version of the conditionality principle alone implies the likelihood principle. So most of the arguments that I have seen against the likelihood principle look at the conditionality principle. I have examined those and none of the ones I have seen against it seem to have even a grain of truth to me. But they obviously have more than a grain of truth to some smart people. That is just to give you a sense of where the discussion on this is in the statistics community.

Genovese I just wanted to point out, in response to what Tom said, an argument in that whole conversation that hasn't received as much of an airing as I think it deserves. James Robins at Harvard School of Public Health makes the argument that our intuition about foundational principles like conditionality, or various related versions of it, has been developed on very low dimensional cases. He puts forward a semi-parametric model in which the conditionality and likelihood principles lead to bad or nonintuitive outcomes, such as that any procedure with good frequentist performance would violate the principles. It is a pretty interesting argument. The original paper [Robins, J.M. and Ritov, Y. (1997), "A Curse of Dimensionality Appropriate (CODA) Asymptotic Theory for Semiparametric Models", Statistics in Medicine, 16, 285-319.] is difficult reading but there is a volume of vignettes in Statistics with a nice summary [Robins, J. and Wasserman, L. (2002). "Conditioning, Likelihood, and Coherence: A review of some foundational concepts" in Statistics in the 21st Century, (eds.) Raftery, A.E., Tanner, M.A., and Wells, M.T., Chapman & Hall/CRC, 431-443].

Loredo Yes, I know the argument. I don't find it intuitively unreasonable. OK. So I think there are people on both sides.

Shawhan Tom, I am just hoping for clarification of what you mean by sufficient statistics in the explanation you just gave. The rest of it I thought was very clear.

Loredo If the dependence of the likelihood function on the parameters is determined entirely by some smaller dimensional piece of the data (smaller than the dimension of the data set itself), that piece of the data is called a sufficient statistic. Most of the nice models that we are familiar with in textbooks are ones that have minimal sufficient statistics, so the number of pieces is equal to the number of parameters. That is a nice situation to have.

Friedman I guess when you get old your mind wanders back to your youth. As a young student in High Energy Physics back when High Energy Physics was a young discipline, there were two documents that sort of laid out statistics to be used in High Energy Physics. One was by Frank Solmitz, and that laid out things in general terms, but also discussed in a fair amount of detail the Bayes-Frequentist controversy. The other was by a fellow named Jay Orear who, in a small note, introduced me to the likelihood principle which really opened my eyes. And at this Conference and in watching you do your work I see that most of what they laid down is pretty much still what you are doing, still arguing about cuts and still using the χ^2 test which was introduced by Solmitz. So I just wanted to point out that statistics has progressed in the last 30 years since Jay Orear and Frank Solmitz wrote those papers. I realize that the statistics literature is fairly impenetrable. But at least for those of us around here, I would just like to echo Persi's offer that if you have a problem, don't just come to us on how to do a t test when you have censored data type focused questions, but just sort of general questions "How do you go about this or that?" We will be more than happy to help you.

Lyons OK, clearly that is not the end of the discussion, and I am sorry to cut it short just when it was heating up, but we do have the tea break and the Conference Summary coming. Also Jerry's and Persi's offers seem a good place to end this Panel Discussion. So thank you all very much.

Panel Discussion Questions

The following questions were submitted by PHYSTAT2003 participants for the Panel Discussion session.

Question 1: Is it worth worrying about whether a method is Bayesian or Frequentist? Is it OK to mix Bayes and Frequentist methods in a single analysis?

Question 2: Is the likelihood principle sensible? If so, why; if not, why?

Question 3: How to perform 'Goodness of Fit' tests with sparse data in many dimensions.

Question 4: How can we have a measure of goodness of fit in one dimension when the requirements for the chi-squared test aren't met? There are two schools of thought working on this. One group tries to wring some sort of meaning out of the value of the likelihood function. The other group tries to find a modified chi-squared statistic with acceptable properties. Is there any hope for either approach?

Question 5: We have seen a plethora of new goodness-of-fit test-statistics (i.e. just of a single hypothesis) proposed in the past few years by physicists. What are the minimum criteria, or essential properties, for a goodness-of-fit test-statistic to possess in order to be suitable for general use? (General use would include, for example, evaluating the quality of an unbinned fit to data in N dimensions with M free parameters.)

Question 6: Neyman and Fisher disagreed sharply about hypothesis testing. Neyman believed it was always necessary to specify an alternative; Fisher, who championed goodness-of-fit tests, thought otherwise. What do today's statisticians think about the question of whether or not an alternative is needed? If one is needed, why? If one is not needed, why?

Question 7: We want to calculate upper limits from "unsuccessful" search experiments (looking for rare or new phenomena), in order to quantitatively measure progress in sensitivity, and compare experiments and techniques. We prefer to have frequentist coverage properties, but also desire inclusion of systematic uncertainties. Bayesian techniques do this naturally, but aren't invariant under reparameterization, and, in the absence of observations, produce limits notably dependent on priors. And we haven't found each arguments for choice of a default prior to be convincing.

So some say we are asking the wrong question, and shouldn't worry about properties of limits on things we haven't observed. Any comments or suggestions? Do any other fields worry about this?

Should we just use a range of priors and draw a wider line to describe our limits, or somehow combine a suite of standard priors?

Question 8: Do 'good' priors exist in one and in many dimensions?

Question 9: It seems to me that the practical difference between objective and subjective Bayesians amounts to this: Objective Bayesians use a formal rule to determine in which parameter the prior should be flat, whereas a subjectivist Bayesian makes that choice herself by virtue of her subjective choice of a prior in some parameterization that she considers meaningful. Admittedly this is a caricature, but is it fair?

Question 10: When using a mixture model and the EM (Expectation/Maximisation) algorithm, is there a way to know how many components should be used in the model?

Question 11: Imagine I have a data set and 2 models. One with say 5 parameters and one with more than 5 parameters (say 6 or 7). What is a 'correct' way to choose between the two models? To be more specific: for example for both models I do the likelihood analysis, I compute likelihood surface, chi-squared, probability distributions etc etc. How shall I use this information to select the more probable model? Is this actually a question that statistics can answer? How can one apply the 'Occam razor' to select simpler models? Is this the correct thing to do? How to address the problem if a) one of the 2 models is a subset of the other one; or b) completely different. How does the choice depend on priors? Can it be made prior-independent?

Question 12: What is the relative efficacy of different approaches (likelihoods, neural nets, or other) to the problem of multi-variable event classification?

Question 13: To what extent are 'Hypothesis Testing' and 'Parameter Determination' equivalent?

Question 14: Physics results are quoted as a point estimate plus an error interval. What is the point estimate which should be combined with the error interval? Should it be bias corrected? Is efficiency an issue? What about variable transformation properties?

Question 15: Ways of overcoming problems related to integer data.

Question 16: How can 'blind analysis' be applied in the case of discovering in a distribution a peak which you did not expect to see? Is this not the most interesting case?

Question 17: Are confidence intervals intended to form a basis for decisions or are they just produced for contemplation?

Question 18: Based on the talks you heard, where do you feel our statistical practice would most benefit from improvement? Bonus: how?

Question 19: What is the most effective way for a physicist to get collaborations started with local statisticians? Institutes? Joint graduate student support? One to one talking?

On Multivariate Goodness–of–Fit and Two–Sample Testing

Jerome H. Friedman
Department of Statistics and Stanford Linear Accelerator Center,
Stanford University, Stanford, CA 94305

It is shown how classification learning machines can be used to do multivariate goodness–of–fit and two–sample testing.

1. INTRODUCTION

In the goodness–of–fit testing problem one is given a data set of N measured observations $\{\mathbf{x}_i\}_{i=1}^N$ each of which is presumed to be randomly drawn independently from some probability distribution with density $p(\mathbf{x})$. The goal is to test the hypothesis that $p(\mathbf{x}) = p_0(\mathbf{x})$, where $p_0(\mathbf{x})$ is some specified reference probability density. Ideally, the test should have power against all alternatives. That is as the sample size N becomes arbitrarily large, $N \to \infty$, the test will reject the hypothesis for all distributions $p \neq p_0$ at any non zero significance α level.

A related problem is two–sample testing. Here one has two data sets: $\{\mathbf{x}_i\}_{i=1}^N$ drawn from $p(\mathbf{x})$, and $\{\mathbf{z}_i\}_{i=1}^M$ drawn from $q(\mathbf{z})$. The goal is to test the hypothesis that $p = q$, again with power against all alternatives; as $N \to \infty$ and $M \to \infty$ the test will always reject when $p \neq q$. Two–sample testing can be used to do goodness–of–fit testing. A random sample $\{\mathbf{z}_i\}_{i=1}^M$ is drawn from the reference distribution $q = p_0$ and then a two–sample test is performed on $\{\mathbf{x}_i\}_{i=1}^N$ and $\{\mathbf{z}_i\}_{i=1}^M$.

In univariate (one–dimensional) problems each observation $\mathbf{x}_i$ (and $\mathbf{z}_i$) consists of only a single measurement. In this case there are a wide variety of useful and powerful goodness–of–fit and two–sample testing procedures. Some of these can be extended to two or perhaps three dimensions if the sample size is large enough. However, when each observation consists of many measured attributes $\mathbf{x}_i = \{x_{i1}, x_{i2}, \cdot\cdot\cdot, x_{in}\}$ (and $\mathbf{z}_i = \{z_{i1}, z_{i2}, \cdot\cdot\cdot, z_{in}\}$) , for large n, these tests rapidly loose power because all finite samples are sparse in high dimensional settings owing to the "curse–of–dimensionality" (Bellman 1961).

2. MACHINE LEARNING CLASSIFICATION

The purpose of a learning machine is to predict (estimate) the unknown value of an attribute y given a set of jointly measured values $\mathbf{x}$ of other attributes. The quantity y is called the "output" or "response" variable, and $\mathbf{x} = \{x_1, \cdot\cdot\cdot, x_n\}$ are referred to as the "input" or "predictor" variables. In the binary classification problem, the response variable realizes two values, i.e. $y \in \{-1, 1\}$, respectively labeling the observations from each of two classes. The goal is to produce a model $F(\mathbf{x})$ that represents a score reflecting confidence that $y = 1$, given a set of joint values for the predictor variables $\mathbf{x}$. This score can then be used in a decision rule to obtain a corresponding prediction

$$\hat{y}(\mathbf{x}) = \begin{cases} 1 & \text{if } F(\mathbf{x}) > t^* \\ -1 & \text{otherwise.} \end{cases}$$

Here t^* is a threshold whose value is chosen to minimize the error rate.

There are a variety of ways one can go about trying to find a good predicting function $F(\mathbf{x})$. In predictive or machine learning a "training" data base $\{y_i, \mathbf{x}_i\}_{i=1}^N$ of N previously solved cases is used for which the values of all variables (response and predictors) have been jointly measured. A "learning machine" is applied to these data in order to extract (estimate) a good scoring function $F(\mathbf{x})$. There are a great many commonly used learning machines. These include linear/logistic regression, neural networks, kernel methods, decision trees, support vector machines, etc. Many are intended for use with large numbers of predictor variables. For descriptions of a wide variety of such learning procedures see Hastie, Tibshirani and Friedman 2001.

3. TWO–SAMPLE TESTING

Binary classification procedures can be used for two–sample testing. A predictor variable training data set is created by pooling the two samples

$$\{\mathbf{u}_i\}_{i=1}^{N+M} = \{\mathbf{x}_i\}_{i=1}^N \cup \{\mathbf{z}_i\}_{i=1}^M.$$

Those observations that originated from the first sample ($1 \leq i \leq N$) are assigned a response value $y_i = 1$ while those from the second sample ($N+1 \leq i \leq N+M$) are assigned $y_i = -1$. A binary classification learning machine is applied to this training data to produce a scoring function $F(\mathbf{u})$. This is then used to score each of the observations $\{s_i = F(\mathbf{u}_i)\}_{i=1}^{N+M}$.

Consider the two sets of score values $S_+ = \{s_i\}_{i=1}^N$ and $S_- = \{s_i\}_{i=N+1}^{N+M}$. These are the scores respectively assigned by the learning machine $F(\mathbf{u})$ to the first sample $\{\mathbf{x}_i\}_{i=1}^N$ and the second sample $\{\mathbf{z}_i\}_{i=1}^M$. Each of these sets of numbers $S_\pm$ can be viewed as a random sample from respective probability distributions with densities $p_+(s)$ and $p_-(s)$. Consider a *univariate* two–sample test T for the equality of these densities $p_+(s) = p_-(s)$. Let $\hat{t}$ represent the value of the corresponding test statistic

$$\hat{t} = T(\{s_i\}_{i=1}^N, \{s_i\}_{i=N+1}^{N+M}). \tag{1}$$

Examples of commonly applied univariate two–sample tests include chi–squared, Kolmogorov–Smirnov, Mann–Whitney, t–test, etc. This quantity (1) is taken to be the statistic for the *multivariate* two–sample test for the equality of the distributions of $\{\mathbf{x}_i\}_{i=1}^N$ and $\{\mathbf{z}_i\}_{i=1}^M$ $(p = q)$.

