CERN-2011-006
27 September 2011

ORGANISATION EUROPÉENNE POUR LA RECHERCHE NUCLÉAIRE
CERN EUROPEAN ORGANIZATION FOR NUCLEAR RESEARCH

Proceedings of the PHYSTAT 2011 Workshop on Statistical Issues Related to Discovery Claims in Search Experiments and Unfolding

CERN, Geneva, Switzerland 17–20 January 2011

Editors: Harrison. B. Prosper (Florida State University)
Louis Lyons (Imperial College and Oxford)

ISBN 978–92–9083–367–3
ISSN 0007–8328

This report should be cited as:
Proceedings of the PHYSTAT 2011 Workshop on Statistical Issues Related to Discovery Claims in Search Experiments and Unfolding, CERN, Geneva, Switzerland, 17–20 January 2011, edited by H.B. Prosper and L. Lyons, CERN–2011–006

A contribution in this report should be cited as:
[Author name(s)], in Proceedings of the PHYSTAT 2011 Workshop on Statistical Issues Related to Discovery Claims in Search Experiments and Unfolding, CERN, Geneva, Switzerland, 17–20 January 2011, edited by H.B. Prosper and L. Lyons, CERN–2011–006, pp. [first page]–[last page]

Abstract

These Proceedings contain written versions of talks given at the PHYSTAT 2011 Workshop, which was held at CERN, on *Statistical issues related to discovery claims in search experiments* and *Unfolding*. The talks, which include several by professional statisticians, varied from general overviews to talks that focused on issues that arise in specific experimental contexts. In addition to the sessions on discovery claims, the PHYSTAT Workshop included a one-day workshop on *Unfolding*. Written versions of the *Unfolding* talks are also included. These Proceedings should be of interest to particle physicists as well as scientists in other fields.

The articles also appear on the web-site:
`http://indico.cern.ch/conferenceOtherViews.py?view=standard&confId=107747`

Preface

The PHYSTAT 2011 Workshop, which took place at CERN in January 2011, was the latest of a series of meetings addressing statistical issues that arise in the analysis of data in High Energy Physics (HEP) and in related fields. It was held almost exactly 11 years after the first meeting of this series, also at CERN, on Confidence Limits. The current meeting was in two parts. The first 3 days concentrated on problems that arise in analyses searching for new phenomena, which sometimes result in discovery claims, but more usually set limits on non-observed effects. These topics reprise those from the PHYSTAT-LHC meeting at CERN in June 2007, and which had been timed to address the statistical issues shortly before real data from the LHC became available. Now that the LHC is operating, and with the prospect of the discovery of new physics, such issues are even more relevant. These Proceedings include a report on a smaller Workshop held at Banff in July 2010 on these topics. The last day of the PHYSTAT 2011 meeting was devoted to the topic of unfolding detector effects from observed distributions, which was the first time this topic received serious attention at a PHYSTAT meeting.

About 160 people attended the meeting, the majority being experimental physicists. Because of the focus on a single topic for each part of the meeting, it was felt inappropriate to have parallel sessions. An important aspect of the meeting was the time between the sessions when vigorous discussions took place. The CERN cafeteria was the scene of much statistical debate.

There are many people who deserve my warmest thanks for making the meeting such a success. First and foremost are my fellow organisers Michelangelo Mangano and Albert de Roeck. Without them, this meeting simply would not have happened. In particular, they were responsible for all the local arrangements, from funding to food. Anastasia Dolya was a fantastic secretary. To her we are all grateful, and to Michelle Connor who provided very helpful support.

The presence of professional statisticians added enormously to the value of the meeting. We especially appreciated the invited talks given by several of them and their active participation throughout the meeting and in the discussions. Their patience in explaining the subtleties of their subject to non-statisticians was much appreciated.

Many thanks to all our speakers, both for the invited talks and the contributed ones. It was clear that considerable effort was involved in preparing them. With such a full set of talks, it was important that the speakers kept to time – thanks to the Chairmen[1] for ensuring that this happened. Harrison Prosper and I are also grateful for the care which went into producing the write-ups of the talks for the Proceedings, especially for the one that arrived before the deadline for contributions. We also thank the members of the Committee who reviewed the talks that appear in these Proceedings. The Committee members were also helpful both at the planning stage of PHYSTAT 2011, and during the meeting itself.

Harrison Prosper deserves a very big 'Thank you' for taking on the large and onerous task of producing these Proceedings. We are all most grateful to him for devoting a significant part of his Sabbatical to doing this.

CERN very kindly provided the funding for this meeting, and also many of the facilities. We were pleased to have the meeting here again.

[1] This word is strictly correct, as all the chairpersons were male.

Finally best wishes to everyone for every success with their analyses, and for many exciting discoveries in the near future.

Louis Lyons
Blackett Lab., Imperial College, London SW7 2BW, UK
and *Particle Physics, Oxford OX1 3RH*
e-mail: l.lyons@physics.ox.ac.uk

Search and Discovery

Unfolding

Appendix

Search and Discovery

Open issues in the wake of Banff 2010

Luc Demortier
The Rockefeller University, New York, NY 10065, USA

Abstract
We review, in some cases very succinctly, statistical issues in the formulation of discovery procedures for high energy physics. This includes alternatives to p-value tests, the look-elsewhere effect, measurement sensitivity, implicit statistical models, parton density uncertainties, reference priors, profile likelihood methods, and extreme value theory.

1 Introduction

From 11 to 16 July 2010, a group of statisticians and physicists met at the Banff International Research Station in the Canadian Rockies to debate statistical issues related to the significance of discovery claims. Although these discussions did not lead to a miraculous consensus on how to claim or not claim discoveries, progress was made in understanding some questions and in learning about potentially useful statistical techniques that are not yet known in the high energy physics community. Section 2 starts with a critical look at the way we quantify evidence against a given hypothesis, and why the almost exclusive use of p-values in our field may not be optimal. Section 3 discusses how to report a failure to discover, and the importance of measurement sensitivity for this. Difficulties arising from likelihood functions that cannot be written down analytically are explored in section 4, and the question of parton density uncertainties is summarized in section 5. Finally, some technical advances in profile likelihood techniques and reference priors are briefly described in sections 6 and 7 respectively, and the potential usefulness of extreme value theory is mentioned in section 8.

2 Discovery claims

Discovery claims in high energy physics are almost universally based on p-value calculations, regardless of the type of hypothesis that is being tested. Equally universal is the discovery threshold, which is set at five standard deviations, corresponding to a Type-I error rate of 2.87×10^{-7}. This threshold was chosen a long time ago [1], based on a back-of-the-envelope estimate of the probability of a false discovery claim in the vast number of histograms examined by all high energy physicists in the course of one year. Since then, statisticians have given ample warning that the evidence contained in a dataset for or against a given hypothesis depends strongly on the type of hypothesis being tested, on the formulation of alternatives, on sample size, on the dimensionality of the problem, and on the stopping rule. Thus it may be time to question the universality of high energy physics procedures in this regard, or at least to explore alternatives. As it turns out, these alternatives can produce results that are quite different from those obtained with p-values.

Another reason to explore alternatives is that p-values are easily misinterpreted, if not by the physicists who produce them, then almost certainly by the public at large. The two most common misinterpretations are that a p-value represents the posterior probability of a hypothesis, or the odds against it, in light of the data. Since these concepts of posterior probability and odds actually belong to the Bayesian paradigm, it is natural to turn to the latter in a search for alternative testing methods. To contrast p-values with Bayesian measures of evidence we start with a well-known paradox formulated by Lindley in 1957 [2].

2.1 Lindley's paradox

Suppose first that we have n measurements $X_1, X_2, \ldots, X_n$ distributed according to a Gaussian with unknown mean μ and known variance σ^2. The likelihood can be reduced to:

$$\mathcal{L}(\mu) \;=\; \frac{e^{-\frac{1}{2}\left(\frac{\bar{x}_o-\mu}{\sigma/\sqrt{n}}\right)^2}}{\sqrt{2\pi}\,\sigma/\sqrt{n}}, \tag{1}$$

where $\bar{x}_o$ is the observed average of all the measurements. We are interested in testing

$$H_0:\ \mu=\mu_0 \quad \text{versus} \quad H_1:\ \mu=\mu_1, \tag{2}$$

with $\mu_1 > \mu_0$. A sufficient test statistic is the average $\bar{X}$, and large values of $\bar{X}$ indicate deviation from H_0 in the direction of H_1. The p-value is therefore:

$$p_0 \;=\; \mathbb{P}\big[\bar{X} \geq \bar{x}_o \big| H_0\big] \;=\; 1-\Phi(z_o), \tag{3}$$

where $\Phi(z)$ is the cumulative standard normal distribution and $z_o \equiv (\bar{x}_o-\mu_0)/(\sigma/\sqrt{n})$ is the number of standard deviations away from H_0. For a Bayesian analysis we must first assign prior probabilities π_0 to H_0 and π_1 to H_1, with $\pi_0+\pi_1=1$. A typical non-informative choice is $\pi_0=\pi_1=1/2$, but the argument works for any value $\pi_0>0$. The posterior probability of H_0 is:

$$p(H_0\,|\,\vec{x}) \;=\; \frac{\pi_0\,\mathcal{L}(\mu_0)}{\pi_0\,\mathcal{L}(\mu_0)+\pi_1\,\mathcal{L}(\mu_1)} \;=\; \left[1+\frac{\pi_1}{\pi_0}\,e^{\left(\frac{\mu_1-\mu_0}{\sigma/\sqrt{n}}\right)\left(\frac{\bar{x}_o-(\mu_0+\mu_1)/2}{\sigma/\sqrt{n}}\right)}\right]^{-1}. \tag{4}$$

Note how this posterior couples the measurement sensitivity, $(\mu_1-\mu_0)/(\sigma/\sqrt{n})$, with the evidence contained in the data, $z_{\rm Bayes} \equiv [\bar{x}_o-(\mu_1+\mu_0)/2]/(\sigma/\sqrt{n})$. If either quantity is zero, the posterior probability of H_0 reduces to π_0. Rewriting the posterior in terms of z_o,

$$p(H_0\,|\,\vec{x}) \;=\; \left[1+\frac{\pi_1}{\pi_0}\,e^{-\frac{1}{2}\left(\frac{\mu_1-\mu_0}{\sigma/\sqrt{n}}\right)^2+\left(\frac{\mu_1-\mu_0}{\sigma/\sqrt{n}}\right)z_o}\right]^{-1}, \tag{5}$$

shows that for a fixed p-value p_0 (or equivalently, a fixed z_o value), the posterior probability of H_0 goes to 1 as the sample size n increases. With α the Type-I error rate, it could happen that a frequentist finds $p_0<\alpha$ and rejects H_0, whereas a Bayesian concludes that the evidence in the data supports H_0. The reason for this discrepancy is clear: evidence in the p-value sense is measured by z_o, which only takes H_0 into account, whereas evidence in the Bayes sense is measured by $z_{\rm Bayes}$, which takes both H_0 and H_1 into account. As the measurement resolution $\sigma/\sqrt{n}$ improves, the only way to keep z_o fixed is to increase the number of standard deviations between the data $\bar{x}_o$ and H_1. Eventually the Bayesian evidence will favor H_0.

Within the Neyman-Pearson theory of testing, the alternative hypothesis H_1 influences the test via the Type-II error rate β, the probability of incorrectly rejecting H_1. As the sample size increases, keeping α fixed allows β to become arbitrarily small, thereby shifting the emphasis from protecting H_0 (the usual goal of an experimenter) to protecting H_1. This can be avoided by letting α decrease as the sample size increases.