3.1. Null distribution

In order to test the "null" hypothesis $p = q$ it is necessary to know the distribution $H_0(t)$ of (1) when the hypothesis is in fact true. One rejects the null hypothesis at significance level α if the value $\hat{t}$ actually observed is greater than or equal to the $1 - \alpha$ quantile of $H_0(t)$, assuming smaller values of t represent greater likelihood of $p = q$. For commonly applied *univariate* two–sample tests the corresponding null distributions are known and have been tabulated. These distributions are valid for the multivariate application provided that separate independent data sets are respectively used for training the learning machine and evaluating the scores (1).

When the same data is used for both training and subsequent scoring, these univariate null distributions are not valid. In this case one can perform a permutation ("Fisher's exact") test. Let $\{j(i)\}_{i=1}^{N+M}$ represent a random permutation of the integers $\{1, 2, \cdot\cdot\cdot, N + M\}$. One constructs a data set $\{y_{j(i)}, \mathbf{u}_i\}_{i=1}^{N+M}$ in which the actual response values $\{y_i\}_{i=1}^{N+M}$ are randomly permuted among the predictors $\{\mathbf{u}_i\}_{i=1}^{N+M}$. These data are then used to train the learning machine, score the observations, and compute the test statistic (1). This random permutation process is repeated many (say P) times producing a set of test statistic values $\{\hat{t}_l\}_{l=1}^P$. One can then reject the null hypothesis with significance level α if the value $\hat{t}$ computed form the original data $\{y_i, \mathbf{u}_i\}_{i=1}^{N+M}$ is greater than or equal to the $1 - \alpha$ quantile of $\{\hat{t}_l\}_{l=1}^P$. This is valid for any number of random permutations P , but power increases with increasing P, reaching a diminishing return for large enough values.

4. GOODNESS–OF–FIT TESTING

As noted in Section 1, two–sample testing can be used to perform goodness–of–fit tests. One draws an artificial ("Monte Carlo") sample $\{\mathbf{z}_i\}_{i=1}^M$ from the reference distribution $q = p_0$ and tests the hypothesis $p = q$, where $p(\mathbf{x})$ is the unknown probability density of the data sample $\{\mathbf{x}_i\}_{i=1}^N$. The test is valid for any size M of the Monte Carlo sample, but power increases with increasing M, reaching a diminishing return for $M >> N$.

In two–sample testing a null distribution $H_0(t)$ is constructed by repeated random permutations of the responses $\{y_i\}_{i=1}^{N+M}$ over the predictors $\{\mathbf{u}_i\}_{i=1}^{N+M}$. This is valid for the goodness–of–fit application as well. However in the goodness–of–fit context there is an alternative method for creating a null distribution that can increase power at the expense of increased computation. One repeatedly draws many (say P) independent Monte Carlo samples of size M from the reference distribution. Each of these Monte Carlo samples $\{\mathbf{z}_i^{(l)}\}_{i=1}^M$ is used, along with the actual data $\{\mathbf{x}_i\}_{i=1}^N$, for training the learning machine and subsequent scoring to produce a test statistic value $\hat{t}_l$ from (1). This produces a set of values $\{\hat{t}_l\}_{l=1}^P$ that can be used as a null distribution to test the hypothesis $p = p_0$ in the usual manner.

The permutation procedure used with two–sample testing to construct a null distribution conditions on the observed data values $\{\mathbf{x}_i\}_{i=1}^N$ and $\{\mathbf{z}_i\}_{i=1}^M$; only information from the labels $\{y_i = \pm 1\}_{i=1}^{N+M}$ that identify the sample from which each observation originated is used. When used for goodness–of–fit testing this conditions on the values of the single Monte Carlo sample $\{\mathbf{z}_i\}_{i=1}^M$ drawn from the reference distribution $q = p_0$. Goodness–of–fit testing using repeated Monte Carlo samples as described above does not involve such conditioning and thereby uses information from the values of $\{\mathbf{z}_i\}_{i=1}^M$, as well as the labels $\{y_i = \pm 1\}_{i=1}^{N+M}$, in testing the null hypothesis. Using this additional information has the potential for increased power at the expense of having to generate many Monte Carlo samples, instead of just one.

5. DISCUSSION

As noted in the introduction, a desirable property of goodness–of–fit and two–sample tests is power against all alternatives to the null hypothesis. This will be the case provided that the chosen leaning machine is universal. That is, as the number of observations used to train it grows arbitrarily large, $N, M \to \infty$, an "optimal" scoring function $F(\mathbf{u})$ is produced that is a strictly monotone function of $\Pr(y = +1 \mid \mathbf{u})$. Some examples of universal learning machines are decision trees, neural networks, and support vector machines based on appropriate kernels. Additionally, a consistent univariate test statistic must be used in (1). That is, as $N, M \to \infty$ they will always reject the null hypothesis when $p_+(s) \neq p_-(s)$.

This notion of power against all alternatives applies in the asymptotic limit of infinite data. It has at best limited meaning in actual finite data applications. With finite data, tests based on different types of (even universal) learning machines will have differential power against different alternative distributions $p \neq p_0$ or $p \neq q$. Depending upon the actual data distribution(s) $p(\mathbf{x})$ (and $q(\mathbf{z})$) encountered in a particular application, some learning machines will have more power than others. Thus, the power of these tests can be highly sensitive to the learning machine employed. Particular choices depend on the types of potential differences between the distributions that are deemed most important to detect. For example, if the distributions tend to be different on a large fraction of the variables, near–neighbor or kernel methods will provide high power. On the other hand if they tend to differ on only a relatively small number of variables, decision trees will provide greater sensitivity.

Some multivariate two–sample tests based on near–neighbors have an advantage in that the permutation null distribution can be computed analytically. For these tests repeated learning machine training and scoring based on randomly generated permutations is not required (see Friedman and Rafsky 1979 and 1983).

In contrast to the dependence on the particular learning machine employed, the multivariate procedures described here are not likely to be very sensitive to the choice of a univariate test statistic (1).

It should be noted that as a data analytic procedure hypothesis testing extracts very little information from the data. This summary information can be encoded in a single binary bit: $b = 0/1 \Rightarrow$ accept/reject the null hypothesis. This represents a rather terse summary of a data set often consisting of many millions of bits. Furthermore, such tests will nearly always reject given enough data. Null hypotheses are seldom strictly true. It is unlikely that the hypothesized reference distribution $p_0(\mathbf{x})$, or the distribution of the second sample $q(\mathbf{z})$, will be *exactly* equal to that of the observed data $p(\mathbf{x})$. Especially if a universal learning machine is employed, enough data will detect the differences however small between them.

If the null hypothesis cannot be rejected then, at least for the size of the samples used, little additional information concerning the nature of the differences between the distributions is likely to be obtainable. However, rejection should serve as a signal to examine the data further in a attempt to extract the ways in which the distributions differ. Some learning machines such as neural networks, near–neighbor and kernel methods, and support vector machines are “black box” procedures that produce little or no interpretable information. Thus, they are not appropriate for this part of the exercise. Other methods such as decision trees are highly interpretable. For example, a decision tree produces sequences of simple inequalities (“cuts”) that identify joint values of the measured variables $\mathbf{x}$ for which $p(\mathbf{x}) >> p_0(\mathbf{x})$, $p(\mathbf{x}) << p_0(\mathbf{x})$, and $p(\mathbf{x}) \simeq p_0(\mathbf{x})$. Such information might yield considerable insight into the mechanism that produced the data.

References

[1] Bellman, R. E. (1961). *Adaptive Control Processes.* Princeton University Press.

[2] Friedman, J. and Rafsky, L. (1979). Multivariate analogs of the Wald–Wolfowitz and Smirnov two–sample tests. Annals of Statist. **7**, 697.

[3] Friedman, J. and Rafsky, L. (1983). Graph–theoretic measures of multivariate association and prediction. Annals of Statist. **11**, 377.

[4] Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning.* Springer.

Some Basic Statistical Equations for Histogram Fitting

S.I. Redin
Budker Institute of Nuclear Physics, 11 Lavrentieva Avenue, Novosibirsk 630090, Russia and Department of Physics, Yale University, New Haven, CT 06520, USA

In this paper we consider fitting of a histogram, say a time distribution, with a fit function $f(\boldsymbol{x}, t)$, where $\boldsymbol{x} \equiv \{x_i\}$ is a vector of fit parameters. We start with the equation for statistical fluctuations of fit parameters as functions of fluctuations in the number of counts in the histogram channels, $\mathcal{N}_n$, and discuss the following topics:

- statistical errors and correlations of fit parameters;
- bias of fit parameters for $\chi^2 = \sum_n \frac{(f-\mathcal{N}_n)^2}{\sigma_n^2}$ minimization and for a binned likelihood function fit; and possible improvement of χ^2 fit;
- comparison of fit parameters obtained from fitting the full set of histogram channels with those obtained from fitting some subset of channels only (set-subset relations); set-subset relations for χ^2 values;
- improvement of errors of fit parameters due to additional (external) knowledge of linear combination[s] of those parameters;
- systematic shift of fit parameters due to neglected background.

Most of the presented equations are being used for the muon (g-2) experiment [1] data analysis.

1. STATISTICAL ERRORS AND CORRELATIONS

As a result of statistical fluctuations in individual histogram channels, minimization of

$$\chi^2 \equiv \sum_n \frac{\big(f(\boldsymbol{x}, t_n) - \mathcal{N}_n\big)^2}{\sigma_n^2} = \sum_n \frac{\big(f(\boldsymbol{x}, t_n) - \mathcal{N}_n\big)^2}{f(\boldsymbol{x}, t_n)} \quad (1)$$

gives a vector of "optimal" fit parameters $\boldsymbol{x}$, shifted with respect to the "true" value $\boldsymbol{x}_\circ$ by some $\Delta\boldsymbol{x} = \boldsymbol{x} - \boldsymbol{x}_\circ$. Vector $\Delta\boldsymbol{x}$ is a function of fluctuations of the number of counts in a histogram's channels. Its elements can be derived from the χ^2 minimization requirement $\partial\chi^2/\partial x_i = 0$, which in general gives a system of nonlinear equations. The solution can be found by successive approximations $\Delta x_i = \Delta x_i^\circ + \Delta x_i^1 + \ldots$ with the leading approximation being

$$\Delta x_i^\circ = \sum_j \left(\mathcal{A}^{-1}\right)_{ij} \sum_n \frac{f'_j}{f}(\mathcal{N}_n - f) \quad (2)$$

which is of order $\frac{\mathcal{N}_n - f}{\mathcal{N}_n}$. The next-to-leading approximation Δx_i^1 is of order $\left(\frac{\mathcal{N}_n - f}{\mathcal{N}_n}\right)^2$, etc. Here

$$f'_i \equiv \frac{\partial f}{\partial x_i} \quad \text{and} \quad \mathcal{A}_{ij} = \sum_n \frac{f'_i f'_j}{f} \quad (3)$$

Important properties of fluctuations of the number of counts in a histogram's channels, $\mathcal{N}_n - f(t_n)$, are

$$\big\langle \mathcal{N}_n - f(t_n)\,\big\rangle = 0 \quad (4)$$

$$\Big\langle \big(\mathcal{N}_n - f(t_n)\big)\ \big(\mathcal{N}_m - f(t_m)\,\big)\Big\rangle = f(t_n)\,\delta_{nm} \quad (5)$$

where $\langle \ldots \rangle$ means average over an ensemble of similar histograms (ensemble average). Eqs. (2) to (5) are the basic elements in evaluation of various ensemble averages. The most fundumental is the correlation of fit parameters:

$$\begin{aligned}
&\langle \Delta x_i\ \Delta x_j\rangle \approx \langle \Delta x_i^\circ\ \Delta x_j^\circ\rangle \\
&= \sum_{ab} \left(\mathcal{A}^{-1}\right)_{ia} \left(\mathcal{A}^{-1}\right)_{jb} \sum_{nm} \frac{f'_a}{f}\frac{f'_b}{f}\Big\langle(\mathcal{N}_n - f)(\mathcal{N}_m - f)\Big\rangle \\
&= \sum_{ab} \left(\mathcal{A}^{-1}\right)_{ia} \left(\mathcal{A}^{-1}\right)_{jb} \sum_n \frac{f'_a f'_b}{f} \\
&= \sum_{ab} \left(\mathcal{A}^{-1}\right)_{ia} \left(\mathcal{A}^{-1}\right)_{jb} \mathcal{A}_{ab} = \left(\mathcal{A}^{-1}\right)_{ij} \qquad (6)
\end{aligned}$$

As a specific, but most practical, case of eq. (6), one can immediately obtain equations for the statistical errors of fit parameters:

$$\sigma_i^2 \equiv \big\langle\,(\Delta x_i)^2\big\rangle = \left(\mathcal{A}^{-1}\right)_{ii} \quad (7)$$

Equations (1) to (5) may be used to evaluate the mean (ensemble average) value of χ^2 itself:

$$\big\langle\chi^2\big\rangle = N_{ch} - L \quad (8)$$

and the mean square variation of χ^2, $\sigma_{\chi^2}^2$, with respect to $\langle\chi^2\rangle$:

$$\sigma_{\chi^2}^2 \equiv \Big\langle\big(\chi^2 - \langle\chi^2\rangle\big)^2\Big\rangle = 2N_{ch} - 2L \quad (9)$$

Here N_{ch} and L are number of histogram channels and number of fit parameters, respectively.