What happens if we remove the advantage the Bayesian approach draws from looking at a *precise* alternative hypothesis? Suppose we replace test (2) by:

$$H_0:\ \mu=\mu_0 \quad \text{versus} \quad H_1':\ \mu>\mu_0. \tag{6}$$

Here the alternative hypothesis is completely vague: the lack of focus on H_1 that characterizes the p-value has been incorporated in H_1'. This is the situation examined by Lindley in his famous paradox. A p-value analysis of this test yields the same result as before, namely Eq. (3). On the other hand,

a Bayesian analysis requires that, in addition to the hypothesis priors π_0 and π_1, we specify a prior distribution $g(\mu)$ for μ under H_1'. The actual form of $g(\mu)$ does not matter much because we will be taking the limit $n \to \infty$. We only assume that $g(\mu)$ is continuous and integrates to 1 over $\mu > \mu_0$. At large values of n and for positive z_o the posterior probability of H_0 is then given by:

$$p(H_0 \,|\, \vec{x}) \;=\; \frac{\pi_0 \,\mathcal{L}(\mu_0)}{\pi_0 \,\mathcal{L}(\mu_0) + \pi_1 \int_{\mu>\mu_0} \mathcal{L}(\mu)\, g(\mu)\, d\mu} \;\simeq\; \left[1 + \frac{\pi_1}{\pi_0}\, \frac{\sqrt{2\pi}\,\sigma}{\sqrt{n}}\, e^{\frac{z_o^2}{2}}\, g(\mu_0)\, \Phi(z_o)\right]^{-1}. \qquad (7)$$

Thus we find again that for a fixed p-value evidence z_o against H_0, the posterior probability of H_0 goes to 1 at large n, hence the paradox. A striking aspect of this paradox is that it arises in the large-sample limit, where the Bayesian and frequentist paradigms often agree in problems of point and interval estimation.

2.2 Resolution

The statistics literature on Lindley's paradox is extensive, and many resolutions have been proposed [3]. A recurring theme in this literature is that the choice of test procedure should depend on one's *prior* beliefs in the hypotheses being tested. For test (6) one can imagine three possibilities:

1. due to past experience or compelling theoretical arguments, there is a concentration of prior belief on H_0;
2. H_0 is not particularly believable, but represents a valuable simplification of our description of the physics process under study;
3. H_0 is not particularly believable, but is stated for convenience (e.g. the hypothesis we are really interested in is $\mu \leq \mu_0$, but $\mu = \mu_0$ is easier to analyze).

When searching for new physics, test (6) is rather common, with μ representing, for example, the production rate of a new particle. Our prior beliefs regarding H_0 and H_1 can then be characterized as follows:

- Even though the physical theory underlying H_0 (the standard model of particle physics) describes a vast body of previous observations extremely well, we know that it is incomplete, and that somewhere it predicts something that will not be observed. Fundamentally the theory is wrong.
- However, we do not know where the breakdown will occur. There are many predictions that can be tested. Furthermore, if the test at hand should be the one to detect a breakdown, there may be more than one physics explanation that could incorporate the alternative hypothesis. It is also possible that the correct physics explanation hasn't been formulated yet.

Which of the three prior belief structures does this situation correspond to? If we leave aside the third case (misspecification of H_0), it could be argued that we have strong belief in the (limited) validity of the standard model (case 1), or that we only view the standard model as a useful simplification of a more fundamental theory (case 2). Each of these views receives its own treatment within the Bayesian paradigm and leads to further insights into Lindley's paradox. It is also possible to accomodate both views in a single treatment.

2.2.1 Case 1: the null hypothesis enjoys strong prior belief

An important insight here is that it is very rare that one tests a true point null hypothesis. Even if the theoretical hypothesis is a point (e.g. the production rate of the Higgs boson is exactly zero because the Higgs boson does not exist), there are always unknown measurement biases that cause the actually tested hypothesis to be "fuzzy". Without arguing this point in detail, it is relatively easy to see how it leads to a resolution of Lindley's paradox [4].

Suppose that by $H_0 : \mu = \mu_0$ we really mean to approximate the hypothesis $H_0' : \mu_0 \leq \mu \leq \mu_0 + \epsilon$ for some small positive ϵ that describes the unknown biases. The test is therefore:

$$H_0' : \mu_0 \leq \mu \leq \mu_0 + \epsilon \quad \text{versus} \quad H_1'' : \mu > \mu_0 + \epsilon, \tag{8}$$

and the p-value is:

$$p_0' \;=\; \sup_{\mu_0 \leq \mu \leq \mu_0 + \epsilon} \mathbb{P}\big[\bar{X} \geq \bar{x}_o \big| H_0'\big] \;=\; \sup_{\mu_0 \leq \mu \leq \mu_0 + \epsilon} \left[1 - \Phi\Big(\frac{\bar{x}_o - \mu}{\sigma/\sqrt{n}}\Big)\right] \;=\; 1 - \Phi\Big(\frac{\bar{x}_o - \mu_0 - \epsilon}{\sigma/\sqrt{n}}\Big). \tag{9}$$

For the Bayesian analysis we suppose that there is a continuous, proper prior $\pi(\mu)$ that peaks inside H_0', such that $\pi_0 = \int_{H_0'} \pi(\mu)\, d\mu$. The posterior probability of H_0' is:

$$p(H_0' \,|\, \vec{x}_o) \;=\; \frac{\int_{\mu_0}^{\mu_0 + \epsilon} \mathcal{L}(\mu)\, \pi(\mu)\, d\mu}{\int_{-\infty}^{+\infty} \mathcal{L}(\mu)\, \pi(\mu)\, d\mu}. \tag{10}$$

At large enough n the likelihood $\mathcal{L}(\mu)$ concentrates around $\bar{x}_o$. Solving equation (9) for $\bar{x}_o$ yields:

$$\bar{x}_o \;=\; \mu_0 + \epsilon + \frac{\sigma}{\sqrt{n}} \Phi^{-1}(1 - p_0'). \tag{11}$$

Hence for fixed p_0' the likelihood concentrates at the edge of H_0' as n becomes large. For a smooth prior $\pi(\mu)$ the numerator of posterior (10) can therefore be approximated by:

$$\int_{\mu_0}^{\mu_0 + \epsilon} \mathcal{L}(\mu)\, \pi(\mu)\, d\mu \;\simeq\; \pi(\mu_0 + \epsilon) \left[\Phi\Big(\frac{\mu_0 + \epsilon - \bar{x}_o}{\sigma/\sqrt{n}}\Big) - \Phi\Big(\frac{\mu_0 - \bar{x}_o}{\sigma/\sqrt{n}}\Big)\right] \;\simeq\; \pi(\mu_0 + \epsilon)\, p_0', \tag{12}$$

where the approximation is valid in the limit where n goes to infinity while p_0' remains constant. A similar calculation for the denominator of (10) yields $\pi(\mu_0 + \epsilon)$. Taking the ratio, we find that $p(H_0' \,|\, \vec{x}_o) \to p_0'$, thus resolving the paradox.

Both Lindley's paradox and the above resolution are formulated in the large-sample limit. However, in problems of practical interest it is rare that one is able to specify ϵ, and in finite samples it is not possible to determine how close the p-value will be to the posterior probability of the null hypothesis H_0 without knowing the prior $\pi(\mu)$. Unfortunately it is notoriously difficult to construct objective priors for testing a precise hypothesis against a vague one (as in equation 6). The problem is that objective priors often tend to be improper. To circumvent this problem, ref. [4] studies lower bounds on Bayes factors and posterior probabilities over wide classes of proper priors. The surprising result is that even these lower bounds are significantly larger than the corresponding p-values, indicating that the latter overestimate the evidence against the null hypothesis. Furthermore, p-values cannot be "recalibrated" for a variety of reasons: the calibration would depend on the sample size, on the postulated probability density of the observations, on the stopping rule of the experiment, and on the type of null hypothesis being tested.

2.2.2 Case 2: the null hypothesis provides a useful simplification

Again we consider test (6), but this time we assume that, although belief in H_0 is not particularly high, this hypothesis embodies a *useful* simplification of the theory that describes the observations [5]. In other words, μ_0 is special in terms of a utility function rather than in terms of prior belief. Let $u(d_i, \mu)$ be the utility of choosing d_i when μ is the value of the parameter of interest, where d_i represents the decision to accept H_i. It seems reasonable to require that the gain in the utility of accepting H_1 be an increasing function of the distance $\delta(\mu, \mu_0)$ between μ and μ_0. For simplicity we set:

$$u(d_1, \mu) - u(d_0, \mu) \;=\; \delta(\mu, \mu_0) - \delta_0, \tag{13}$$

where δ_0 is a constant, which can be interpreted as a penalty for using the more complicated model implied by H_1 when the simpler H_0 would suffice (since $u(d_1, \mu_0) = u(d_0, \mu_0) - \delta_0$). Rejecting H_0 is the optimal decision when it leads to an expected gain in utility:

$$\mathbb{E}\big[u(d_1, \mu) - u(d_0, \mu) \,\big|\, \vec{x}\big] \;>\; 0 \quad \text{or} \quad U(\vec{x}) \;\equiv\; \mathbb{E}\big[\delta(\mu, \mu_0) \,\big|\, \vec{x}\big] \;>\; \delta_0, \tag{14}$$

where the expectation is taken with respect to the posterior distribution of μ. For the Gaussian model (1) used in Lindley's paradox, an appropriate choice of δ is the Mahalanobis distance

$$\delta(\mu, \mu_0) \;=\; \Big(\frac{\mu - \mu_0}{\sigma}\Big)^2, \tag{15}$$

and an appropriate prior for μ is the reference prior, which in this case is the indicator function of the set $\mu \geq \mu_0$. The posterior expected utility can then be written as:

$$U(\vec{x}) \;=\; \int_{\mu_0}^{+\infty} \frac{e^{\frac{1}{2}\left(\frac{\bar{x}_o - \mu}{\sigma/\sqrt{n}}\right)^2}}{\sqrt{2\pi}\,(\sigma/\sqrt{n})\,\Phi(z_0/\sqrt{2})} \Big(\frac{\mu - \mu_0}{\sigma}\Big)^2 d\mu \;=\; \frac{1}{n}\bigg[1 + z_o^2 + \frac{z_o\, e^{-z_o^2/2}}{\sqrt{2\pi}\,\Phi(z_o)}\bigg]. \tag{16}$$

Since one rejects H_0 whenever $U(\vec{x}) > \delta_0$, and since $U(\vec{x})$ is a one-to-one function of z_o, it is possible to choose δ_0 so as to make this procedure identical to the p-value test $p_0 < \alpha$. However, if we consider the situation in Lindley's paradox, where n is increased while z_o stays constant, agreement between the two procedures for a fixed penalty δ_0 can only be achieved if α decreases with n. Thus we are led to the same conclusion that we obtained by considering the Type-II error rate β of test (2).

Note that in this utility based approach it is perfectly possible to use objective priors, even if they are improper. It is also possible to put a finite prior weight on the null hypothesis, thereby obtaining a treatment that mixes the first two cases in our description of possible belief structures for test (6). Further details on this methodology can be found in ref. [6].

2.3 Application to the look-elsewhere effect

To illustrate the discrepancy between p-value and Bayesian measures of evidence, we briefly consider the problem of searching for a resonance peak somewhere in a spectrum of finite width. Since the location of the peak is not known a priori, the significance of an interesting local excess must be corrected for the fact that a background fluctuation like the observation could have occurred *anywhere* in the spectrum. This is the look-elsewhere effect (LEE).