2. BIAS OF FIT PARAMETERS

Another important quantity is bias of fit parameters $\langle \Delta x_i \rangle$. Since ensemble averaging of Δx_i° in eq. (2) vanishes, $\langle \Delta x_i \rangle \approx \left\langle \Delta x_i^1 \right\rangle$. Ensemble averaging of Δx_i^1 for minimization of $\chi^2 \equiv \sum_n \frac{(\mathcal{N}_n - f)^2}{\sigma_n^2}$ gives:

$$\begin{aligned} \left\langle \Delta x_i^1 \right\rangle \;=\; & -\frac{1}{2} \sum_{jkl} \left(\mathcal{A}^{-1}\right)_{ij} \left(\mathcal{A}^{-1}\right)_{kl} \sum_n \frac{f_j' f_{kl}''}{f} \\ & + \xi \sum_j \left(\mathcal{A}^{-1}\right)_{ij} \times \sum_n \frac{f_j'}{f} \qquad (10) \\ & - \xi \sum_{jkl} \left(\mathcal{A}^{-1}\right)_{ij} \left(\mathcal{A}^{-1}\right)_{kl} \sum_n \frac{f_j' f_k' f_l'}{f^2} \end{aligned}$$

where $\xi = \frac{1}{2}$ for $\sigma_n^2 = f(\boldsymbol{x}_\circ, t_n)$ and $\xi = -1$ for $\sigma_n^2 = \mathcal{N}_n$. For a one parameter fit eq. (10) reads:

$$\left\langle \Delta x^1 \right\rangle = -\frac{1}{2}\, \sigma^4 \sum_n \frac{f' f''}{f} + \xi\, \sigma^2 \sum_n \frac{f'}{f} - \xi\, \sigma^2 \sum_n \frac{f'^{\,3}}{f} \quad (11)$$

The corresponding equations for the likelihood function fit are:

$$\left\langle \Delta x_i^1 \right\rangle = -\frac{1}{2} \sum_{jkl} \left(\mathcal{A}^{-1}\right)_{ij} \left(\mathcal{A}^{-1}\right)_{kl} \sum_n \frac{f_j'\, f_{kl}''}{f} \quad (12)$$

and

$$\left\langle \Delta x^1 \right\rangle = -\frac{1}{2}\, \sigma^4 \sum_n \frac{f'\, f''}{f} \quad (13)$$

which are the same as eqs. (10) and (11) with $\xi = 0$. Simple estimates show that the second term on the right side of eqs. (10), (11) in general supersedes other terms by a factor of order $N_{ch} \gg 1$. Thus the likelihood function fit, which does not contain such a term, generally has the smallest bias. The χ^2 fit, with $\sigma_n^2 = f$, has in general smaller bias (by a factor -2) than the fit with $\sigma_n^2 = \mathcal{N}_n$, though using the former is more complicated technically.

Detailed study reveals the reason for the differences in the biases as being the next-to-leading term in the Taylor expansion of likelihood and χ^2 functions in series $\left(\frac{f - \mathcal{N}_n}{\mathcal{N}_n}\right)^m$. As was shown in [2], simple modifications of the χ^2 function may reduce its bias to the level of bias for the likelihood function fit. Perhaps the simplest of such modifications is $\chi^2 = \left(\frac{(f-\mathcal{N}_n)^2}{\mathcal{N}_n} - \frac{2}{3}\frac{(f-\mathcal{N}_n)^3}{\mathcal{N}_n^2}\right)$ which is, in fact, nothing else but the first two terms of the Taylor expansion of the likelihood function (up to a factor of 2).

3. SET-SUBSET RELATIONS FOR THE χ^2 FIT

We denote $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ to be vectors of fit parameters obtained from the χ^2 minimization fit for the full set of histogram channels Ω_1, and for some subset Ω_2, respectively. Denote $\boldsymbol{x}_\circ$ to be vector of "true" values of fit parameters, common for both Ω_1 and Ω_2, and $\Delta\boldsymbol{x}_1 \equiv \boldsymbol{x}_1 - \boldsymbol{x}_\circ$, $\Delta\boldsymbol{x}_2 \equiv \boldsymbol{x}_2 - \boldsymbol{x}_\circ$. Then

$$\Delta x_{1i} \;=\; \sum_j \left(\mathcal{A}_1^{-1}\right)_{ij} \sum_{n \in \Omega_1} \frac{f_j'}{f} (\mathcal{N}_n - f) \qquad (14)$$

$$\Delta x_{2i} \;=\; \sum_k \left(\mathcal{A}_2^{-1}\right)_{ik} \sum_{m \in \Omega_2} \frac{f_k'}{f} (\mathcal{N}_m - f) \qquad (15)$$

$$\text{where } \mathcal{A}_{1ij} \;=\; \sum_{n \in \Omega_1} \frac{f_i'\, f_j'}{f} \text{ and } \mathcal{A}_{2ik} = \sum_{m \in \Omega_2} \frac{f_i'\, f_k'}{f}. \; (16)$$

The ensemble average of the difference $x_{1i} - x_{2i}$ vanishes in first approximation: $\langle x_{1i} - x_{2i} \rangle = \langle \Delta x_{1i} - \Delta x_{2i} \rangle \approx \langle \Delta x_{1i}^\circ \rangle - \langle \Delta x_{2i}^\circ \rangle = 0$. The mean square value of $x_{1i} - x_{2i}$ is:

$$\left\langle (x_{1i} - x_{2i})^2 \right\rangle = \sigma_{1i}^2 - 2\, \langle \Delta x_{1i}\, \Delta x_{2i} \rangle + \sigma_{2i}^2 \qquad (17)$$

Since Ω_2 is a subset of Ω_1, the correlation term $\langle \Delta x_{1i}\, \Delta x_{2i} \rangle$ in eq. (17) is equal to σ_{1i}^2. That follows directly from eqs. (14) to (16) and the properties of basic fluctuations given in eqs. (4) and (5). Thus

$$\left\langle (x_{1i} - x_{2i})^2 \right\rangle = \sigma_{2i}^2 - \sigma_{1i}^2 \qquad (18)$$

It is interesting to note that corresponding set-subset relation for the χ^2 values is

$$\left\langle \left(\chi_1^2 - \chi_2^2 - \left\langle \chi_1^2 - \chi_2^2 \right\rangle \right)^2 \right\rangle = 2N_{ch\,1} - 2N_{ch\,2} \quad (19)$$

which is the same as eq. (18) if one substitutes corresponding errors from eq. (9), although the real derivation of eq. (19)[1] is different and more complicated than that of eq. (18) for the fit parameters.

4. χ^2 FIT WITH INCORPORATION OF EXTERNAL LIMITED KNOWLEDGE OF FIT PARAMETERS

It might happen that some of the fit parameters are known with limited precision from other experiments, independently of our fit. If K linear combinations $F_k(\boldsymbol{x}) = \sum_{i=1} C_{ik} x_i$ $(k = 1, \ldots, K)$ are known to be

[1] We have such a derivation in [3]

$F_{*k} \pm \sigma_{F_k}$, it is reasonable to add $\sum_k \frac{(F(\boldsymbol{x}) - F_{*k})^2}{\sigma^2_{F_k}}$ to the right side of eq. (1) and use the resulting expression for the χ^2 fit. In such a case eq. (2) becomes

$$\Delta x_i = \sum_j \left(\mathcal{A}_c^{-1}\right)_{ij} \left(\sum_n \frac{f'_j}{f} (\mathcal{N}_n - f) + \sum_k \frac{C_{jk}}{\sigma^2_{F_k}} \Delta F_{*k} \right) \qquad (20)$$

where $(\mathcal{A}_c)_{ij} \equiv \mathcal{A}_{ij} + \sum_k \frac{C_{ik} C_{jk}}{\sigma^2_{F_k}}$ and $\Delta F_{*k} \equiv (F_k - F_{*k})$. The corresponding equation for the correlation matrix for such a case is

$$\langle \Delta x_i \; \Delta x_j \rangle = (\mathcal{A}_c)^{-1}_{ij} \qquad (21)$$

5. SYSTEMATIC SHIFT OF FIT PARAMETERS DUE TO NEGLECTED BACKGROUND

Suppose we have some low level background $h(t)$ admixed to the data, which is otherwise unambiguously described by a multi-parameter function $f(\boldsymbol{x}; t)$. The background might be small enough to evade observation "by eye" or even to spoil χ^2 considerably. Nevertheless fitting the histogram with the function $f(\boldsymbol{x}; t)$ alone will give parameter values x_i , shifted with respect to the "true" values $x_{i\circ}$ by some $\delta x_i = x_i - x_{i\circ}$. These systematic shifts δx_i can be found from the χ^2 minimization requirement $\partial \chi^2 / \partial x_i = 0$:

$$\begin{aligned} \frac{\partial \chi^2}{\partial x_i} &= 2 \sum_n \frac{f - \mathcal{N}_n}{f} \, f'_i \;\; \approx \; 2 \sum_{j=1}^{L} \delta x_j \sum_n \frac{f'_j \, f'_i}{f} \\ -2 \sum_n \frac{h}{f} \, f'_i \;\; &= \; 2 \sum_{j=1}^{L} \mathcal{A}_{ij} \, \delta x_j - 2 \sum_n \frac{h}{f} \, f'_i = 0 \end{aligned} \qquad (22)$$

and hence

$$\delta x_i = \sum_j \left(\mathcal{A}^{-1}\right)_{ij} \times \sum_n \frac{h}{f} \, f'_j \qquad (23)$$

with the same matrix $\mathcal{A}$ as in eq. (3). Here we use $f(\boldsymbol{x}, t_n) \approx f(\boldsymbol{x}_\circ, t_n) + \sum_j f'_j \, \delta x_j$ and $\mathcal{N}_n \approx f(\boldsymbol{x}_\circ, t_n) + h(t_n)$.

6. REPLACE SUMS BY INTEGRALS

For practical reasons, it is convenient to replace sums over the histogram's channels in the equations above by integrals in the following way:

$$\sum_n g(t_n) \;\; = \;\; \frac{1}{b} \sum_n g(t_n) \, \Delta t \;\; \approx \;\; \frac{1}{b} \int g(t) \, dt \qquad (24)$$

where $b = \Delta t$ is the width of a histogram channel and g is an arbitrary combination of the fit function and its derivatives. In turn, b can be eliminated in favor of the total number of events, N_{tot}, via the equation

$$N_{tot} = \sum_n \mathcal{N}_n \approx \sum_n f(t_n) \approx \frac{1}{b} \int f(t) \, dt \qquad (25)$$

Thus eq. (24) can be rewritten as

$$\sum_n g(t_n) \; \approx \; \frac{N_{tot}}{\int f(t) \, dt} \; \int g(t) \, dt \qquad (26)$$

Replacing sums by integrals in many cases allows one to obtain *analytical* expressions for the important statistical quantities discussed in this paper and makes their analysis easy.

7. STATISTICAL EQUATIONS FOR THE MUON g-2 EXPERIMENT

For the muon g-2 experiment [1], the time distribution of decay electrons plays the central role. It may be approximated by the 5-parameter function $G(t) = N_\circ \, e^{-t/\tau} \, [1 + A \, \cos(\omega t + \phi)]$, where A, ω and ϕ are the amplitude, frequency and phase of $g-2$ oscillations, respectively; τ is the muon lifetime in the lab frame and $N_\circ$ is the normalization constant. The value of ω, obtained from the χ^2 fit, is used to evaluate the muon $g-2$ value.