The statistician R.B. Davies computed the LEE correction to p-values in 1987 [7]. Suppose that for each value of the resonance location $\theta \in [A, B]$, the test statistic $S(\theta)$ is (asymptotically) chisquared with s degrees of freedom. Davies derived the following formula for the LEE-corrected tail probability:

$$\mathbb{P}\left[\sup_{A \leq \theta \leq B} S(\theta) \;>\; u\right] \leq \mathbb{P}(\chi^2_s > u) \;+\; \langle N(u) \rangle, \tag{17}$$

where $\langle N(u) \rangle$ is the expected number of upcrossings of the level u by the process $S(\theta)$. LEE-corrected p-values are typically obtained via Monte Carlo simulation, which can be very time consuming for large values of u. Ref. [8] solves this problem by providing an analytical formula for the scaling of $\langle N(u) \rangle$ with u. The computation can then be considerably shortened by performing it at some low value of u and using the formula to extrapolate to the observed value.

For a simple example that doesn't require the full generality of Davies's result, consider the spectrum of observed Poisson counts shown in the left panel of Fig. 1. We assume that the background noise is the same in all bins, and that any signal can only appear in one bin. The p-value in any given bin i is

$$p(n_{o,i}) \;=\; \sum_{n=n_{o,i}}^{\infty} \frac{\mu^n}{n!}\, e^{-\mu}, \tag{18}$$

where μ is the background level and $n_{o,i}$ is the observed count in bin i. We are interested in the most significant effect, as identified by the smallest p-value in the spectrum, say $p_{\min}$. If the total number of bins examined is B, the LEE-corrected significance is:

$$p_{LEE} \;=\; \mathbb{P}\Big[\min_{1\le i\le B} p(n_{o,i}) \le p_{\min} \,\Big|\, H_0\Big] \;=\; 1-(1-p_{\min})^B. \tag{19}$$

Note that p_{LEE} is larger than $p_{\min}$.

The Bayesian calculation starts with the likelihood function:

$$\mathcal{L}(\eta,\ell) \;=\; \prod_{i=1}^{B} \frac{[\mu+\eta\,\delta_{i\ell}]^{n_{o,i}}}{n_{o,i}!}\, e^{-\mu-\eta\,\delta_{i\ell}}, \tag{20}$$

where η is the signal magnitude and ℓ its bin number. We wish to test

$$H_0 : \eta = 0 \quad \text{versus} \quad H_1 : \eta > 0. \tag{21}$$

The problem has one nuisance parameter, the signal location ℓ. In the absence of any information about ℓ, we take its prior to be uniform: $\pi(\ell) = 1/B$. If the value of η was specified under H_1, the posterior probability of H_0 would be:

$$\begin{aligned} p(H_0\,|\,\vec{n}_o)_{LEE} \;&=\; \frac{\pi_0 \sum_\ell \mathcal{L}(0,\ell)\,\pi(\ell)}{\pi_0 \sum_\ell \mathcal{L}(0,\ell)\,\pi(\ell) + (1-\pi_0)\sum_\ell \mathcal{L}(\eta,\ell)\,\pi(\ell)} \\ &=\; \left[1+\frac{1-\pi_0}{\pi_0}\frac{1}{B}\sum_{l=1}^{B}\Big(1+\frac{\eta}{\mu}\Big)^{n_{o,\ell}}\, e^{-\eta}\right]^{-1}, \end{aligned} \tag{22}$$

where π_0 is the prior probability of H_0. How can we handle the fact that η is actually *not* specified under H_1? The preferred option is a subjective Bayesian analysis: introduce a proper prior for η under H_1 and integrate it out. A second option is to do an objective Bayesian analysis by constructing a "neutral" prior for η; however this prior needs to be proper, otherwise the posterior probability of H_0 will be undefined. Methods for doing this are described in ref. [9]. A third option is the utility-based approach of section 2.2.2. Finally, one could simply plot $p(H_0\,|\,\vec{n}_o)_{LEE}$ from equation (22) as a function of η to get a sense of the variation of the Bayesian evidence regarding H_0. This is shown in the right panel of Fig. 1 for $\pi_0 = 1/2$. It is quite remarkable that, *even at its minimum,* the posterior probability of H_0 is still about an order of magnitude higher than the p-value. Of course one could reduce this discrepancy by lowering π_0, but this would mean that a substantial fraction of the evidence against H_0 is due to one's prior opinion about H_0 rather than to the data.

The plot also shows the effect of the LEE correction on both the p-value and the posterior probability. For this particular example, the effect is about the same on both quantities.

3 Measurement sensitivity

So far we have concentrated our attention on the interpretation of evidence supporting discovery claims. For tests such as (6), where the alternative hypothesis specifies a *range* of values for the parameter of interest, an equally important and difficult issue is what to report when no discovery can be claimed. If the p-value (3) is greater than the significance threshold α, we accept H_0. However, this does not mean that all values of μ under H_1' are now rejected: there are values of μ that our experiment is not sensitive to, and others that the data won't allow us to exclude. One way to investigate this is to test individual values of μ under H_1':

$$H_1'[\mu_1] :\; \mu = \mu_1 \quad \text{versus} \quad H_0 :\; \mu = \mu_0, \tag{23}$$

where, as before, $\mu_1 > \mu_0$. A $(1-\gamma)$ C.L. upper limit μ_u can then be defined as the largest value of μ_1 that is not rejected by the test at some significance level γ.

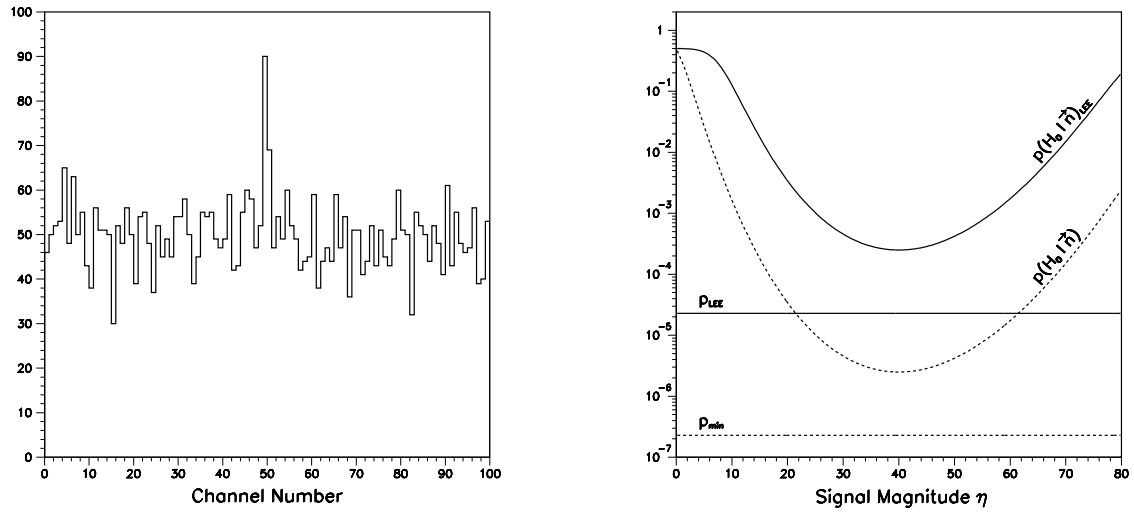


Fig. 1: Left: spectrum of Poisson counts used to illustrate the look-elsewhere effect on p values and posterior probabilities. Right: posterior probability of the background-only hypothesis as a function of the tested signal magnitude η, with and without LEE correction, compared with the corresponding p values.

In the frequentist approach to testing, μ_u can be obtained by solving the p-value equation $p_1(\mu_u) = \gamma$, where

$$p_1(\mu_1) \;=\; \mathbb{P}\big[\bar{X} < \bar{x}_o \,\big|\, H_1'[\mu_1]\big] \tag{24}$$

is the p-value for testing $H_1'[\mu_1]$. For the Gaussian likelihood (1), the upper limit derived this way is given by:

$$\mu_u \;=\; \bar{x}_o + \frac{\sigma}{\sqrt{n}}\,\Phi^{-1}(1-\gamma). \tag{25}$$

Due to measurement resolution effects it may happen that $\bar{x}_o$ is such that the upper limit μ_u falls below the lower boundary μ_0 of the physical parameter space. In this case the upper limit is unphysical and the corresponding interval is empty: all values of μ are excluded, regardless of the actual measurement sensitivity.

This problem has been known for at least twenty-five years [10]. As recently emphasized by Bob Cousins, the underlying issue is lack of conditioning in the standard frequentist approach, which, in the presence of physical boundaries, yields what is known in the statistics literature as "relevant subsets" [11]. These are subsets of sample space with respect to which the conditional coverage of a confidence interval procedure is consistently above or consistently below the nominal coverage for *all* parameter values.

Many solutions have been proposed over the years. Here we only mention those that are based on solid statistical principles. The first one is to calculate a Bayesian upper limit: the resulting intervals are never empty, but they require the choice of a prior and typically do not achieve exact frequentist coverage. The second solution is to do a frequentist construction with a so-called "unified" ordering rule, such as the likelihood-ratio ordering rule of ref. [12]. This procedure has coverage and never yields empty intervals, but there are cases where the behaviour of interval length as a function of the observations is unsatisfactory. In addition, it only accomodates one confidence level where high energy physicists typically require three: one for the discovery significance ($1.0 - 2.87 \times 10^{-7}$), one for the upper limit (95%) reported in the absence of a discovery claim, and one for the two-sided interval (68%) reported with a discovery claim. A third possibility is to modify the statistical model of the measurement, in particular its error structure [13]. For the Gaussian example, one typically assumes that the standard

deviation is known exactly and is independent of the mean, neither of which may be true. Finally, some astrophysicists have recently proposed to keep reporting the standard frequentist upper limit, but to complement it with a minimum sensitivity bound, defined as the smallest parameter value that one would have a pre-specified probability of detecting at a pre-specified level of significance if it was the true value [14]. As indicated by its definition, the construction of this sensitivity bound requires two pre-specified numbers; in addition, the handling of nuisance parameters is not trivial.

There is at present no consensus on the optimal method.

4 Implicit statistical models

High energy physics measurements are complex in the sense that we typically do not know the exact analytical dependence of the likelihood function on some parameters of the model. All we have is the underlying stochastic mechanism, which we can simulate with a Monte Carlo algorithm. This difficulty occurs for both nuisance and interest parameters.

As illustration, consider the measurement of the mass μ of a new particle. The data sample consists of a signal component (events containing the new particle) and an irreducible background component. If we have an event by event estimator X of μ, the likelihood has the form:

$$\mathcal{L}(\mu) \;=\; \prod_{i=1}^{N} \Big[(1-\epsilon_b)\, f_s(x_i\,;\,\mu) \;+\; \epsilon_b\, f_b(x_i) \Big] \;\times \ldots, \tag{26}$$

where ϵ_b is the background contamination of the sample, and f_s and f_b are the signal and background distributions of X. These distributions are usually approximated by histograms from Monte Carlo simulations, which may be smoothed or fitted with parametric representations. In addition, the f_s distribution must be constructed on a grid of μ values supplemented with interpolation. This is inefficient since a lot of time is wasted modeling $f_s(x\,;\,\mu)$ at μ values far from the maximum-likelihood estimate (MLE). Finally, f_s and f_b also depend on nuisance parameters such as energy scales, initial and final state radiation, parton densities, etc. Generalizing the above approach to multiple parameters quickly becomes impractical [15].

Over the years, a number of ingenious but somewhat dubious shortcuts were invented by high energy physicists to take nuisance parameters into account. An example shortcut is to evaluate the shift $\Delta\mu$ in the MLE of μ induced by a one-sigma variation of a given nuisance parameter, and then to replace the likelihood by its convolution with a Gaussian with standard deviation $\Delta\mu$:

$$\mathcal{L}(\mu) \;\to\; \tilde{\mathcal{L}}(\mu) \;\equiv\; \int \mathcal{L}(\mu') \, \frac{e^{-\frac{1}{2}\left(\frac{\mu-\mu'}{\Delta\mu}\right)^2}}{\sqrt{2\pi}\,\Delta\mu} \, d\mu' \tag{27}$$

When there is more than one nuisance parameter, $\Delta\mu$ is replaced by the sum in quadrature of the individual shifts. The validity of this method has never been studied in detail.

In the next two subsections we examine approaches, one Bayesian and the other frequentist, that may be useful for handling implicit models.

4.1 Approximate Bayesian computation methods

In the Bayesian paradigm, the likelihood is integrated over the nuisance parameters, a feature that lends itself well to Monte Carlo computations. Implicit statistical models can be analyzed with the help of so-called ABC methods (Approximate Bayesian Computation). The goal is to *approximate* the posterior distribution $\pi(\mu \,|\, x) \propto p(x \,|\, \mu)\, \pi(\mu)$. All we need is a suitable distance function $d(x_a, x_b)$ between two datasets x_a and x_b. Let x_{obs} be the observed dataset. The simplest ABC algorithm is the ABC rejection sampler:

1. Sample $\mu^\star$ from $\pi(\mu)$.
2. Simulate a dataset $x^\star$ from $p(x \mid \mu^\star)$.
3. If $d(x_{obs}, x^\star) \leq \epsilon$, accept $\mu^\star$, otherwise reject.
4. Return to step 1.

The output of an ABC algorithm is a sample of parameters $\mu^\star$ from a distribution $\pi(\mu \mid d(x_{obs}, x^\star) \leq \epsilon)$. If ϵ is sufficiently small, this distribution will be a good approximation to the posterior $\pi(\mu \mid x_{obs})$. A delicate issue is the choice of distance function $d(x_a, x_b)$. There is no general theory for this, and the choice must be made on a case-by-case basis.

There exist other ABC algorithms, which are more efficient than the rejection sampler and even work with improper priors [16].

When combining the results from different experiments, common uncertainties and the resulting correlations must be taken into account. This seems doable with ABC methods, although the generation of Monte Carlo samples (an industry in itself) will have to be carefully coordinated between experiments.

4.2 Decision-Theoretic Methods

In the frequentist paradigm one is interested in procedures that have coverage for all values of the interest and nuisance parameters. Other requirements besides coverage are needed to specify unique procedures.

For the construction of confidence intervals, one approach, based on decision-theoretic ideas, is known as minimax expected size (MES): it minimizes the maximum expected size of the confidence set over parameter space. In a Monte Carlo implementation of MES, parameter values are drawn at random from the parameter space, and a dataset is simulated for each parameter value. Each simulated dataset is compared to the observed dataset using a likelihood ratio test. Inverting the likelihood ratio test minimizes the probability of including false values in the confidence region, which in turn minimizes the expected size of the confidence region. This Monte Carlo algorithm does not require explicit knowledge of the likelihood function, only of the data generating mechanism [17]. In addition, it is well suited for handling physical boundaries in parameter space.

At present the Bayesian approach via ABC methods seems a lot more flexible than the above frequentist method, since ABC methods produce an approximation to the posterior itself. The decision-theoretic procedure only produces confidence intervals, and only of the MES type (no choice of ordering rule).

5 Parton Density Function Uncertainties

Currently the parton densities are determined by a fit to ~ 35 datasets with a total of ~ 3000 data points. The standard parametrization uses ~ 25 parameters, and the fit quality is characterized by a χ^2 value. Uncertainties on the parton densities are derived from a $\Delta\chi^2$ procedure, but the standard $\Delta\chi^2 = 1$ rule yields clearly unrealistic uncertainties. Instead, 90% C.L. uncertainties are obtained via $\Delta\chi^2 = 100$ or 50, depending on the group doing the fit.

These uncertainties are not yet understood from a statistical point of view. Some suggestions were made at Banff to improve this situation:

- A decision-theoretic approach such as MES.
  This may be of value for quantifying the uncertainty in the pdf estimates.
- A random effects model.
  Assume that the theory does not quite fit each experiment, resulting in underestimated prediction errors. Propose as solution that the theory parameter is slightly different in each experiment, and all these individual parameters are constrained to the formal parameter of the theory via some distributional assumptions (such as a multivariate-t prior).

– A closure test.
First verify that for data generated from the theoretical distributions, the $\Delta\chi^2 = 1$ criterion yields reasonable uncertainties. Then study how inferences are affected by biasses in theory and/or data.

6 Profile Likelihood Methods

Using results due to Wilks and Wald, ref. [18] derives a comprehensive set of asymptotic formulae, based on the profile likelihood, for use in searches for new physics.

An interesting technique introduced in that paper is the so-called Asimov dataset, which is in a sense the most representative dataset of an ensemble: when one uses it to evaluate the estimators for all parameters, one obtains the true parameter values. Asimov datasets can be used to simplify the estimation of measurement sensitivities and to compute Jeffreys' prior.

7 Reference Priors

Uniform priors have been the norm in high energy physics for a long time, partly because they *seem* reasonable (by the principle of indifference), and partly because the corresponding posterior intervals sometimes exhibit reasonable frequentist behaviour. However, they are also known to suffer from two major drawbacks: they give inconsistent results if the parametrization of the problem is changed, and they are not guaranteed to yield proper posteriors.

Reference priors have been developed over the past thirty years with the aim of providing a "standard" for presenting and comparing measurements of quantities about which little or no prior knowledge is available. Similarly to other standards (e.g. lengths and weights), the reference prior standard was designed with some rational considerations in mind: the algorithm is based on information theory and is very generally applicable; reference posteriors are invariant under one-to-one transformations of the parameter of interest, have good frequentist coverage properties, and avoid the so-called marginalization paradoxes that plague other non-informative constructions. In high energy physics, reference priors are now available for cross section measurements, when partial information is available for acceptances and background sources [19].

There are still some important issues however:

1. Can reference posterior inferences be reported by themselves, or should they be reported only as part of a sensitivity analysis? If the latter, how should one choose alternative priors?
2. The general definition of reference priors involves the taking of limits, and this must be done carefully in order to avoid infinities. The standard approach is to use sequences of nested compact sets that converge to the whole parameter space. Unfortunately there is no unique way of choosing these compact sets, and there is no guarantee that different choices lead to the same result, or even that all choices lead to a proper posterior. This ambiguity prevents us from designing a completely general numerical algorithm.
3. How should we handle implicit statistical models? Can we combine ABC methods with numerical algorithms for computing reference posteriors?

8 Extreme value theory

Let $X_1, X_2, X_3, \ldots$ be independent and identically distributed random variables. Whereas central limit theory is concerned with the behavior of the partial sums $X_1 + X_2 + \ldots + X_n$ as $n \to \infty$, extreme value theory studies the behavior of the sample extremes $\max\{X_1, X_2, \ldots, X_n\}$ as $n \to \infty$. This theory has many applications, for example to the question of how high dikes should be built in the Netherlands to protect land below sea level from storm surges that drive the seawater level up along the coast.

In high energy physics we are often interested in extreme events, that is, collision events in which some measurable quantity takes on a very large value. Extreme value theory may help here, by providing a solid basis for extrapolating from measurements at lower values of the quantity of interest [20].

9 Acknowledgements

I wish to thank the organizers and participants of both the 2010 Banff meeting and the 2011 Phystat workshop for many interesting talks and productive discussions.

References

[1] A. H. Rosenfeld, "Are there any far-out mesons or baryons?," in *Meson spectroscopy: a collection of articles*, C. Baltay and A. H. Rosenfeld, eds., W. A. Benjamin, Inc., New York, Amsterdam, 1968, pg. 455.

[2] D. V. Lindley, "A statistical paradox," Biometrika **44**, 187 (1957).

[3] See for example G. Shafer, "Lindley's paradox," (with discussion) J. Amer. Statist. Assoc. **77**, 325 (1982).

[4] J. O. Berger and M. Delampady, "Testing precise hypotheses," Statist. Sci. **2**, 317 (1987); `http://www.stat.duke.edu/~berger/papers/p-values.pdf`.

[5] M. J. Bayarri, "Comment," Statist. Sci. **2**, 342 (1987).

[6] J. Bernardo, these proceedings. Note that Bernardo uses a loss function instead of our utility function; one is simply the negative of the other.

[7] R. B. Davies, "Hypothesis testing when a nuisance parameter is present only under the alternative," Biometrika **74**, 33 (1987).

[8] E. Gross and O. Vitells, "Trial factors for the look-elsewhere effect in high energy physics," Eur. Phys. J. C**70**, 525 (2010); arXiv:1005.1891v3 [physics.data-an].

[9] J. O. Berger, "A comparison of testing methodologies," *PHYSTAT LHC Proceedings on "Statistical issues for LHC physics,"* CERN Yellow Report CERN-2008-001 (7 March 2008), pg. 8.

[10] V. L. Highland, "Estimation of upper limits from experimental data," Temple University preprint C00-3539-38 (1987).

[11] R. J. Buehler, "Some validity criteria for statistical inferences," Ann. Math. Statist. **30**, 845 (1959).

[12] G. J. Feldman and R. D. Cousins, "Unified approach to the classical statistical analysis of small signals," Phys. Rev. D **57**, 3873 (1998).

[13] G. Casella, "Comment," Statist. Sci. **17**, 159 (2002).

[14] V. L. Kashyap *et al.*, "On computing upper limits to source intensities," Astrophysical J. **719**, 900 (2010); arXiv:1006.4334v1 [astro-ph.IM]. Caveat: the terminology of this paper interchanges the concepts of *upper limit* and *upper bound* as understood in high energy physics.

[15] P.J. Diggle and R.J. Gratton, "Monte Carlo methods of inference for implicit statistical models," J. R. Statist. Soc. B**46**, 193 (1984).

[16] T. Toni *et al.*, "Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems," J. R. Soc. Interface **6**, 187 (2009).

[17] C. M. Schafer and P. B. Stark, "Constructing confidence regions of optimal expected size," J. Amer. Statist. Assoc. **104**, 1080 (2009); `http://www.stat.cmu.edu/~cschafer/cmspbs.pdf`.

[18] G. Cowan *et al.*, "Asymptotic formulae for likelihood-based tests of new physics," arXiv:1007.1727v2 [physics.data-an] (2010).

[19] L. Demortier, H. B. Prosper, and S. Jain, "Reference priors for high energy physics," Phys. Rev. D **82**, 034002 (2010).

[20] L. de Haan and A. Ferreira, "Extreme value theory: an introduction," Springer (2006).

Discovery: A Statistical Perspective

David R. Cox
Nuffield College, Oxford, OX1 1NF, UK

Abstract
A general statistical formulation of discovery problems is sketched and important distinctions drawn. Procedures for checking for the existence of a signal are analyzed from different viewpoints.

1 Introduction

Discovery is taken to mean finding and verifying a rare signal against a noisy background. There are many variants and formulation is critical. Key elements are the following:

- the reference frame in which signals are defined
- the statistical properties of the noise
- the temporal sequence of data collection
- the statistical character of the signal
- the frequency of occurrence of signals
- the multi-stage character of the search process

Thus the reference framework may be a set of bins, the order of which is essentially ignored, an ordered set of histogram bins, or a one or higher dimensional continuum, corresponding to energy, time, spectral frequency or to one or more spatial dimensions. Another possibility is the linked use of two or more reference frames. Thus a weak signal when data are ordered by energy level and a weak signal when ordered by some other feature might, under some circumstances, become a strong signal if it could be shown that the same originating events were involved in each case.

Typical noise processes are either Gaussian processes or Poisson processes and it may be important to allow for errors in estimating their properties. Conventional statistical thinking would, if there is extensive observation of the background, tend to rely on the empirical variability of the background rather than on *a priori* assumptions about its form. This would, for example, cover the possibilities of over- or under-dispersion relative to the Poisson distribution.

Observation may be in one step or by the gradual accretion of frequencies over time.

The signal may be a single blip, or a set of occurrences at nearby points in the reference frame.