Application of statistical equations, discussed in this paper, to the 5-parameter function fit $G(t)$ gives:

- statistical errors and correlations

$$\sigma_\omega \;\; = \;\; \frac{\sqrt{2}}{\tau \, A \, \sqrt{N_{tot}}} \qquad (27)$$

$$\langle \Delta\omega \, \Delta\phi \rangle \;\; = \;\; -\frac{2}{\tau^2 \, A^2 \, N_{tot}} \; (t_s + \tau) \qquad (28)$$

where t_s is the histogram start time. For an estimated $N_{tot} \approx 10^{10}$ total events in the whole experiment, $\sigma_\omega / \omega \sim 0.3 \cdot 10^{-6} = 0.3$ ppm. Correlations of ω with other parameters, except for the phase, vanish. In fact, $\langle \Delta\omega \, \Delta\phi \rangle$ also vanishes when $t_s = -\tau$, which may be achieved by appropriate choice of the time origin. This technical trick makes calculations easy.

- The correlation of frequency ω and phase ϕ allows one to use possible external knowledge of ϕ (at time t', with error σ_F) to improve the precision of ω. This would result in

$$\sigma_\omega \;\; = \;\; \sigma_{\omega\circ} \left(1 + \frac{\sigma_F^{-2}}{(\tau \sigma_{\omega\circ})^{-2} + \sigma_F^{-2}} \, \frac{(t' - t_s - \tau)^2}{\tau^2} \right)^{-1/2} \qquad (29)$$

where $\sigma_{\omega\circ} = \frac{\sqrt{2}}{\tau A \sqrt{N_{tot}}}$ is the statistical error of ω from the χ^2 fit alone, see eq. (27).

- systematic shift due to neglected background $h(t)$:

$$\delta\omega \;\; = \;\; - \; \frac{2}{e \, N_\circ \, A \, \tau^3} \int \frac{t \, h(t) \, \sin(\omega t + \phi)}{1 + A \, \cos(\omega t + \phi)} \, dt \qquad (30)$$

For $\delta\omega$ for a particular example of $h(t)$, see [4].

• bias of fit parameters (the leading second term in eq. (10) for $\Delta\omega$):

$$\begin{aligned} \frac{\Delta\omega}{\omega} &\approx \frac{1}{2\omega}\left(\frac{\sqrt{2}}{\tau A\sqrt{N}}\right)^2 \frac{N_{ch}}{\Delta T} \\ &\times \int \frac{N_\circ e^{-t/\tau} A\, t\, \sin(\omega t+\phi)}{N_\circ e^{-t/\tau}[1+A\cos(\omega t+\phi)]}\, dt \\ &\approx \frac{N_{ch}}{\omega^2\, \tau^2\, A\, N_{tot}} \end{aligned} \tag{31}$$

where ΔT is the total range of the histogram in time units. For $N_{tot} = 10^{10}$, $\Delta\omega/\omega$ is ~ 0.1 ppb, which is negligible compare to the statistical error ~ 0.3 ppm. However if for technical or other reasons one wants to split statistics into, say, 1000 parts, fit them separately and obtain the final result by averaging those 1000, the net result might be biased by 0.0001 ppm × 1000 = 0.1 ppm, which is not negligible.

• Set-subset relations: we use "standard" eqs. (18) and (19) for histogram start time change in the course of our systematic study, but we also have derived and use a more elaborate and specific equation for similar changes of the energy threshold (E_{thr}) of decay electrons,

$$\left\langle(\omega_1-\omega_2)^2\right\rangle = \sigma^2_{\omega 2} - \sigma^2_{\omega 1}\left(2\,\frac{A_1}{A_2}\,\cos(\phi_1-\phi_2)-1\right) \tag{32}$$

where $g-2$ amplitude A and phase ϕ are functions of E_{thr}. For the case of $A_1 = A_2$ and $\phi_1 = \phi_2$, eq. (32) coincides with eq. (18). For more details see [4].

References

[1] G.W. Bennett et al., Phys. Rev. Lett. **89**, 101804 (2002), and references therein.

[2] S.I. Redin, Internal g-2 note #418, 2002, unpublished.

[3] S.I. Redin, Internal g-2 note #410, 2002, unpublished.

[4] S.I. Redin et al., Proc. of the Conf. on Advanced Statistical Techniques in Particle Physics, p. 242-247, Durham, England, 2002.

Incorporating Systematics and Statistical Uncertainties into Exclusion Limits

S.I.Bityukov
IHEP, Protvino, 141281, Russia
N.V. Krasnikov
INR RAS, Prospect 60-letiya Octyabrya 7a, Moscow, 117312, Russia

It is important to know the range in which a planned experiment can exclude the presence of a signal at a given confidence level $1-\epsilon$. We propose to use the probability of making a correct decision in future hypothesis testing about observation of a signal in planned experiments as the confidence level in the determination of exclusion limits.

1. INTRODUCTION

It is important to know the range in which a planned experiment can exclude the presence of a signal at a given confidence level $1-\epsilon$. It means that we plan to have an uncertainty not more than ϵ in our conclusion about observation or non-observation of a signal. The estimation of this uncertainty in future hypothesis testing allows the determination of exclusion limits.

Let us consider a statistical hypothesis

$$H_0\text{: } \textit{new physics is present in Nature}$$

against an alternative hypothesis

$$H_1\text{: } \textit{new physics is absent in Nature.}$$

The value of uncertainty is defined by the probability to reject H_0 when it is true (Type I error)

$$\alpha = P(reject\ H_0 | H_0\ is\ true)$$

and the probability to accept H_0 when H_1 is true (Type II error)

$$\beta = P(accept\ H_0 | H_0\ is\ false).$$

There are different approaches to the construction of exclusion regions in planned experiments [1–5]. For example, Hernandez et al. [1] propose the following criteria for the definition of exclusion limits:

$$\beta < \Delta \text{ and } \frac{\alpha}{1-\beta} < \epsilon\ ,$$

i.e. the experiment will observe with probability at least $1-\Delta$ at most a number of events such that the limit obtained at the $1-\epsilon$ confidence level excludes the corresponding signal.

In a recent note [4], the authors also propose to construct the exclusion region using two values: the magnitudes of significance level (α) and power of test $(1-\beta)$ in hypothesis testing.

We propose to construct the exclusion region using only one value: the estimator

$$\hat{\kappa} = \frac{\hat{\alpha} + \hat{\beta}}{2} \qquad (1)$$

of the uncertainty [3] $\kappa = \alpha + \beta$, when testing H_0 versus H_1 with an equal-tailed test. It is the probability of making an incorrect choice in favor of one of the hypotheses in future hypothesis testing. Here $\hat{\alpha}$ and $\hat{\beta}$ are the estimators of possible Type I error (α) and Type II error (β) obtained by direct calculations. Then $\hat{\kappa}$ is independent of which hypothesis is chosen as H_0, and which is H_1. The estimator $\hat{\kappa}$ differs from the estimator $\tilde{\kappa} = \hat{\kappa}/(1-\hat{\kappa})$ which was used in [6]. For Poisson distributions, we propose to use an equal probability test [7] as a good approximation to the equal-tailed test for estimation of $\hat{\kappa}$. Note that our approach is close to that proposed in ref [5].

Ref [8] suggests a Monte Carlo method for taking into account several types of systematics in construction of confidence limits. We consider here systematics which do not have statistical properties and hence cannot be taken into account by traditional methods for estimating their influence on exclusion limits.

2. THE PROBABILITY OF MAKING A CORRECT DECISION

Suppose that the probability of observing n events in an experiment is described by the function $f(n;\mu)$ with parameter μ, and that we know the expected numbers of signal and background events (μ_s and μ_b respectively).

Let us specify what we mean by the probability of making a correct decision about the presence or absence of a new phenomenon in a planned experiment. Let us define the criterion for the hypothesis choice and calculate the probability of making a correct decision. This is possible, because we construct the critical region in such a way that the probability of an incorrect choice in favor of one of the hypotheses is independent of whether H_0 or H_1 is true. We consider two conditional distributions of probabilities

$$\begin{cases} f_0(n) = f(n;\mu_s+\mu_b), \\ f_1(n) = f(n;\mu_b) \end{cases} \tag{2}$$

We suppose that any prior suppositions about H_0 and H_1 can be included in $f_0(n)$ and $f_1(n)$. After choosing a critical region in some way, we can estimate the Type I ($\hat{\alpha}$) and Type II errors ($\hat{\beta}$). In the case of applying the equal-tailed test, their combination Eq. 1 is the probability of making incorrect choice in favour of one of the hypotheses.

In actuality we must estimate the random value $\kappa = \alpha+\beta = \hat{\kappa}+e$, where $\hat{\kappa}$ is a constant and e is a stochastic term. α is the fraction of incorrect decisions if H_0 is true. Then β is absent because H_1 is not realised in Nature. Correspondingly, β is the fraction of incorrect decisions if H_1 takes place; then α is absent. If H_0 is true, the Type I error equals $\hat{\alpha}$ and the error of our estimator (Eq. 1) is $\hat{e} = \hat{\kappa} - \hat{\alpha} = \dfrac{\hat{\alpha}+\hat{\beta}}{2} - \hat{\alpha} = -\dfrac{\hat{\alpha}-\hat{\beta}}{2}$. Similarly, if H_1 is true, the Type II error equals $\hat{\beta}$ and the error of the estimator is $\hat{e} = \hat{\kappa} - \hat{\beta} = \dfrac{\hat{\alpha}+\hat{\beta}}{2} - \hat{\beta} = \dfrac{\hat{\alpha}-\hat{\beta}}{2}$. Thus the stochastic term takes the values $\pm\dfrac{\hat{\alpha}-\hat{\beta}}{2}$. If we require $\hat{\alpha} = \hat{\beta}$, both errors of the estimation are equal to 0 $\quad(\hat{\kappa}-\hat{\alpha} = \hat{\kappa}-\hat{\beta} = 0)$. As a result the estimator (Eq. 1) gives the probability of making an incorrect decision in future hypothesis testing.

Accordingly, $1-\hat{\kappa}$ is the probability to make a correct choice with the given critical value. Note that the equal probability test gives results close to the equal-tailed test in the case of Poisson distributions and we use an equal probability test henceforth.

Let the probability of observing n events in an experiment be described by a Poisson distribution with parameter μ, i.e.

$$f(n;\mu) = \frac{\mu^n}{n!}e^{-\mu}. \tag{3}$$

Then the Type I and II errors can be written as:

$$\begin{cases} \hat{\alpha} = \sum\limits_{i=0}^{n_c} f(i;\mu_s+\mu_b) = \sum\limits_{i=0}^{n_c} f_0(i), \\ \hat{\beta} = 1 - \sum\limits_{i=0}^{n_c} f(i;\mu_b) = 1 - \sum\limits_{i=0}^{n_c} f_1(i), \end{cases} \tag{4}$$

where n_c is a critical value.

$\hat{\kappa}$ has a minimum if we choose n_c such that $f_0(n_c) = f_1(n_c)$. (For the discrete Poisson distribution, $n_c =$ largest integer i such that $f_0(i) \leq f_1(i)$). This follows directly from

$$\hat{\kappa} = \frac{\hat{\alpha}+\hat{\beta}}{2} = \frac{1}{2}(1 - \sum_{i=0}^{n_c}(f_1(i) - f_0(i))). \tag{5}$$

The value of $\hat{\kappa}$ decreases as i increases from 0 up to n_c. As soon as $f_0(i) > f_1(i)$, the value of $\hat{\kappa}$ increases. Thus $\hat{\kappa}$ will have its minimal value when applying the equal probability test [3], and

$$n_c = [\frac{\mu_s}{ln(\mu_s+\mu_b) - ln(\mu_b)}], \tag{6}$$

where square brackets mean the integer part of a number.

3. SIGNAL SIGNIFICANCE AND EXCLUSION LIMITS

$\hat{\kappa}$ plays the role of ϵ in the definition of the confidence level and, correspondingly, of the significance S of an excess of signal events above background [10] in planned experiments. In the case of Poisson distributions the definition of significance as

$$\hat{\kappa} = \frac{1}{\sqrt{2\pi}}\int_{S_{12}}^{\infty} e^{-\frac{x^2}{2}}dx. \tag{7}$$

leads to the formula [3, 6, 7]

$$S_{12} = \sqrt{\mu_s+\mu_b} - \sqrt{\mu_b}. \tag{8}$$

A factor two is needed (i.e. $S = 2S_{12}$) to correspond with common practice. As shown in [11] this approximation has good statistical properties as the significance for Poisson distributions.

Let us define the exclusion limit of a planned experiment: *the planned experiment can exclude the presence of a signal at a given confidence level ϵ if the probability of a wrong decision about the observation of the signal will be equal to or less than ϵ.*

Thus to determine the exclusion limit μ_s we must solve equations (9) for μ_s, with given ϵ and μ_b, where n_c is determined by eqn. (6).

$$\begin{cases} \hat{\alpha} = \sum\limits_{i=0}^{n_c} f(i;\mu_s+\mu_b) = \sum\limits_{i=0}^{n_c} f_0(i), \\ \hat{\beta} = 1 - \sum\limits_{i=0}^{n_c} f(i;\mu_b) = 1 - \sum\limits_{i=0}^{n_c} f_1(i), \\ \hat{\kappa} = \dfrac{\hat{\alpha}+\hat{\beta}}{2} \leq \epsilon. \end{cases} \tag{9}$$

4. SYSTEMATICS OF THEORETICAL ORIGIN

We consider here forthcoming experiments to search for new physics. In this case we must take into account the systematic uncertainty which have theoretical origin without statistical properties. For example, two loop corrections for most reactions at present are not known. It means that we can only estimate the scale of influence of background uncertainty on the observability of signal, i.e. we can determine the admissible level of uncertainty in theoretical calculations for a given experiment proposal.

Suppose the background cross section is known to be in the interval $(\sigma_b, \sigma_b(1+\delta))$, and hence the average number of background events lies in the interval $(\mu_b, \mu_b(1+\delta))$.

As we know nothing about possible values within this range, we consider the worst case. Eqs. (9) for the exclusion limit are modified to

$$\begin{cases} n_c = [\dfrac{\mu_{lim}}{ln(\mu_{lim}+\mu_b(1+\delta)) - ln(\mu_b(1+\delta))}], \\ \hat{\alpha} = \displaystyle\sum_{i=0}^{n_c} f(i;\mu_{lim}+\mu_b(1+\delta)) \\ \hat{\beta} = 1 - \displaystyle\sum_{i=0}^{n_c} f(i;\mu_b(1+\delta)) \\ \hat{\kappa} = \dfrac{\hat{\alpha}+\hat{\beta}}{2} \leq \epsilon, \end{cases} \qquad (10)$$

where μ_{lim} is the exclusion limit at the worst background level $\mu_b \cdot (1+\delta)$.

5. CONCLUSIONS

We have considered a possible approach for the determination of exclusion limits for planned experiments. We propose to use the probability of making an incorrect decision in future hypothesis testing about the observation of the new phenomenon. We also propose an approach to take into account systematics of a theoretical origin.

Acknowledgments

We would like to thank Louis Lyons, V.A. Matveev and V.F. Obraztsov for their interest and useful comments. We are very grateful to A. Nikitenko for very fruitful discussions. We also wish to thank S.S. Bityukov, R.D. Cousins, Yu.P. Gouz, G. Kahrimanis, V.V. Smirnova, V.A. Taperechkina for discussions and E.A. Medvedeva for help in preparing the paper. S.B. is grateful to Arla LeCount and Jim Linnemann. This work has been supported by grant RFBR 03-02-16933.

References

[1] J.J.Hernandez, S.Navas and P.Rebecchi, “Estimating exclusion limits in prospective studies of searches”, Nucl.Instr.&Meth. A **378** (1996) 301.

[2] T.Tabarelli de Fatis and A.Tonazzo, “Expectation values of exclusion limits in future experiments”, Nucl.Instr.&Meth. A **403**(1998) 151.

[3] S.I. Bityukov and N.V. Krasnikov, “New physics discovery potential in future experiments,” Modern Physics Letters **A13** (1998) 3235.

[4] L. Fleysher et al., “Exclusion regions and their power”, physics/0308067, August, 2003.

[5] G. Punzi, “Sensitivity of searches for new signals and its optimization”, these Proceedings.

[6] S.I. Bityukov and N.V. Krasnikov, Proc. of Workshop on Confidence limits, eds. F. James, L. Lyons, Y. Perrin, CERN-2000-005, 2000, p. 219.

[7] S.I.Bityukov and N.V. Krasnikov, Proc. of Conf. “Advanced statistical techniques in particle physics”, eds. M.R. Whalley, L. Lyons, Durham, UK, 2002, p.77; hep-ph/0204326.

[8] J.Conrad et al., “Coverage of Confidence Intervals for Poisson Statistics in Presence of Systematic Uncertainties”, Proc. of Conf. “Advanced statistical techniques in particle physics”, eds. M.R. Whalley, L. Lyons, Durham, UK, 2002, p.58.

[9] S.I.Bityukov and N.V.Krasnikov, *On the observability of a signal above background*, Nucl.Instr.&Meth. **A452** (2000) 518.

[10] A.G.Frodesen, O.Skjeggestad and H.Toft, *Probability and Statistics in Particle Physics,* UNIVERSITETSFORLAGET, Bergen-Oslo-Tromso, 1979, p.408.

[11] V.Bartsch and G.Quast, “Expected signal observability at future experiments”, CMS Internal Note 2003/039. August, 2003.

Bayesian Separation of Independent Sources in Astrophysical Radiation Maps Using MCMC

E.E. Kuruoğlu and P.M. Comparetti
Istituto di Scienza e Tecnologie dell'Informazione, CNR, via G. Moruzzi 1, 56124 Pisa, Italy.

In this work, we present a novel approach to the recovery of independent sources of radiations in sky maps. The work is motivated by the need to resolve CMB and other specific sources of radiation from a mixture of sources in the observations that are to be made by the Planck satellite after its launch in 2007. In particular, we present a numerical approach for Bayesian estimation namely Markov Chain Monte Carlo (MCMC) that exploits available prior information about radiation sources and hence differs from most of other work in the literature which are generally blind. MCMC provides large flexibility in modelling the problem and avoids analytical difficulties by resorting to numerical techniques. Results demonstrate the success and the flexibility of the approach.

1. INTRODUCTION

ESA's Planck satellite, which is to be launched in 2007, will provide 9 all-sky maps ranging in frequency from 30 GHz to 900 GHz, and in angular resolution from 30 to 4.5 arcminutes. Celestial microwave radiation is generated by various astronomical sources, and the measured signals are superimpositions of the source signals, corrupted by measurement noise. Source signals include the cosmic microwave background (CMB), the thermal Galactic dust radiation, the synchrotron radiation (caused by the interaction of the electrons with magnetic field of the galaxy) and the free-free radiation (due to the thermal bremstrahlung from hot electrons when accelarated by ions in the interstellar gas); among which CMB is of paramount importance since it is a relic radiation remaining from the first instant light was able to travel in the universe and therefore contains the picture of the very early universe. In addition, the measurement of the anisotropies in the CMB will place fundamental constraints on models for the evolution of large scale structure in the universe. Each of the other source signals is also of interest in cosmology and astrophysics. Our goal is to reconstruct these signals.

We implement a Markov Chain Monte Carlo (MCMC) algorithm to perform Bayesian source separation, with application to the separation of signals of different origin in sky radiation maps. The problem is formulated as the separation of an instantaneous linear mixing

$$\mathbf{x} = \mathbf{A}\mathbf{s} + \mathbf{n} \tag{1}$$

where $\mathbf{s}$ contains the astrophysical sources, $\mathbf{x}$ houses the observations over various frequency channels and $\mathbf{A}$ is the mixing matrix where each row contains the frequency response of each source at a certain frequency channel.

The problem has been dealt with using other methods by several researchers in the literature including Baccigalupi et al. [2000] and Maino et al. [2002] who implemented the FastICA algorithm and its noisy version which had limited success in the presence of significant noise. A source model was introduced in Kuruoglu et al. [2003] implementing the Independent Factor Analysis (IFA) technique which also included the noise in the mixing model. Despite this added flexibility, IFA uses a fixed source model which lacks freedom in modelling source model parameters and moreover could not deal with non-stationary noise which is the case in our problem. In this work, we aim at overcoming these limitations by providing a full Bayesian analysis equipped with statistical priors for the source parameters and utilizing the prior information about the mixing as well.

Other related work include Snoussi et al. [2001] which uses an Expectation-Maximization (EM) algorithm for obtaining the mixing parameters to obtain a locally optimal maximum a posteriori estimate for the mixing matrix and other model parameters, then solves the problem of estimating the source signals given the parameter estimates, while in our method all unknowns, including the source signals, are estimated jointly, providing a globally optimal minimum mean squared error (MMSE) estimate. Miskin and MacKay [2000] use a variational Bayesian approach which uses several approximations to avoid analytical difficulties. MCMC approach, on the other hand, does not make any analytical approximations and bypasses the analytical difficulties by using a Bayesian sampling strategy.

Since the MCMC methods provide samples from the full posterior distribution, one can easily infer other functions of the parameters and their uncertainties. The great flexibility of the sampling approach allows us to make appropriate modelling choices for our problem. We can therefore relax the assumption that noise is stationary, and work under the realistic assumption that antenna noise is Gaussian but non-stationary, with a different but known variance at each pixel.

2. MARKOV CHAIN MONTE CARLO

For a detailed description of the Markov Chain Monte Carlo approach the reader is referred to Gelman et al. [1995].

In this work, we tried various sampling schemes including Metropolis-Hastings but here we present results with Gibbs sampling which we preferred because of computational convenience. In this case, the samples are generated using the conditional posterior as the sampling kernel. For assessing convergence, we generated several parallel Metropolis chains and followed both inter and intra-sequence variances. We made decisions on convergence looking at the ratio of the marginal posterior variance of the estimand which is a weighted average of inter and intra-sequence variances to the intra-sequence variance (Gelman et al. [1995]).

We assume a Gaussian mixture model for the sources. In that sense, the source model is similar to the one in Kuruoglu et al. [2003]; however, unlike that work here the model parameters are assigned prior distributions. In particular, the means of the Gaussian components are assigned Gaussian priors, the variances inverse Gamma distributions and the index parameter which chooses between Gaussian components is assigned a Dirichlet prior. The priors are chosen to be conjugate to simplify the computations.

We also exploited the prior knowledge about the mixing matrix, the elements of which are related through equations of black body radiation as detailed in Baccigalupi et al. [2000]. This information reduces the unknowns in the matrix. The remaining elements are assigned non-informative priors.

3. SIMULATION STUDIES

To test the performance of our approach, we use synthetic but realistic radiation maps of CMB generated using Seljak and Zaldarriaga [1996], and synchrotron and galactic dust extrapolated from the low resolution images obtained during the COBE mission. We use the 30 GHz, 70GHz and 143 GHz channels. The original images are mixed using the matrix

$$A = \begin{bmatrix} 1.26 & 37.0 & 0.13 \\ 1.14 & 2.91 & 0.55 \\ 0.78 & 0.34 & 1.79 \end{bmatrix}. \qquad (2)$$

Noise with an average SNRs of $12.56\mu K$ (at 30GHz), $11.307\mu K$ (at 70GHz) and $1.839\mu K$ (at 143GHz) was added to the mixtures.

The original images, the mixtures and the MMSE estimates obtained by the Gibbs-MCMC technique are given in Figure 1 where for each pixel, its marginal posterior mean value is shown. As seen from the figure, the method has been very successful in separating the signals. The success of the method in the presence of noise and its clear superiority to FastICA (Hyvarinen and Oja [1997]) can be seen in Table I where respective signal-to-interference ratios (SIR) are given.

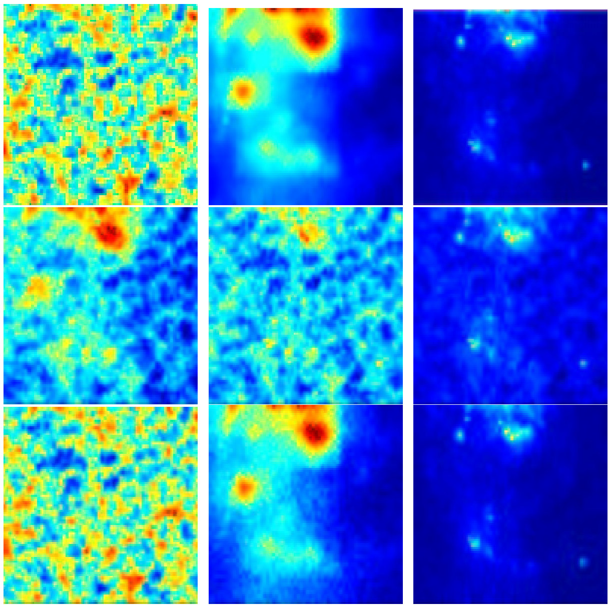


Figure 1: top row: original images of CMB, synchrotron and galactic dust, middle row: artificial mixtures over 143GHz, 70GHz and 30 GHz channels with additive noise, bottom row: estimates using MCMC

Table I Comparison of performances of separation by MCMC and FastICA by SIR of the estimates.