A distinction crucial to the following discussion is between two main situations. First there may be no signal present or just one. In the second situation there are likely to be a limited but nonzero number of signals present and the challenge is to find as many as possible of them with few false alarms. The former seems the version more appropriate for current issues in particle physics and therefore we largely concentrate on that.

Examples range from particle physics to genetic epidemiology, especially GEWAS (genome-wide association studies), isolation of faint signals in complex spectra, and aspects of drug development and of plant breeding programmes. All have distinctive features.

2 Simplest Formulation

We start with the simplest formulation. At each of a large number n of sites a one-sided test of significance yields the set of p-values $p_1, \ldots, p_n$. For simplicity we assume the underlying test distributions to be essentially continuous and moreover the different tests to be statistically independent.

This is an explicitly significance test based formulation in which only a null hypothesis is specified formally, together with a criterion judged sensitive to relevant departures. There is no statistically formulated model of the signal, so that no estimation question arises. Of course richer formulations allow richer solutions.

Schweder and Spjotvoll (1982) discussed the analysis and interpretation of the empirical distribution of the p-values. The analysis could be done in terms of an equivalent standard normal variable, $t = \Phi^{-1}(1-p)$, where $\Phi(.)$ denotes the standard normal cumulative distribution function, or in terms of the transform $z = -\log p$. In a sense all are equivalent, but there are substantial advantages to the last because it places the region of interest prominently in large values, rather than crowded near zero for p and in the upper tail of a normal distribution for t. Use of z or a near equivalent appears a favoured method in analysing GEWAS, following Wellcome Trust Case Control Consortium (2007).

For more careful analysis and for dealing with generalizations it is helpful to use the Renyi decomposition. Under the global null hypothesis, that is that there is no signal present, the random variables $(z_1, \ldots, z_n)$ are independently exponentially distributed with unit mean. The Renyi decomposition is that the corresponding ordered values $(z_{(1)} \leq z_{(2)} \leq \ldots \leq z_{(n)})$ can be represented in terms of a another set of independent unit exponential variables $(v_1, v_2, \ldots, v_n)$ in the form

$$z_{(1)} = v_1/n, z_{(2)} = v_1/n + v_2/(n-1), \ldots, z_{(n)} = v_1/n + v_2/(n-1) + \cdots + v_n.) \tag{1}$$

That is, as the ordered values evolve in sequence they form a nonstationary random walk with exponentially distributed steps.

Two immediate consequences are first that the ordered z may be plotted against the expected values of the decomposition, Eq. (1), namely $1/n, 1/n + 1/(n-1), \ldots, 1/n + 1/(n-1) + \ldots + 1$. Interest lies in the upper end of the plot. Secondly the largest value, $z_{\max} = z_{(n)}$ is such that its associated significance level is

$$P(z_{\max} \geq z) = 1 - (1 - e^{-z})^n$$

and this is to a close approximation

$$1 - \exp(-ne^{-z}). \tag{2}$$

This is the standard allowance for selecting the most significant of many test statistics and the second form shows virtual equivalence to using the Gumbel distribution of extreme value theory for $z_{\max}$.

3 Extensions

The most immediate use of the above results is to show the assembled p-values graphically, in particular leading to evidence for an isolated signal if $z_{\max}$ is large. We now sketch a number of extensions and modifications.

Efron in series of papers, synthesized and extended in a monograph (Efron, 2010), has emphasized that the distribution of the test statistic under the real subject-matter null hypothesis may not be that specified by statistical theory. Indeed if interest lies in the 5σ region proposed in the particle physics context it is hardly plausible that the probabilistic interpretation associated with the Gaussian distribution will hold quantitatively. The implication for testing for a single extreme point is that $z_{\max}$ should be compared not with its expected value from the exponential distribution but with an extrapolation from previous values in the plot.

More formally if the p's are independent and identically distributed but not uniformly distributed the Renyi decomposition applies to a nonlinear function of the v's so that the plot mentioned above should be a smooth curve and if the upper section can be treated as locally linear the largest order statistics should have the form

$$\eta_0 + \eta_1\{v_1/n + v_2/(n-1) + \cdots + v_{n-k+1}/(n-k)\},$$

where η_1 is the local slope of the plot at large values.

This can be estimated from

$$\hat{\eta}_1 = \{2z_{n-1} + z_{n-2} + \ldots + z_{n-k} - (k+1)z_{n-k-1}\}/k.$$

Here the effective slope of the plot is estimated bearing in mind the special structure of the order statistics as a random walk. Then $(z_n - z_{n-1})/\hat{\eta}_1$ has under the null hypothesis the standard variance ratio or F distribution with degrees of freedom $(2, 2k)$. Choice of k will in general be based on inspection of the plot.

Next if the z are based on histogram bins, choice of bin size may be problematic. One resolution is to take a number of bin sizes $h, 2h, \ldots$ and to take the largest z, denoted again by $z_{\max}$. This will again have a Gumbel distribution, that is leading to a significance level of

$$1 - \exp(-n^* e^{-z_{\max}}). \tag{3}$$

Here n^* is an effective sample size, intermediate between the number of small bins and the total number of bins, the latter corresponding to the Bonferroni bound. The constant n^* could perhaps be calculated theoretically but would probably best be found by simulation; note that finding a single constant by simulation is much easier than studying the extreme tails of a distribution!

A further possibility is that the test statistics have some correlation structure, for example that of a stationary time series. Then the z-plot may be nonlinear with a nonunit slope and $z_{\max}$ will have a Gumbel distribution, Eq. (3), that is with a modified effective sample size, n^{**}, where in extreme value theory n/n^{**} is termed the extremal index. Such correlation might arise from the noise process. Another way would be if the signal is expected to be spread over a number of bins suggesting the use of a statistic, in Gaussian form, in the one sided-case of

$$S_m = T_m + aT_{m-1} + a^2 T_{m-2} + \ldots.$$

Here T_m is a standard Gaussian test statistic from bin m. After dividing S_m by its standard error it can be converted to z-form and then plotted as before.

Thus in summary the use of $z = -\log p$ provides a graphical analysis and a range of test procedures adaptable to various situations.

Figure 1, very kindly provided by Professor Brad Efron, illustrates such a plot based on a genetical application in which 6033 genes were examined for a possible connection with prostate cancer. A full description is given by Efron (2010, Section 2.1). The upward curvature of the plot shows a very clear departure from the theoretical null hypothesis distribution. A direct interpretation would be that the clear departure from linear form indicates that an appreciable number of genes are overexpressed relative to the null hypothesis. That there is a single fairly smooth curve implies that the data do not point to an isolated single anomalous point and that a possible explanation for all the data is essential agreement with the theoretical global null hypothesis distribution.

4 False Discovery Rate

The previous discussion has concentrated on testing the global null hypothesis, that is on the question of whether there is a signal at all. When it is likely that there is a small number of signals present, Benjamini and Hochberg (1995) suggested controlling the false discovery rate. There are two slightly different definitions of this. Let S signals be declared present of which F are false. Then the false discovery rate is defined as either

$$E(F/S \mid S > 0) \quad \text{or} \quad E(F/S).$$

The justification for the former definition is that if $S = 0$ there can be no false detection! In effect Benjamini and Hochberg's procedure uses cut off level among the $z_{(n-k)}$. Their formal analysis is in the

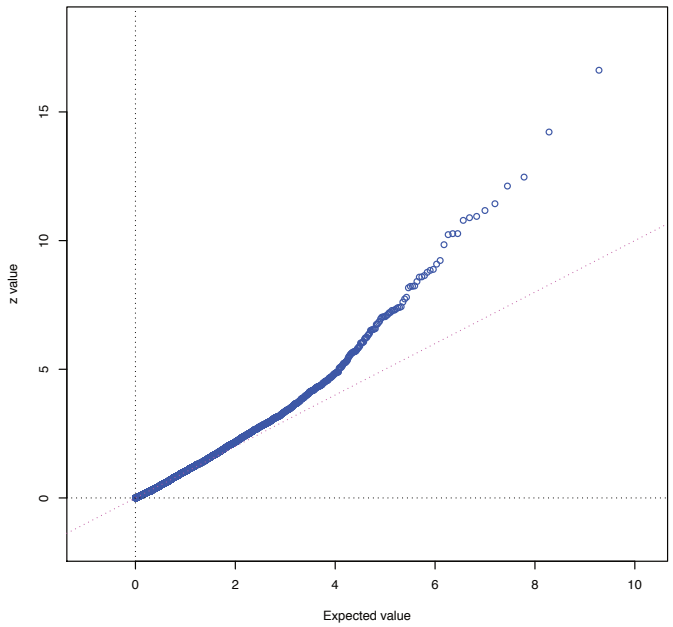


Fig. 1: Diagram provided by Professor Brad Efron arising from Efron (2010, Section 2.1). 6033 genes each tested for possibly being more active in prostate cancer patients. p-values converted to ordered z values where $z = -\log p$. Plotted against expectation assuming universal null hypothesis of no genuine effects. Upward curvature at right suggests some genes significantly overexpressed relative to null hypothesis.

spirit of Neyman-Pearson theory; it does not distinguish between overwhelming signals and those that only just emerge over threshold. This could be remedied to some extent by using several false discovery rates.

5 Bayesian Approach

We now turn briefly to a parallel Bayesian discussion. In the work summarized in his monograph Efron (2010) has developed a fully nonparametric empirical Bayes procedure in terms of two distributions, one for the null hypothesis and one applying when there is a signal, and a prior probability that the null is false. All are estimated empirically and in the monograph a number of examples of effective application are given. These do, however, require substantial amounts of data and a nontrivial number of signals. Cox and Wong (2004) gave a much simpler discussion in which the two associated distributions are normal with unit variance, the mean being zero under the null hypothesis and unknown under the signal. There are thus only two unknown parameters to be estimated; it was shown by simulation that satisfactory false recovery rates can be achieved. In both these procedures each potential signal is assigned an estimated posterior probability of being real.

To apply the procedure of Cox and Wong (2004) to the detection of a single signal we must somehow assign numerical values to the prior probability that there is a signal at a given site and to the mean of the corresponding normal distribution.

Suppose then that with n sites the prior probability of a signal at a particular site is α/n and that the corresponding normal distribution has mean μ_s. The total number of signals present thus has a Poisson distribution with mean α and the prior probability that the global null hypothesis holds is $e^{-\alpha}$. Thus if, for example, $\alpha = \log 2 = 0.69$ the prior probability of the global null hypothesis is one-half.

If the maximum normal-based test statistic is $t_{\max}$ the corresponding posterior log odds that there is a real signal, that is the log of the ratio of the probability of a real signal to the probability of a null

signal, is approximately

$$t_{\max}(\mu_s - t_{\max}/2) + \log(\alpha/n).$$

There is in this context no possibility of estimating μ_s from the data under analysis. The final term will typically be small and negative and, for given $t_{\max}$, the first term has maximum $t_{\max}^2/2$ achieved when $\mu_s = t_{\max}$. If $t_{\max} = \mu_s = 5$, the first term gives an odds ratio of approximately 10^5 compared with roughly $10^{-6.5}$ for the corresponding p-value, illustrating the superficially less extreme answers from the Bayesian formulation.

6 Discussion

A number of important issues are ignored in the preceding discussion. Typically discovery will be a multi-step procedure, starting often with preliminary data processing. Uncertainty at all stages will need consideration, although formal synthesis into a single measure of overall uncertainty may be neither necessary nor feasible. An extreme case of such a discovery process is traditional plant breeding in which a large number of varieties are reduced to a very small number in a series of trials involving progressively fewer and fewer varieties. Here issues of statistical significance at the separate stages are virtually irrelevant.

In many applications data are accrued over time and any signal will gradually emerge from noise. Analysis will proceed over time probably until a notable result can be reported and often data collection will continue past that. In a Bayesian formulation, provided the prior distribution and the general formulation do not change over time, there is no need to account for the repeated analysis.