	CMB	Synchrotron	Dust
MCMC	19.5	18.2	27.6
FastICA	4.5	11.4	10.5

4. CONCLUSIONS

In this short contribution, we presented a numerical Bayesian approach, namely MCMC with a Gibbs sampling scheme, for the separation of the mixtures of astrophysical sources in radiation maps. The method provides a very flexible source model and includes noise in the mixing model in contrast with some previous work. It also exploits available prior information about the sources and the mixing and obtains a posterior which gives one the freedom of inference for various statistical variables of interest. The simulation results demonstrate the success of the method. Currently, we are exploring different estimators and are testing our method over different regions of the sky which show varied characteristics to provide comparisons with other existing techniques.

Acknowledgments

Authors would like to thank the Planck teams in Trieste and Bologna for providing the images.

References

C. Baccigalupi, L. Bedini, C. Burigana, G. De Zotti, A. Farusi, D. Maino, M. Maris, F. Perrotta, E. Salerno, A. Tonazzini, et al., Monthly Notices of the Royal Astronomical Society **318**, 769 (2000).

D. Maino, A. Farusi, C. Baccigalupi, F. Perrotta, A. Banday, L. Bedini, C. Burigana, G. De Zotti, K. Gorski, and E. Salerno, Monthly Notices of the Royal Astronomical Society **334**, 53 (2002).

E. Kuruoglu, L. Bedini, E. Salerno, and A. Tonazzini, Neural Networks **16**, 479 (2003).

H. Snoussi, G. Patanchon, J. Macias-Prez, A. Mohammad-Djafari, and D. J., in *AIP Proceedings of Workshop on Bayesian and Maximum-Entropy Methods* (2001), pp. 125–140.

J. Miskin and D. MacKay, in *Advances in Independent Component Analysis*, edited by M. Girolami (Springer-Verlag, 2000).

A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian Data Analysis* (Chapman & Hall, UK, 1995).

U. Seljak and M. Zaldarriaga, Astrophysical Journal **469**, 437 (1996).

A. Hyvarinen and E. Oja, Neural Computation **9**, 1483 (1997).

Using α-stable Distributions to Model the $P(D)$ Distribution of Point Sources in CMB Sky Maps

D. Herranz and E.E. Kuruoğlu
Istituto di Scienza e Tecnologie dell'Informazione. CNR, via G. Moruzzi 1, 56124 Pisa, Italy
L. Toffolatti
Departamento de Física, Universidad de Oviedo, c/ Calvo Sotelo s/n, 33007 Oviedo, Spain

We present a new approach to the statistical study and modeling of point source counts in astronomical images. The approach is based on the theory of α-stable distributions. We show that the non-Gaussian distribution of the intensity fluctuations produced by a generic point source population – whose number counts follow a simple power law – belongs to the α-stable family of distributions. With the α-stable model it is possible to totally describe the non-Gaussian distribution with a few parameters which are closely related to the parameters describing the source counts. Using statistical tools available in the signal processing literature, we show how to estimate these parameters in an easy and fast way. Then we apply the method to Cosmic Microwave Background (CMB) observations where point sources appear as superimposed to the cosmological signal as well as the instrumental noise, and propose a method to statistically disentangle these contributions. In the case of the Planck mission, our technique is able to determine the parameters of the dominant point source populations with relative errors $< 5\%$ for the 30 GHz and 857 GHz channels. The formalism and methods presented here can be useful also for other experiments in other frequency ranges such as X-rays or radio Astronomy.

1. INTRODUCTION

The study of the fluctuations in the Cosmic Microwave Background (CMB) radiation has become one of the milestones of modern cosmology not only for the relevance of the study of the CMB anisotropies in itself but also for the unique opportunity it provides for the study of the physical sources (foregrounds) that are superimposed to the CMB radiation. Among the different foregrounds that appear in CMB observations, extragalactic point sources (EPS) are specially difficult to deal with. While the brightest EPS can be individually detected in CMB sky maps, the vast majority of them are faint and remain unresolved, creating a diffuse foreground that is frequently referred as 'confusion noise'. Due to the intrinsic diversity of the galaxies that contribute to this 'confusion noise', it is impossible to establish a single spectral behavior characterizing it, thus hampering the performance of classical component separation methods that use multi-frequency observations to separate the different foregrounds. Therefore, the study of the EPS foreground is a difficult task in CMB Astronomy.

There are two traditional ways of determining the main statistical properties of the EPS population. One possibility is to detect the brightest point sources in a given data set, e.g. using a linear filter to detect them (see for example [12]), and then obtain parameters such as the number counts as a function of the observed flux. This approach is limited by the current low sensitivity of detectors at CMB frequencies. The other possibility is to directly study the statistical properties of the confusion noise distribution which, in general, is mixed with the signal coming from CMB and the other foregrounds plus instrumental noise. It is a well-known fact that the intensity distribution given by unresolved sources is strongly non-Gaussian and shows long positive tails. The statistical study of the EPS confusion noise is generally performed using statistical indicators such as the moments up to a certain degree (see, e.g., [7, 8]). Anyway, the lack of an analytical form for the probability density function (*pdf*) makes it difficult to determine the optimal statistics for this kind of studies. In particular, it is not clear which moments are necessary to characterize the distribution.

In this contribution, we will focus in the application of a relatively novel formalism, the α-stable distributions, to model the *pdf* of the intensity fluctuations due to point extragalactic sources. α-stable distributions are known to be very efficient in modeling random processes with long non-Gaussian tails. Moreover, we will show that the *pdf* of the confusion noise generated by EPS whose number counts follow a simple power law, observed with a filled-aperture instrument, must follow exactly an α-stable distribution. The great advantage is that α-stable distributions are completely described by a small number of parameters instead of an infinite number of moments. Optimal techniques already existent in the signal processing literature are easy to adapt to directly extract the main parameters of the source distribution (namely, the slope of the number counts power law and its normalization) using straightforward statistical estimators specifically designed to deal with α-stable distributions. Finally, the methods can be generalized for dealing with mixtures of signals, as is the case when the EPS population is added to CMB signal and instrumental noise.

2. SOURCE COUNTS AND $P(D)$ DISTRIBUTION

Let us consider a population of EPS whose differential number counts can be described in a power law form: $n(S) = kS^{-\eta}$, $S > 0$, where η is the *slope* of the differential counts power law, k is called its *normalization* and S is the observed flux. The sources are now observed with an instrument whose angular response is $f(\theta, \phi)$. Let us now define the *deflection* D as the fluctuation field that is observed, that is $D = I - \langle I \rangle$, where I is the intensity at a given point (time) and $\langle I \rangle$ is its average value. The characteristic function $\psi(w)$ of the deflection probability distribution $P(D)$ was studied in [1] among others and, after some straightforward calculations, can be expressed as

$$\psi(w) = \exp\left\{ i\mu w - \gamma \left| w \right|^{\alpha} \left[1 + i\beta \mathrm{sgn}(w) \tan\left(\frac{\alpha\pi}{2}\right) \right] \right\}, \tag{1}$$

where the parameters α, β, γ and μ relate to the physical parameters of the EPS distribution and of the detector through

$$\beta = \frac{1}{\pi} \Gamma\left(\frac{1+\alpha}{2}\right) \Gamma\left(\frac{1-\alpha}{2}\right) \cos\left(\frac{\alpha\pi}{2}\right) = 1, \tag{2}$$

$$\alpha = \eta - 1, \quad \gamma = \frac{\pi^{3/2} k \Omega_e}{2^{\alpha+1} \Gamma\left(\frac{\alpha+1}{2}\right) \Gamma\left(\frac{\alpha+2}{2}\right) \sin\left(\frac{\alpha\pi}{2}\right)}, \tag{3}$$

$$\mu = \frac{k\Omega_e}{1-\alpha} \lim_{a \to 0^+} a^{1-\alpha}, \tag{4}$$

and where $\Omega_e = \int \left[f(\theta, \phi) \right]^{\eta - 1} d\Omega$ is a geometrical factor called *effective beam solid angle*. The second equality in eq. (2) is due to the properties of the Γ function but we keep β as a 'variable' in eq. (1) for reasons that will be clear in the next section. The previous equations are valid for $1 < \eta < 3$ and can be obtained from eq. (8) in [1] just by rearranging terms (except for a $[2\pi]^{\alpha}$ term that corresponds to a different choice of the normalization of the beam and that is not relevant). The utility of expressing the characteristic function of the $P(D)$ in this way will be clear in the next section.

Equation (1) has an important drawback: to obtain the *pdf* of the deflections, $P(D)$, it is necessary to make the inverse Fourier transform of $\psi(w)$ which, in general, cannot be evaluated analytically. Although it can be performed numerically, the computational cost can be high if many different realizations are needed for a particular task. This has hampered the study of the $P(D)$ in the past but, as we will show in the following, it is not necessary to work with the *pdf* in all the cases. As we will show in the following sections, the study of the characteristic function itself can be insightful enough.

3. α-STABLE DISTRIBUTIONS

In section 2 we have reformulated the expression for $\psi(w)$ given by [1] so that it appears as in eqs. (1) to (4). The reason for doing so is that eq. (1) has exactly the same expression than the characteristic function of a family of distributions called in the statistical signal processing literature α-*stable* distributions. In fact, α-stable distributions are *defined* by characteristic functions such as the one in eq. (1). In this section we will very briefly overview the main properties of this kind of distributions. For a detailed description of α-stable distributions and its mathematical foundation, see [9].

The α-stable are a family of distributions that include the Gaussian distribution as a particular case. The α-stable distributions furnish tractable examples of impulsive behavior (i. e. the presence of heavy non-Gaussian tails in the *pdf*). While in a general case the full description of a non-Gaussian distribution requires the knowledge of all the cumulants of the distribution, in the case of α-stable distributions the distribution is uniquely described by means of only four parameters μ, α, β and γ. Among these parameters, α is a measure of the degree of impulsivity (non-Gaussianity) of the distribution: lower values of α corresponding to more non-Gaussian cases. The parameter β gives an idea of the asymmetry of the distribution, $\beta = 0$ corresponding to symmetric distributions whereas $\beta = \pm 1$ indicates maximum asymmetry. The parameter γ indicates the dispersion of the distribution around its maximum. The parameter μ is a simple shift in the position of the maximum.

When $\alpha = 2$, the α-stable corresponds to a Gaussian distribution with dispersion $\gamma = \sigma_g^2/2$. Then, α-stable distributions include as a particular case the Gaussian distribution, and yet they are able to describe a wider range of cases where non-Gaussianity and long tails of the distribution are present. The α-stable distributions can be shown to be the limit distribution of natural noise processes under realistic assumptions pertaining to their generation mechanism and propagation conditions ([6]).

So far we have shown that the $P(D)$ distribution, originated by an EPS power law whose number counts follow a power law, is a member of the α-stable family of distributions. Moreover, from eq. (2) we know that for this case $\beta = 1$ (i.e. the distribution shows a tail only to positive values), so we need to know only three parameters α, γ and μ to have a full statistical knowledge of the $P(D)$ distribution. As already said, μ is a simple shift in the position of the maximum and its value does not affect the shape of the $P(D)$. Therefore, the knowledge of only two statistical parameters, α and γ, gives us a complete description of the statistics of the $P(D)$ distribution. From eq. (3) we can see that the values of α and γ depend on the beam of the experiment, that is usually well-known, and the EPS number counts parameters k and η. Hence, if we are able to determine α and γ we will determine as well the properties of the EPS number counts.

The α-stable distributions have been thoroughly studied in the signal processing literature. In particular, a great effort has been devoted to design statistically optimal methods to estimate the α-stable parameters α and γ (see for example [5]). As a result of this effort, in the signal processing literature there is a plethora of available methods to perform statistical inference on α-stable environments. In this work we will focus on the application of existent techniques for α-stable parameter extraction in order to obtain optimal estimators of the parameters describing the differential number counts of the EPS population, namely the slope η and the normalization k.