No appreciable allowance for repeated analysis is required in frequentist theory, provided the formulation is in terms of confidence intervals with a target set for their width. The main formulation in previous sections is in terms of pure significance testing with no probability model of behaviour in the presence of a signal. Then allowance for the effect of repeated testing on the p-value is in general needed, but in the case of very extreme levels such as 5σ this allowance is likely to be negligible.

A more serious difficulty in application is the clash between the desire for the degree of objectivity achieved by precise prior specification of the procedure of analysis and the need to learn from the unexpected. In the present context it may be appropriate to concentrate on the single bin histogram approach possibly with allowance for data-dependent choice of bin width in the way outlined above.

It is a pleasure to thank Brad Efron for Fig. 1 and for helpful comments and Louis Lyons for comments on the paper and for many discussions of these issues.

References

[1] Y. Benjamini, Y. Hochberg, *J. R. Statist. Soc.* B **57**, 289-300 (1995).

[2] D.R. Cox and M.Y. Wong, *J. R. Statist. Soc.* B **66**, 395-400 (2004).

[3] B. Efron, *Large-scale inference.* IMS Monograph (Cambridge University Press, 2010).

[4] T. Schweder and E. Spjotvoll, *Biometrika* **69**, 493-502 (1982).

[5] Wellcome Trust Case Control Consortium, *Nature* **447**, 661-682 (2007).

The Bayesian Approach to Discovery

James Berger
Duke University, Durham NC, USA

Abstract
The Bayesian approach to discovery is essentially the Bayesian approach to hypothesis testing. This is discussed, through a pedagogical example that illustrates the approach and the differences with non-Bayesian approaches to testing, and through the Bayesian formulation of the generic HEP problem. Ensuing discussion focuses on the potential value of the Bayesian approach in dealing with the highly problematical 'look-elsewhere' effect, and the major challenge to the Bayesian approach, which is the choice of suitable prior distributions for unknown model parameters. A brief discussion of Bayesian unfolding is also included.

1 Introduction

'Discovery' can mean many things, from discovery of a completely anticipated entity such as the Higgs boson, to discovery of completely unanticipated new physics. Bayesian analysis is relevant to both types of discovery, but here we focus primarily on the former, in that it is easier to discuss Bayesian analysis for anticipated events because it (typically) reduces to Bayesian hypothesis testing.

Even the discussion of Bayesian hypothesis testing is, however, rather spotty. We begin with a simple pedagogical example of Bayesian testing, to set the notation and emphasize important distinctions with non-Bayesian approaches. The Bayesian formulation of the generic HEP problem is then introduced, to provide a vehicle for discussion of the look-elsewhere effect (multiple testing in statistical language). We partly focus on this aspect of the Bayesian approach because of its potential for dealing with one of the most troubling issues in discovery in HEP.

A major difficulty in the implementation of Bayesian hypothesis testing is the choice of the needed prior distributions of unknown parameters. This is considerably more problematical than in Bayesian estimation (e.g., in the choice of upper confidence limits), since standard objective priors are not available. A complete discussion of this issue is beyond the scope of this paper, but some of the basic issues involved are discussed.

Finally, an idea concerning unfolding (deconvolution) is discussed, because it also relates to a long-studied Bayesian problem.

2 A pedagogical example of Bayesian testing

As background to later developments, we review the ideas of Bayesian testing, as discussed in [1] through a simple example; we borrow many of the details from that reference.

Suppose the data, X, is the number of events observed in time T that are characteristic of Higgs boson production in an LHC particle collision experiment. The probabilistic model for the data is that X has density

$$\text{Poisson}(x \mid s+b) = \frac{(s+b)^x e^{-(s+b)}}{x!},$$

where s is the mean rate of production of Higgs events in time T in the experiment and b is the (assumed known) mean rate of production of such events from background sources in time T. Two specific values of X and b that we will follow through various analyses are

Case 1: $x = 7$ and $b = 1.2$; *Case 2:* $x = 6$ and $b = 2.2$.

The main purpose of the experiment is supposedly to determine whether or not the Higgs boson exists which, in terms of the probability model for the data, is typically phrased as testing $H_0 : s = 0$ versus $H_1 : s > 0$. Thus H_0 corresponds to 'no Higgs.'

The p**-value** in this example, corresponding to observed data x, is

$$p = P(X \geq x \mid b, s = 0) = \sum_{m=x}^{\infty} \text{Poisson}(m \mid 0 + b)\,.$$

For the two cases,

Case 1: $p = 0.00025$ if $x = 7$ and $b = 1.2$; *Case 2:* $p = 0.025$ if $x = 6$ and $b = 2.2$.

There is general agreement that a small p-value indicates that something unusual has happened, but that the p-value does not have a direct quantitative interpretation as evidence against the null hypothesis. Thus Luc Demortier observed in his talk at the Phystat 07 conference:

> "In any search for new physics, a small p-value should only be seen as a first step in the interpretation of the data, to be followed by a serious investigation of an alternative hypothesis. Only by showing that the latter provides a better explanation of the observations than the null hypothesis can one make a convincing case for discovery."

The **Bayes factor** of H_0 to H_1 in our ongoing example is given by

$$B_{01}(x) = \frac{\text{Poisson}(x \mid 0 + b)}{\int_0^\infty \text{Poisson}(x \mid s + b)\pi(s)\, ds} = \frac{b^x\, e^{-b}}{\int_0^\infty (s + b)^x\, e^{-(s+b)}\pi(s)\, ds}\,; \tag{1}$$

in the *subjective Bayesian approach*, the prior density, $\pi(s)$, is chosen to reflect the beliefs of the investigators (e.g., it could reflect the standard model predictions for the signal given information about the mass of the Higgs) while, in the *objective Bayesian approach*, it is chosen conventionally and nominally reflects a lack of knowledge concerning s.

A reasonable objective (though proper) prior here is the *intrinsic prior* $\pi^I(s) = b(s + b)^{-2}$ (see [1]). For this prior, the Bayes factor is given by

$$B_{01} = \frac{b^x\, e^{-b}}{\int_0^\infty (s + b)^x\, e^{-(s+b)} b(s + b)^{-2}\, ds} = \frac{b^{(x-1)}\, e^{-b}}{\Gamma(x - 1, b)}\,,$$

where Γ is the incomplete gamma function. The result for the two cases is

Case 1: $B_{01} = 0.0075$ (recall $p = 0.00025$); *Case 2:* $B_{01} = 0.26$ (recall $p = 0.025$)

The objective choice of prior probabilities of the hypotheses is $\Pr(H_0) = \Pr(H_1) = 0.5$, in which case

$$\Pr(H_0 \mid x) = \frac{B_{01}}{1 + B_{01}}\,.$$

For the two cases in the example,

Case 1: $\Pr(H_0 \mid x) = 0.0075$ (recall $p = 0.00025$); *Case 2:* $\Pr(H_0 \mid x) = 0.21$ (recall $p = 0.025$).

In addition to the uncertainty in the hypotheses, there is also uncertainty in s, given that H_1 were true. The complete posterior distribution is thus determined by

- $\Pr(H_0 \mid x)$, the posterior probability of the null hypothesis;
- $\pi(s \mid x, H_1)$, the posterior distribution of s under H_1.

For Case 1 in the example, Figure 1 presents these two parts of the full posterior distribution. One way of thinking of this is that the vertical bar gives the probability that one has just observed noise, while the density part says where s is likely to be if there is a discovery.

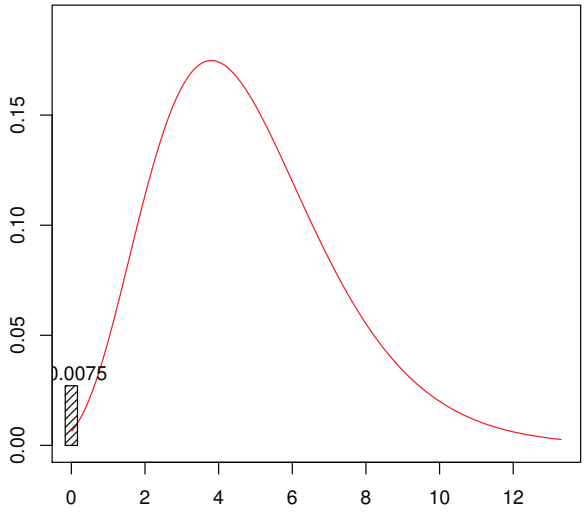


Fig. 1: For Case 1, $\Pr(H_0 \mid x)$ (the vertical bar), and the posterior density for s given $x = 7$ and H_1.

A useful summary of the complete posterior is $\Pr(H_0 \mid x)$ and C, a (say) 95% posterior confidence interval for s under H_1. For the two cases, and with C chosen to be an equal-tailed 95% posterior confidence interval (i.e., omitting 2.5% of the posterior mass on the left and the right)

Case 1: $\Pr(H_0 \mid x) = 0.0075$ and $C = (1.0, 10.5)$; *Case 2:* $\Pr(H_0 \mid x) = 0.21$ and $C = (0.2, 8.2)$.

C could, alternatively, be chosen to be a one-sided confidence bound, if desired.

Note that confidence intervals alone are *not* a satisfactory inferential summary. In Case 2, for instance, the 95% confidence interval does not include 0, and so many mistakenly believe that one can accordingly reject $H_0 : s = 0$. But, the full posterior distribution also has a probability of 0.21 that $s = 0$, which would hardly imply a confident rejection.

The Bayesian error probabilities given in the previous section differed from the corresponding p-values by factors of 30 and 10 in the two cases, respectively. It might be tempting to say that there is something wrong with the Bayesian analysis, but even a pure likelihood analysis reveals the same effect. In particular (following [2]), note that a lower bound on the Bayes factor over all possible priors can be found by choosing $\pi(s)$ to be a point mass at $\hat{s}$ (the maximum likelihood estimate), yielding

$$B_{01}(x) = \frac{\text{Poisson}(x \mid 0 + b)}{\int_0^\infty \text{Poisson}(x \mid s + b)\pi(s)\, ds} \geq \frac{\text{Poisson}(x \mid 0 + b)}{\text{Poisson}(x \mid \hat{s} + b)} = \min\{1, \left(\frac{b}{x}\right)^x e^{x-b}\}\,. \qquad (2)$$

In 'likelihood language,' this says that, for the given data, the likelihood of H_0 relative to the likelihood of H_1 is at least the bound on the right hand side of (2). For the two cases, this bound is

Case 1: $B_{01} \geq 0.0014$ (recall $p = 0.00025$); *Case 2:* $B_{01} \geq 0.11$ (recall $p = 0.025$),

so that a serious discrepancy remains even when the prior is eliminated. This is partly due to the fact that the p-value is based on the probability of the tail area of the distribution, rather than the probability of the actual data. For further discussion of the discrepancy (and problems with interpretation of p-values) see [1]. It is also shown there how the objective Bayesian posterior probabilities are also the optimal conditional frequentist error probabilities, so that both Bayesian and frequentist philosophies support the conclusion that p-values cannot in any sense be viewed as actual error rates.

3 A Bayesian formulation of the basic HEP problem

Following [3], a more complete model for the basic HEP problem can be summarized as follows. Let N be the observed Poisson number of events. The events are independent and each has characteristics ('marks' in the Poisson process world) X_i, $i = 1, \ldots, N$. These events could arise from only a background Poisson process, having mean b and density $f_b(x)$ (hypothesis H_0). In addition to background, there could also be an additive signal Poisson process with mean s and density $f_s(x)$ (hypothesis H_1). It is then of interest to test

H_0 : N has mean b, and the X_i have density $f_b(x) > 0$, versus

H_1 : N has mean $b + s$, and the X_i have density $(\gamma f_b(x) + (1-\gamma) f_s(x))$, where $\gamma = \frac{b}{(b+s)}$.