3.1. Point source parameter extraction using α-stable distributions

As mentioned in the introduction, most statistical studies of the $P(D)$ have been performed in the past by calculating moments of the observed deflections and then fitting them to theoretical models. However, it can be proved that these 'classical' methods based on the study of integer order moments of the distribution (variance, skewness, etc) have bad convergence properties in α-stable environments. The intuitive reason for this is that the moments of order ≥ 2 of the $P(D)$ are not well defined (except for the case $\alpha = 2$), as can be seen from eq. (1) just by remembering that the moments of the distribution are related to the derivatives of the characteristic function through

$$i^n M_n = \left[\frac{d^n\psi}{dw^n}\right]_{w=0}, \tag{5}$$

where M_n is the moment of order n. Therefore, integer moment-based methods are not reliable as a means to learn information on the EPS population from the $P(D)$.

Fortunately, from eq. (3) the values of the EPS power law k and η can be directly obtained if we are able to estimate first the α-stable parameters α and γ. Over the past years a number of efficient estimators for the parameters of α-stable distributions have been developed. In [5] several groups of techniques to determine the parameters α and γ are described, for example the so-called *logarithmic moments estimators* and the *fractional-lower order moments estimators*. These methods exploit the fact that while integer order moments are not well defined for this kind of distribution, it is possible to calculate non-integer order moments that are well defined from the mathematical point of view. By means of the mentioned estimators, given an image of the sky containing EPS confusion noise it is easy to estimate the parameters α and γ (and conversely k and η) in an unbiased and efficient way.

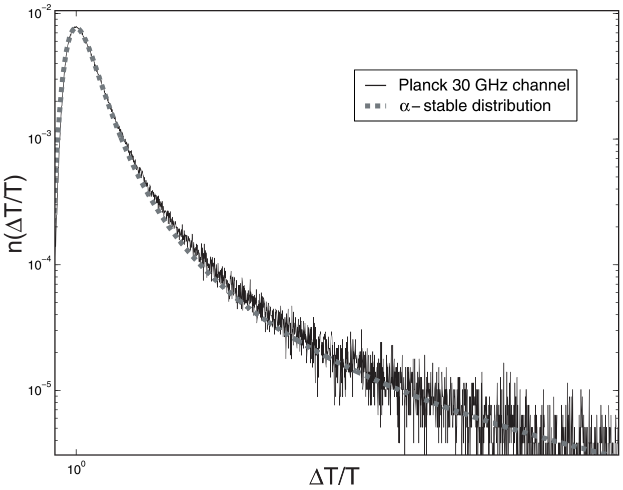


Figure 1: $P(D)$ distribution of the EPS at the 30 GHz channel (solid line) and its corresponding α-stable model (gray dotted line).

A complete description of these estimators and their performance is beyond the scope of this work, and can be found in [5]. The application of these methods to the study of EPS have been studied in profundity in [3].

When the α-stable distribution generated by the EPS is 'corrupted' with another signal (for example CMB) the problem becomes much more complicated, since the resultant distribution of the mixture is not a pure α-stable but the convolution of the α-stable *pdf* with the *pdf* of the other signals (in the case of the CMB, a Gaussian). In that case, the estimators mentioned above are not any longer optimal and it is necessary to use directly the expression for the characteristic function. As an example, let us consider the case in which the α-stable signal is mixed with a Gaussian component (as in the case of the mixture of CMB, EPS and instrumental noise). In that case, the resultant characteristic function will be similar to the one in eq. (1) but an additional term $\sigma^2 w^2/2$ will appear inside the exponential (σ^2 is the variance of the Gaussian component). It is possible to simplify the characteristic function using a simple transformation of the data called centro-symmetrization ([5]). If $\{X_i,\ i = 1,\ldots,N\}$ is the sequence of data, we can generate a centro-symmetrized sequence of data $\{X_i^S\}$ just by making $X_j^S = X_{2j} - X_{2j-1}$ for all $j = 1,\ldots,N/2$. In that case, after centro-symmetrization the new characteristic function can be proven to be:

$$\psi^S_{mix}(w) = \exp\left[-2\gamma\,|w|^\alpha - \sigma^2 w^2\right], \tag{6}$$

where σ is the dispersion of the Gaussian component, that may be either a known value or one of the parameters to estimate, and the superscript S stands for centro-symmetrization. The parameter extraction in the case of mixtures with characteristic functions such as in eq. (6) has been studied by [4]. One possibility to perform the estimation is to mimimize the distance

$$D_\Theta \equiv \int_{-\infty}^{\infty} \left| \hat{\psi}_N(w) - \psi_\Theta(w) \right|^2 W(w) dw, \qquad (7)$$

with respect to the set of parameters $\Theta = \{\alpha, \gamma, \sigma\}$, where $\hat{\psi}_N(w)$ is the empiric characteristic function $\hat{\psi}_N(w) \equiv \frac{1}{N} \sum_{m=1}^{N} e^{iwx(m)}$ and $W(w)$ is an appropriate weighting function [4]. For example, the choice $W(w) = \exp(-w^2)$ allows the integral (7) to be solved by means of Gauss-Hermite quadratures, which is computationally convenient.

4. APPLICATION TO PLANCK SIMULATED DATA

To test the ideas shortly reviewed in the last sections, we performed realistic simulations of the sky at the nine frequencies that will be covered by Planck, containing CMB, EPS and instrumental noise at the Planck expected levels. The simulations cover the whole sky and use the HEALPix (Hierarchical, Equal Area and iso-latitude) pixelisation scheme. For each channel, the resolution and the beam size correspond to the technical specifications of the mission. CMB emission has been simulated assuming a flat ΛCDM Universe with $\Omega_m = 0.3$ and $\Omega_\Lambda = 0.7$ The C_l's were generated with the CMBFAST code ([10]). The EPS simulations were done using the point source model given by [11] (hereafter, T98 model) for each one of the Planck channels. Finally, Gaussian white noise was added to each channel using the expected noise levels for Planck.

Note that the EPS in the model by [11] do not follow a pure, single power law distribution as presented in section 2. However, the power law is a good approximation for the true behaviour of the number counts over a significant range of fluxes. In [2] the authors have shown that the α-stable model is a good representation of the $P(D)$ generated T98 model in all the Planck channels, specially the lower and higher frequency ones. As an example, in fig. 1, the $P(D)$ of the EPS in the 30 GHz Planck channel is compared with an α-stable model whose parameters have been determined from the T98 simulated sky using the logarithmic estimators mentioned in section 3.1. The model fits the data almost perfectly.

Table I Results for the most significant Planck channels. The values of k are expressed in $10^{-5} \times \mathrm{Jy}^\alpha$ pixel^{-2} units. Values of σ are expressed in $10^{-5} \times \Delta T/T$ units (thermodinamic temperature). The subscript f refers to the best-fit values of the EPS simulations, whereas the subscript e makes reference to values estimated by mimizing the distance (7).

ν (GHz)	η_f	η_e	k_f	k_e	$\sigma_{CMB+n,f}$	$\sigma_{CMB+n,e}$
30	2.26	2.24	5.72	3.83	3.14	3.16
857	2.63	2.71	6.34	5.22	1940	1940

From the simulations containing the mixture of CMB, EPS and instrumental noise we tried to estimate the parameters of the EPS number counts η and k using the α-stable model and minimizing the distance (7) between the empiric characteristic function of the simulated data and the model (eq:charfmix). This gives us the estimates of the α-stable parameters α and γ that can be used in turn to estimate η and k using eq. (eq:gamma). Results are shown in table I for the two most significant Planck channels from the point of view of the EPS: the 30 GHz channel, that is dominated by flat-spectrum radio sources, and the 857 GHz channel, that is dominated by dusty far-IR galaxies. These two channels are very useful to study both kinds of EPS populations in a frequency range where their properties are not well known. Table I shows that our method is able to estimate the parameters of the number counts of the dominating EPS populations with significant accuracy. In particular the slope η is determined with relative errors lower than 5%.

5. CONCLUSIONS

In this work we have introduced the formalism of α-stable distributions as a useful tool for the statistical modelling of the intensity fluctuations due to point sources in astronomical images. We have shown that when the number counts of the sources follow a power law the characteristic function of the resultant distribution is exactly an α-stable one. The α-stable model allows us to describe the $P(D)$ distribution with a few parameters directly related to the parameters of the number counts law and to design statistically optimal and fast estimators to extract these parameters of the EPS populations from $P(D)$ distribution alone or mixed with other astrophysical sources whose *pdf* is known. We have proposed a method to extract the relevant information of the EPS population and the CMB plus noise joint variance in the Planck sky maps using the empirical characteristic function We have applied our technique to *realistic* Planck simulations containing CMB, instrumental noise and extragalactic point sources. The technique succeeds in extracting the α-stable parameters of the EPS distribution as well as the variance of the CMB plus noise contribution in the most relevant frequency channels of Planck, the 30 GHz one (at which the EPS are dominated by radio galaxies) and the 857 GHz one (dominated by dusty galaxies). The method uses all the information in the data, taking into account bright sources as well as very faint ones which contribute to the confusion noise.

The method presented here could also be applied to different fields in Astronomy, including the X-ray background and radio Astronomy. The application to these fields is now under study.

Acknowledgments

DH and LT acknowledge partial financial support from the European Community's Human Potential Programme under contract HPRN-CT-2000-00124 CMBNET. This work has used the software package HEALPix (Hierarchical, Equal Area and iso-latitude pixelisation of the sphere, http://www.eso.org/science/healpix), developed by K.M. Gorski, E. F. Hivon, B. D. Wandelt, J. Banday, F. K. Hansen and M. Bartelmann.

References

[1] Condon, J.J., 1974, ApJ, 188, 279.

[2] Herranz, D., Kuruoğlu, E.E. & Toffolatti, L., A&A submitted (astro-ph/0307114).

[3] Herranz, D., Kuruoğlu, E.E. & Toffolatti, L., 2003, *'Using α-stable distributions to model the point source population in CMB sky maps'*, Technical Report 2003-TR-16, ISTI-CNR.

[4] Ilow, J. & Hatzinakos, D., 1998, Signal Processing, 65, 199.

[5] Kuruoğlu, E.E., 2001, IEEE Trans. Signal Proc., 49, 2192.

[6] Nikias, C.L. & Shao, M., *'Signal Processing with α-Stable Distributions and Applications'*, Wiley & Sons, 1995.

[7] Pierpaoli, E., 2003, ApJ submitted (astro-ph/0301563).

[8] Rubiño-Martín, J.A. & Sunyaev, R.A., 2003, MNRAS submitted (astro-ph/0211430).

[9] Samorodnitsky, G. & Taqqu, M.S., *'Stable non-Gaussian random processes: stochastic models with infinite variance'*, New York, NY: Chapman & Hall, 1994.

[10] Seljak, U. & Zaldarriaga, M, 1996, ApJ, 469, 437.

[11] Toffolatti, L., Argüeso, F., De Zotti, G., Mazzei, P., Franceschini, A., Danese, L. and Burigana, C. 1998, MNRAS, 297, 117.

[12] Vielva, P., Martínez-González, E., Gallegos, J., Toffolatti, L. & Sanz, J. L., 2003, MNRAS, 344, 89.

Recommended Reading List

GENERAL BACKGROUND:

- Textbooks by Eadie et al, Lyons, Barlow and Cowan
- Durham Conference on 'Advanced Statistical Techniques in Particle Physics': `http://www.ippp.dur.ac.uk/Workshops/02/statistics/`
- BaBar Statistics Working Group web page: `http://www.slac.stanford.edu/BFROOT/www/Statistics/bibliography.html`
- CDF Statistics Committee web page: `http://www-cdf.fnal.gov/physics/statistics/statistics.html`
- Statistics section in 'Review of Particle Properties' by Particle Data Group
- R. Cousins, 'Why isn't every Physicist a Bayesian?' Am J Phys 63 (1995) 398
- Van Dyk, David A. (2002). 'Hierarchical Models, Data Augmentation, and Markov Chain Monte Carlo' in Statistical Challenges in Modern Astronomy III (Eds: E. Feigelson and G. J. Babu), Springer, pages 41 - 54.
- Gutti J. Babu, Eric Feigelson, (1996), Astrostatistics: Interdisciplinary Statistics

LIMITS:

- Workshop on Confidence Limits, CERN Yellow Report 2000-005, or `http://cern.web.cern.ch/CERN/Divisions/EP/Events/CLW/`
- FNAL Confidence Limits Workshop: `http://conferences.fnal.gov/cl2k/`
- Ilya Narsky, 'Expected coverage of Bayesian confidence intervals for mean of Poisson statistic in measurements with background', SMUHEP-00-03, hep-ex/0005019

BAYESIAN PRIORS:

- Kass and Wasserman, JASA 91, # 435, 1343 (1996), also as `http://www.stat.cmu.edu/www/cmu-stats/tr/tr583/tr583.html`
- Jim Linnemann's Fermilab talk in `http://conferences.fnal.gov/cl2k/copies/linnemann1.pdf`

SYSTEMATICS:

- Roger Barlow, SLUO Lectures on Statistics and Numerical Methods in HEP, Lecture 5: Systematic Errors. Available via BaBar Statistics Working Group web page: `http://www.slac.stanford.edu/BFROOT/www/Statistics/bibliography.html`
- R. D. Cousins and V. Highland, NIM A320 (1992) 331

GOODNESS OF FIT:

- Joel Heinrich 'Can the likelihood function be used to measure goodness of fit?' CDF/MEMO/BOTTOM/CDFR/5639, and available via CDF Statistics Committee web page: `http://www-cdf.fnal.gov/physics/statistics/statistics_home.html`

COMBINING RESULTS:

- W Mass Summer 01 Conference note: Combined Preliminary Results on the Mass and Width of the W Boson Measured by the LEP Experiments, LEP W Working Group, July 2001
- TGC Summer 01 Conference note: Combined Results for Electroweak Gauge Boson Couplings Measured on the LEP Experiments, LEP TGC Working Group, July 2001.

  Both notes are available from the LEP EW WG pages: `http://lepewwg.web.cern.ch/LEPEWWG/`

Committees

SCIENTIFIC COMMITTEE:

Roger Barlow

Bob Cousins

Glen Cowan

Seth Digel

Brad Efron

Jerry Friedman

Fred James

Dean Karlen

Jim Linnemann

Tom Loredo

Louis Lyons

Bill Murray

Vahe Petrosian

Frank Porter

Harrison Prosper

John Rice

Jeffrey Scargle

Peter Shawhan

Pekka Sinervo

Diego Torres

Guenther Walther

Steve Yellin

LOCAL ORGANIZING COMMITTEE:

Richard Mount

Joseph Perl

Arla LeCount

David Lee

List of Participants

Arroyo, Carlos G.	University of Massachusetts	arroyo@slac.stanford.edu
Askew, Andrew W.	Rice University	askew@physics.rice.edu
Babu, G. Jogesh	The Pennsylvania State University	babu@stat.psu.edu
Baggio, Lucio	INFN and University of Trento (Italy)	baggio@science.unitn.it
Barlow, Roger	Manchester University	Roger.Barlow@man.ac.uk
Beau, Tristan J.	PCC/APC	beau@in2p3.fr
Bellerive, Alain	Carleton University	alainb@physics.carleton.ca
Blobel, Volker H.	University of Hamburg	volker.blobel@desy.de
Bolton, Adam S.	Massachusetts Institute of Technology	bolton@mit.edu
Bonvicini, Giovanni	Wayne State University	giovanni@physics.wayne.edu
Bukin, Alexander D.	Budker Institute of Nuclear Physics	bukin@slac.stanford.edu
Cadonati, Laura	Massachusetts Institute of Technology	cadonati@ligo.mit.edu
Canelli, Florencia	University of Rochester	canelli@fnal.gov
Cardoso, Jean-Francois	CNRS/LTCI and PCC/APC	cardoso@tsi.enst.fr
Chang, George T.	Stanford University	gtchang@stanford.edu
Chen, Xin	Stanford Linear Accelerator Center	xchen@slac.stanford.edu
Connors, Alanna	Eureka Scientific	aconnors@frances.wellesley.edu
Cooke, Mark S.	University of California, Berkeley	mcooke@slac.stanford.edu
Cousins, Robert D.	University of California, Los Angeles	cousins@physics.ucla.edu
Cowan, Glen D.	Royal Holloway, University of London	g.cowan@rhul.ac.uk
Cranmer, Kyle S.	University of Wisconsin-Madison	cranmer@cern.ch
Cristinziani, Markus	Stanford Linear Accelerator Center	markus@slac.stanford.edu
de Leeuw, Jan	UCLA Dept Of Statistics	deleeuw@stat.ucla.edu
Delabrouille, Jacques	PCC - College de France & APC	delabrouille@cdf.in2p3.fr
Demortier, Luc M.	The Rockefeller University	luc@fnal.gov
Dhurandhar, Sanjeev V.	Inter University Centre for Astronomy and Astrophysics	sanjeev@iucaa.ernet.in
Diaconis, Persi	Stanford University	Persi.Diaconis@stanford.edu
Digel, Seth W.	Stanford University	digel@stanford.edu
Dikova, Smiliana D.	Institute of Astronomy Bulgarian Academy of Sciences	skydyn@bas.bg
Dorfan, Jonathan	Stanford Linear Accelerator Center	Jonathan.Dorfan@slac.stanford.edu
Dubois-Felsmann, Gregory	California Institute of Technology	gpdf@hep.caltech.edu
Edwards, Adam J.	Stanford University	aedwards@stanford.edu
Efron, Bradley	Stanford University	brad@stat.stanford.edu
Eigen, Gerald	University of Bergen	gerald@slac.stanford.edu
Feigelson, Eric	Pennsylvania State University	edf@astro.psu.edu
Feldman, Gary	Harvard University	feldman@physics.harvard.edu
Finn, Lee S.	Pennsylvania State University	LSFinn@psu.edu
Friedman, Jerome H.	Stanford University	jhf@stanford.edu

Genovese, Christopher R.	Carnegie Mellon University	genovese@cmu.edu
Giebels, Berrie	Laboratoire Leprince-Ringuet	berrie@slac.stanford.edu
Goldhaber, Alfred Scharff	C.N. Yang Institute for Theoretical Physics	goldhab@insti.physics.sunysb.edu
Goradia, Shantilal G.	Gravity Research Institute, Inc	Shantilalg1@juno.com
Graf, Norman A.	Stanford Linear Accelerator Center	Norman.Graf@slac.stanford.edu
Graham, Matthew J.	California Institute of Technology	mjg@cacr.caltech.edu
Gray, Alexander G.	Carnegie Mellon University	agray@cs.cmu.edu
Groom, Don	Lawrence Berkeley National Lab	deg@lbl.gov
Heinrich, Joel G.	University of Pennsylvania	heinrich@hep.upenn.edu
Hill, Gary C.	University of Wisconsin, Madison	ghill@amanda.wisc.edu
Irwin, John	Stanford Linear Accelerator Center	irwin@slac.stanford.edu
James, Frederick	CERN	f.james@cern.ch
Javerdin, Moutete F.		moutetej@yahoo.com
Jennings, Kristofer	Purdue University	jennings@stat.purdue.edu
Karlen, Dean	University of Victoria	karlen@uvic.ca
Kim, Peter	Stanford Linear Accelerator Center	pck@slac.stanford.edu
Kinoshita, Kay	University of Cincinnati	kayk@physics.uc.edu
Kleijn, Bas	UC Berkeley, Statistics Dept.	kleijn@stat.berkeley.edu
Knuteson, Bruce O.	Massachusetts Institute of Technology	knuteson@mit.edu
Lazzarini, Albert	California Institute of Technology	lazz@ligo.caltech.edu
Lee, Sang-Joon	Rice University	sangjoon@rice.edu
Levi, Ofer	Stanford University	levi@sccm.stanford.edu
Levit, Creon	NASA Ames Research Center	creon@nas.nasa.gov
Linnemann, James T.	Michigan State University	linnemann@pa.msu.edu
Litchfield, Peter J.	Minnesota University	pjl@physics.umn.edu
Lopez, Angel M.	University of Puerto Rico	angel@charma.uprm.edu
Loredo, Thomas J.	Cornell University	loredo@astro.cornell.edu
Loudin, Jim D.	Rice University	blgjimee@rice.edu
Lu, Minghui	University of Oregon	lum@rpi.edu
Lyons, Louis	Oxford University	l.lyons@physics.ox.ac.uk
Madgwick, Darren S.	University of California, Berkeley	dsm@astron.berkeley.edu
Mahabal, Ashish	California Institute of Technology	aam@astro.caltech.edu
Majewski, Stephanie	Stanford University	majewski@stanford.edu
McLachlan, Charles I.	University of Cambridge	cim20@cam.ac.uk
Miquel, Ramon	Lawrence Berkeley National Laboratory	rmiquel@lbl.gov
Mount, Richard P.	Stanford Linear Accelerator Center	richard.mount@slac.stanford.edu
Murray, William J.	CLRC - RAL	w.murray@rl.ac.uk
Mutter, Andreas	Freiburg University	andreas.mutter@physik.uni-freiburg.de
Narsky, Ilya V.	California Institute of Technology	narsky@hep.caltech.edu
Newman, David S.	*Retired*	dsnewman@ix.netcom.com
Nichol, Robert	Carnegie Mellon University	nichol@cmu.edu
Nolan, Patrick L.	Stanford University	Patrick.Nolan@stanford.edu
Pia, Maria Grazia	INFN Genova	Maria.Grazia.Pia@cern.ch
Porter, Frank C.	California Institute of Technology	fcp@hep.caltech.edu
Prevot, Isabelle	SODERN	isabelle_prevot@sodern.fr
Prosper, Harrison B.	Florida State University	harry@hep.fsu.edu
Punzi, Giovanni	Scuola Normale Superiore/INFN-Pisa	giovanni.punzi@pi.infn.it
Quayle, William B.	University of Wisconsin-Madison	quayle@wisconsin.cern.ch

Raja, Rajendran	Fermi National Accelerator Laboratory	raja@fnal.gov
Redin, Sergei I.	Budker Institute of Nuclear Physics	redin@inp.nsk.su
Reid, Nancy	University of Toronto	reid@utstat.utoronto.ca
Ribon, Alberto	CERN	Alberto.Ribon@cern.ch
Rice, John A.	University of California, Berkeley	rice@stat.berkeley.edu
Rochester, Leon S.	Stanford Linear Accelerator Center	lsrea@slac.stanford.edu
Roe, Byron P.	University of Michigan	byronroe@umich.edu
Rolke, Wolfgang A.	University of Puerto Rico - Mayaguez	wolfgang@puerto-rico.net
Roodman, Aaron	Stanford Linear Accelerator Center	roodman@slac.stanford.edu
Santos, Daniel A.	UFRJ	alencar@ccard.com.br
Satpathy, Asish	University of Texas at Austin	satpathy@slac.stanford.edu
Scannapieco, Evan	Kavli Institute for Theoretical Physics	evan@arcetri.astro.it
Scargle, Jeffrey D.	NASA Ames Research Center	jeffrey@cosmic.arc.nasa.gov
Sharma, Ajay		ajay2244@hotmail.com
Shawhan, Peter S.	California Institute of Technology	shawhan_p@ligo.caltech.edu
Shmakova, Marina	Stanford Linear Accelerator Center	shmakova@slac.stanford.edu
Siemiginowska, Aneta L.	CFA/SAO	asiemiginowska@cfa.harvard.edu
Sina, Ramin	University of Maryland	sina@pbar.umd.edu
Sinervo, Pekka K.	University of Toronto	pekka.sinervo@utoronto.ca
Snyder, Arthur E.	Stanford Linear Accelerator Center	snyder@slac.stanford.edu
Stark, Philip B.	University of California, Berkeley	stark@stat.berkeley.edu
Stump, Daniel R.	Michigan State University	stump@pa.msu.edu
Sturrock, Peter A.	Stanford University	sturrock@stanford.edu
Tcheng, Cédric	SODERN	cedric_tcheng@sodern.fr
Terranova, Francesco	Istituto Nazionale di Fisica Nucleare	francesco.terranova@cern.ch
Thiessen, Henry A.	Los Alamos National Laboratory	hat@lanl.gov
van Dyk, David A.	University of California, Irvine	vandyk@stat.harvard.edu
Verde, Licia	University of Pennsylvania	lverde@astro.princeton.edu
Vilalta, Ricardo	University of Houston	vilalta@cs.uh.edu
Wagner, Stephen R.	Stanford Linear Accelerator Center	stevew@slac.stanford.edu
Walther, Guenther	Stanford University	walther@stat.stanford.edu
Wandelt, Benjamin D.	University of Illinois at Urbana-Champaign	bwandelt@uiuc.edu
Woodroofe, Michael B.	University of Michigan	michaelw@umich.edu
Yabsley, Bruce D.	Virginia Tech	yabsley@bmail.kek.jp
Yellin, Steven J.	Stanford University	yellin@slac.stanford.edu
Zarb Adami, Kristian	University of Cambridge	kz202@mrao.cam.ac.uk
Zech, Gunter	Universitaet Siegen	zech@physik.uni-siegen.de
Zyla, Piotr	Lawrence Berkeley National Laboratory	PAZyla@lbl.gov