We will consider the case where $f_b(x)$ and $f_s(x)$ are known but b and s are unknown.

The Bayes factor of H_1 to H_0 for priors $\pi_0(b)$ and $\pi_1(b, s) = \pi_0(b)\pi_1(s \mid b)$ is

$$\begin{aligned} B_{10} &= \frac{\int_0^\infty \int_0^\infty (b+s)^N e^{-(b+s)} \prod_{i=1}^N [\gamma f_b(x_i) + (1-\gamma) f_s(x_i)] \pi_1(b, s)\, ds db}{\int_0^\infty b^N e^{-b} \prod_{i=1}^N [f_b(x_i)]\, \pi_0(b) db} \\ &= \frac{\int_0^\infty \int_0^\infty b^N e^{-(b+s)} \prod_{i=1}^N \left[1 + \frac{s f_s(x_i)}{b f_b(x_i)}\right] \pi_0(b)\pi_1(s \mid b)\, ds db}{\int_0^\infty b^N e^{-b}\, \pi_0(b) db}\,. \end{aligned} \tag{3}$$

Note that, if b is known, this becomes

$$B_{10} = \int_0^\infty e^{-s} \prod_{i=1}^N \left[1 + \frac{s f_s(x_i)}{b f_b(x_i)}\right] \pi_1(s \mid b)\, ds\,.$$

In the absence of (or desire not to utilize) subjective priors, recommended objective choices are the intrinsic priors $\pi_0^I(b) = b^{-1/2}$ and $\pi_1^I(s \mid b) = b(s+b)^{-2}$. The latter is (necessarily) proper, and is justified in the same fashion as the choice in the simpler problem given in Section 2. Since b occurs in both models, it is allowable (and desirable) to utilize the standard objective prior for a Poisson mean, which is $b^{-1/2}$; see Section 5 for discussion.

Note that ignoring the densities f_s and f_b and basing the answer solely on N (as in Section 2) is equivalent to assuming that $f_s \equiv f_b$. It is thus not the case that the simpler analysis simply utilizes less information; it could actually be misleading.

4 Controlling for look-elsewhere effects

A major strength of Bayesian analysis is that it easily (and often automatically) adjusts for look-elsewhere effects. This is illustrated using (3), for a situation of multiple hypothesis testing. The interesting issues involving multiple cuts are then briefly discussed. Finally, the contrasting difficulty of frequentist adjustment for look-elsewhere effects is illustrated.

4.1 Multiple Hypotheses

Suppose N_j of the X_i are in bin B_j, $j = 1, \ldots, m$, and that we assume we have only densities $f_s(B_j)$ and $f_b(B_j)$. Then

$$B_{10} = \frac{\int_0^\infty \int_0^\infty b^N e^{-(b+s)} \prod_{i=1}^m \left[1 + \frac{s f_s(B_j)}{b f_b(B_j)}\right]^{N_j} \pi_0(b)\pi_1(s \mid b)\, ds db}{\int_0^\infty b^N e^{-b}\, \pi_0(b) db}\,.$$

Suppose, in addition, that $f_s(B_j)$ gives probability one to some unknown bin B, with each bin being equally likely. This is equivalent to saying that we are testing the mutually exclusive hypotheses:

H_j : *signal is only in bin* B_j, with the hypotheses having equal prior probability. Then

$$\begin{aligned} B_{10} &= \frac{E^B\left[\int_0^\infty \int_0^\infty b^N e^{-(b+s)} \prod_{i=1}^m \left[1+\frac{sf_s(B)}{bf_b(B_j)}\right]^{N_j} \pi_0(b)\pi_1(s \mid b)\, dsdb\right]}{\int_0^\infty b^N e^{-b}\, \pi_0(b) db} \\ &= \frac{\frac{1}{m}\sum_{j=1}^m \int_0^\infty \int_0^\infty b^N e^{-(b+s)} \left[1+\frac{s}{bf_b(B_j)}\right]^{N_j} \pi_0(b)\pi_1(s \mid b)\, dsdb}{\int_0^\infty b^N e^{-b}\, \pi_0(b) db}\,. \end{aligned}$$

The point here is that the marginal likelihood of hypothesis H_i gets automatically down-weighted by $1/m$, the 'cost' of looking in m different bins. There is no need to make any adjustment for this 'looking elsewhere;' it happens automatically as part of the Bayesian analysis.

4.2 Multiple Cuts

The situation involving multiple cuts is interesting, in that Bayesian analysis does not readily apply. A cut really just produces a subset of the overall data, and there is no natural Bayesian way to separately analyze different subsets of data and combine the analyses for an overall conclusion.

If each of m cuts produces data $X_j, j = 1, \ldots, m$, one could legitimately consider the joint distribution of $(X_1, \ldots, X_m)$, and perform a Bayesian analysis. But this is typically not possible because of the difficulty of determining the dependence between the X_j. Of course, if the cuts were such that the X_j could be considered independent, combined analysis is possible: simply multiply the individual cut likelihoods and proceed.

It seems that use of cuts is a practical necessity, but it is worth noting that they are not inherently needed in Bayesian analysis. Suppose, for instance, that $f_s(x) = 0$ for $x \in \Omega^c$, so that it seems tempting to cut on Ω as this is the only region that can possibly contain the signal. But is this necessary? Note that, in this situation, (3) becomes (with the last expression below following from algebra, not probability logic)

$$\begin{aligned} B_{10} &= \frac{\int_0^\infty \int_0^\infty b^N e^{-(b+s)} \prod_{\{i:x_i\in\Omega\}} \left[1+\frac{sf_s(x_i)}{bf_b(x_i)}\right] \pi_0(b)\pi_1(s \mid b)\, dsdb}{\int_0^\infty b^N e^{-b}\, \pi_0(b) db} \\ &= \int_0^\infty B_{10}(x_\Omega \mid b)\pi_0(b \mid x_\Omega, x_{\Omega^c}) db\,, \end{aligned}$$

where x_Ω (x_{Ω^c}) is the data in Ω (in Ω^c). Thus the Bayesian analysis indeed uses the cut data to find the Bayes factor given b, as would be expected, but it uses all the data to learn more about b, which is clearly desirable (unless, of course, b were different over Ω and Ω^c).

4.3 The difficulty in frequentist control of the look-elsewhere effect

To indicate the difference between the Bayesian and frequentist approaches to controlling look-elsewhere effects, consider the simple multiple testing scenario of testing $H_{0i} : \mu_i = 0$ versus $H_{1i} : \mu_i > 0$, $i = 1, ..., m$, based on data $X_i, i = 1, \ldots, m$, that are normally distributed with mean μ_i, variance 1, and correlation ρ. Furthermore, suppose we know that there is at most one signal.

If $\rho = 0$, one can just do the individual tests at level α/m (Bonferroni) to obtain an overall error probability of α. If $\rho > 0$, however, the situation is more difficult. One natural way to proceed would be to choose the overall decision rule "declare μ_i to be a signal if X_i is the largest value and $X_i > K$," and then compute the corresponding frequentist type I error probability

$$\alpha = \Pr(\max_i X_i > K \mid \mu_1 = \ldots = \mu_m = 0) = E^Z\left[1 - \Phi\left(\frac{K - \sqrt{\rho}Z}{\sqrt{1-\rho}}\right)^m\right],$$

where Φ is the standard normal cdf and Z is a standard normal random variable.

This gives (essentially) the Bonferroni correction when $\rho = 0$, but can be shown to converge to $1 - \Phi[K]$ as $\rho \to 1$, which is the type I error that would result from a single test. Thus the needed frequentist control for multiple testing ranges from the drastic Bonferroni correction to none, depending on the correlations among the data.

In contrast, the Bayesian adjustment for multiple testing does not depend on correlations among the data, and occurs only through the choice of prior probabilities of hypotheses [4]. In the above scenario, for instance, one might assign prior probability 1/2 to no signal, and assign each of the possible alternative hypotheses prior probability of $1/(2m)$ (recall that we are assuming that at most one alternative is true.). The ensuing Bayesian analysis correctly controls for the look-elsewhere effect, regardless of the data distribution.

5 Remarks on choice of priors

A primary difficulty in the Bayesian approach to discovery is the difficulty in choosing prior distributions. Of course, if subjective (or evidence-based) priors are available – and if use of such priors is viewed as appropriate – there is no problem. Typically, however, such priors are not available for, at least, many parameters in the likelihood, and default choices are needed. A brief overview of the situation is given here; more extensive discussions and references can be found in [5, 6].

5.1 Testing versus estimation priors

Unfortunately, the situation in testing (discovery) is quite different from the situation in estimation (e.g., setting confidence limits). For the latter problem, excellent objective priors are available, such as reference priors [7, 8]. These priors are typically improper, which is not a problem for estimation but is often a problem for testing. In (1) for instance, suppose we were to consider the improper prior $\pi(s) = c/\sqrt{s}$, where c is a constant; this is the standard reference prior for estimation. Note that there is no natural choice of c, since the prior is improper and, since the choice of c is irrelevant to estimation, one typically sees the choice $c = 1$. In (1), however, use of $\pi(s) = c/\sqrt{s}$ yields

$$B_{01}(x) = \frac{b^x \, e^{-b}}{c \int_0^\infty (s+b)^x \, e^{-(s+b)} s^{-1/2} \, ds} \,,$$

which is arbitrary since c is arbitrary. In general, parameters that occur in one hypothesis – but not the other – require proper priors.

5.2 Choosing priors for "common parameters"

Often parameters occur in both the likelihoods in the numerator and denominator of a Bayes factors, as does b in (3). Then it is possible (and usually desirable) to use the estimation default priors, e.g. $c/\sqrt{b}$, since c will now cancel in the numerator and denominator.

The difficulty here is in ensuring that the parameters are actually the same in both likelihoods. When the parameters have physical meaning, such as b, this is not an issue. Also, when parameters have what is called the same group invariance structure, one can use the right-Haar priors for those parameters [9].

In other contexts, however, one has to be careful. In variable selection in regression for instance, the meaning of regression coefficients is highly affected by which other coefficients are in the model, and use of a common prior may then be inappropriate. This is usually dealt with by orthogonalizing the parameters (making a transformation so that the partial Fisher information is zero) – see [10].

5.3 Basics of choosing priors for non-common parameters

This is the difficult situation, and there are no ready answers. We have already seen that improper priors cannot be used. Even worse is the (all-too-common) use of vague proper priors.

Example: Suppose X has the normal density $\phi(x \mid \mu, 1)$ with mean μ and variance 1. It is desired to test $H_0 : \mu = 0$ versus $H_1 : \mu \neq 0$ with a Uniform$(-c, c)$ prior for μ where c is large, so that one has a 'vague proper prior.' The Bayes factor is

$$B_{01}(c) = \frac{\phi(x \mid 0, 1)}{\int_{-c}^{c} \phi(x \mid \mu, 1)(2c)^{-1} d\mu} \approx \frac{2c\, \phi(x \mid 0, 1)}{\int_{-\infty}^{\infty} \phi(x \mid \mu, 1) d\mu}$$

for large c, which depends dramatically on the choice of c. One might try specifying a plausible range $c_1 \leq c \leq c_2$ and look at $B_{01}(c)$ over this range, hoping the the conclusion is robust, but this will typically not be the case.

It is interesting to obtain some feel for the sensitivity of Bayes factors to the choice of prior. In this example, consider three prior distributions:

- $\pi^u(\mu)$ is Uniform$(0, 10)$, corresponding, say, to a known upper limit on μ.
- $\pi^e(\mu)$ is Normal$(4, 1)$, an evidence-based prior arising from a previous experiment.
- $\pi^t(\mu)$ is a point mass at 4, the prediction of a new theory.

Table 1 gives the posterior probabilities of the null hypothesis that result from these priors for various values of x (equivalent to the number of σ from zero), assuming that the prior probability of the null hypothesis is 1/2. Indeed the posterior probabilities are quite sensitive to the choice of prior. The view of Bayesians, however, is that this sensitivity is an unavoidable fact of life; the three priors correspond to quite different types of prior knowledge and that this knowledge can have a pronounced effect should not be surprising. The corresponding p-values are also given in Table 1, to emphasize the fact that, while the Bayesian answers are sensitive to the prior, they have much more in common with each other than with the p-value.

Table 1: Posterior probabilities of H_0, given various data x from a $N(\mu, 1)$ distribution, assuming prior probability of 1/2 for H_0 and use of the uniform, normal and point mass priors for μ, as well as the corresponding p-values.

	$\pi^u(\mu)$	$\pi^e(\mu)$	$\pi^t(\mu)$	p-value
$x = 2$	0.54	0.34	0.50	$p = 0.025$
$x = 4$	1.3×10^{-3}	4.7×10^{-4}	3.4×10^{-4}	$p = 3.1 \times 10^{-5}$
$x = 6$	6.0×10^{-8}	5.8×10^{-8}	1.1×10^{-7}	$p = 1.0 \times 10^{-9}$

5.4 Various proposed default priors for non-common parameters

There is a long history concerning suggestions for priors for non-common parameters in hypothesis testing and model selection. We give a brief description of the most commonly used methods here. Much more extensive discussions of this history can be found in [5, 6]. A focus of much of this work is to ensure that priors are appropriately balanced between the hypotheses, often called *predictive matching*.

5.4.1 Fractional Bayes factors

Fractional priors [11] use a fraction γ of the likelihood $L(s)$ as the prior, i.e., $\pi(s) = L(s)^\gamma / \int L(s)^\gamma ds$, with the remaining part of the likelihood, $L(s)^{1-\gamma}$, being treated as the likelihood for the Bayes factor computation. For testing between two models, with likelihoods $L_1(s_1)$ and $L_0(s_0)$, this leads to the Bayes factor

$$B_{10} = \frac{\int L_1(s_1)^{1-\gamma_1} \pi_1(s_1) ds_1}{\int L_0(s_0)^{1-\gamma_0} \pi_0(s_0) ds_0} = \frac{\int L_1(s_1) ds_1 \int L_0(s_0)^{\gamma_0} ds_0}{\int L_0(s_0) ds_0 \int L_1(s_1)^{\gamma_1} ds_1}\,.$$

This is often computationally attractive, and usually does correspond (at least asymptotically) to a real Bayesian analysis if the γ_i are chosen appropriately; the typically recommended choice is $\gamma_i = n/p_i$, where n is the sample size of the data, and p_i is the dimension of the parameter s_i. For discussion of the strengths and weaknesses of this approach, see [5].

5.4.2 Intrinsic priors

Intrinsic priors (see [5] for discussion and earlier references) are generated in a bootstrap fashion using either subsets of the data or artificial data. They are very widely applicable and have excellent Bayesian properties, but can be computationally intensive.

One of the methods of deriving an intrinsic prior is through the *expected posterior* prior construction [12]. For the situation where the data consists of i.i.d. observations from a density $f(x \mid s)$, and for testing $H_0 : s = s_0$ versus $H_1 : s \neq s_0$, the construction is as follows:

- let $\pi^O(s)$ be a good estimation objective prior, so that $\pi^O(s \mid \boldsymbol{x}) = [\prod_{i=1}^n f(x_i \mid s)]\pi^O(s)/m^O(\boldsymbol{x})$ is the resulting posterior, where $\boldsymbol{x} = (x_1, \ldots, x_n)$ and $m^O(\boldsymbol{x}) = \int [\prod_{i=1}^n f(x_i \mid s)]\pi^O(s)\, ds$;
- then the intrinsic prior is $\pi^I(s) = \int \pi^O(s \mid \boldsymbol{x}^*)[\prod_{i=1}^q f(x_i \mid s_0)]\, d\boldsymbol{x}^*$, with $\boldsymbol{x}^* = (x_1, \ldots, x_q)$ being (unobserved) data of the minimal sample size q such that $m^O(\boldsymbol{x}^*) < \infty$.

Note that this will be a proper (not vague proper) prior.

The idea behind this prior is that, if one were handed the data $\boldsymbol{x}^*$ but allowed to use it only for prior construction, one would happily compute $\pi^O(s \mid \boldsymbol{x}^*)$ and use this proper prior to conduct the test. We don't have $\boldsymbol{x}^*$ available, but we could simulate $\boldsymbol{x}^*$ from the null model, and compute the resulting 'average' prior. This is the method used to derive the intrinsic prior $\pi_1^I(s \mid b)$ in Section 1.

5.4.3 Conventional priors

For common specific situations, proper conventional priors have been proposed. For instance, in testing involving the normal linear model, numerous proper default priors have been proposed that depend only on the design matrix of the model being considered. The most popular of these conventional priors are the Zellner-Siow priors [13], which were developed following ideas of [10]. These priors have some very nice properties. In particular, they result in answers that

- are invariant to scale changes in covariates (i.e., the units of measurement used);
- are consistent (i.e., the true model will be selected as $n \to \infty$), if the true model is among those being considered;
- are information-consistent (i.e., will reject the null model as the associated t or F statistics $\to \infty$).

For description and further discussion, see [5].

5.4.4 Approximations

Because of the difficulty in choosing prior distributions and in computing Bayes factors, use of approximations such as BIC (Bayes information criterion [14]) is often considered; BIC neither requires specification of priors nor integration over the likelihood. The accuracy of the approximation is, however, mixed, at best. The approximation does capture part of the influence of prior distribution in a generic way, thus allowing for an Ockham's Razor effect of preferring the more parsimonious of two models that equally well explain the data. But BIC basically ignores constants in the Bayes factor that arise from the priors, and these constants can be arbitrarily large or small, so the approach is by no means a cure-all.

6 A comment on unfolding

The workshop also had a focus on unfolding (deconvolution), which has historically also been a central problem in Bayesian statistics. If one has a density $f(x \mid s)$ of data x, given an unknown parameter s, and a prior distribution $\pi(s)$ for s, the predictive (or marginal) distribution of x is then

$$m_\pi(x) = \int f(x \mid s)\pi(s)\, ds\,.$$

Often one is in a situation of having an estimate $\hat{m}(x)$ for the predictive distribution, and the goal is then to find $\pi(s)$ such that $m_\pi(x)$ is as close as possible to $\hat{m}(x)$, the unfolding problem.

In [15], a very interesting algorithm for attacking the problem is presented. Start with any initial $\pi_0(s)$ that has support everywhere. Then iteratively compute

$$\pi_l(s) = \int \pi_{l-1}(s \mid x)\hat{m}(x)dx\,, \quad \text{where } \pi_{l-1}(s \mid x) = \frac{\pi_{l-1}(s)f(x \mid s)}{\int \pi_{l-1}(s)f(x \mid s)ds}\,.$$

Theorem: $\pi^*(s) = \lim_{l\to\infty} \pi_l(s)$ is the density for which $m_\pi(x)$ is as close as possible to $\hat{m}(x)$ in Kullback-Leibler divergence.

Here is a potentially interesting implementation of this algorithm using particle filtering:

- Represent $\pi_l(s)$ by a collection of *particles* $\{s_i\}$ with weights $\{w_i^{(l)}\}$. (Initialize with a random sample $\{s_i\}$ from $\pi_0(s)$, so the initial weights are equal.)
- Then $\pi_l(s \mid x)$ would be the same collection of particles but with modified weights

$$w_i^{(l)}(x) = \frac{w_i^{(l-1)} f(x \mid s_i)}{\sum_j w_j^{(l-1)} f(x \mid s_j)}\,,$$

  and $\pi_{l+1}(s)$ would be the same collection of particles but with weights

$$w_i^{(l+1)} = \int w_i^{(l)}(x)\hat{m}(x)dx\,.$$

- As one progresses one will need to add new particles adapting to the evolving density, but there are likely techniques in particle filtering for doing this.

Acknowledgements

This work was supported by NSF Grants AST-0507481, DMS-0757549-001, and DMS-1007773.

References

[1] Berger, J. (2008). A comparison of testing methodologies. In *Proceedings of the PHYSTAT-LHC Workshop on Statistical Issues for LHC Physics,* **CERN 2008-001,** 8–19.

[2] Edwards, W., Lindman, H. and Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, **70**, 193–242.

[3] Lockhart, R. (2010). Banff Challenge 2 – statistician's language. Available at http://www.birs.ca/workshops/2010/10w5068/10w5068BanffChallenge2.pdf.

[4] Scott, J. and Berger, J. (2010). Bayes and Empirical-Bayes multiplicity adjustment in the variable-selection problem. *Annals of Statistics* **38**, 2587–2619.

[5] Berger, J. and Pericchi, L. (2001). Objective Bayesian methods for model selection: introduction and comparison (with Discussion). In *Model Selection*, P. Lahiri, ed., Institute of Mathematical Statistics Lecture Notes – Monograph Series, volume 38, Beachwood Ohio, 135–207.

[6] Liang, F., Paulo, R., Molina, G., Clyde, M. A. and Berger, J. O. (2008). Mixtures of g priors for Bayesian variable selection. *J. Amer. Statist. Assoc.*, **103**, 410–423.

[7] Bernardo, J. M. (2005). Reference analysis. Handbook of Statistics 25 (D. K. Dey and C. R. Rao eds.). Amsterdam: Elsevier, 17–90.

[8] Demortier, L. (2005). Bayesian reference analysis for particle physics. *PHYSTAT05 Proceedings on "Statistical Problems in Particle Physics, Astrophysics and Cosmolgy"*. Imperial College Press.

[9] Berger, J., Pericchi, L, and Varshavsky, J. (1998). Bayes factors and marginal distributions in invariant situations. *Sankhyā*, **A 60**, 307–321.

[10] Jeffreys, H. (1961). *Theory of Probability*, London: Oxford University Press.

[11] O'Hagan, A. (1995). Fractional Bayes factors for model comparisons. *Journal of the Royal Statistical Society*, Ser. B, **57**, 99–138.

[12] Pérez, J.M. and Berger, J. (2002). Expected posterior prior distributions for model selection. *Biometrika*, **89**, 491–512.

[13] Zellner, A. and Siow, A. (1980). Posterior Odds Ratios for Selected Regression Hypotheses. In Bayesian Statistics: Proceedings of the First International Meeting held in Valencia (Spain) (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.) 585–603. University of Valencia.

[14] Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–464.

[15] Shyamalkumar, N.D. (1996). Cyclic I_0 projections and its applications to statistics. Technical Report 96-24, Department of Statistics, Purdue University.

Bayes and Discovery: Objective Bayesian Hypothesis Testing

José M. Bernardo
Universitat de València, Spain

Abstract
Hypothesis testing is formulated from a decision theoretical viewpoint. The combined use of intrinsic discrepancy, an invariant information-based loss function, and conventional reference priors provides an objective Bayesian solution to precise hypothesis testing problems which easily integrates with the standard formulation of objective Bayesian point and region estimation.

1 Precise Hypothesis Testing

Let $\boldsymbol{z}$ be the available data which are assumed to have been generated as one random observation from model $\mathcal{M}_{\boldsymbol{z}} = \{p(\boldsymbol{z} \,|\, \boldsymbol{\omega}), \boldsymbol{z} \in \boldsymbol{\mathcal{Z}}, \boldsymbol{\omega} \in \boldsymbol{\Omega}\}$. Often, but not always, data will consist of a random sample $\boldsymbol{z} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ from some distribution $q(\boldsymbol{x} \,|\, \boldsymbol{\omega})$, with $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$; in this case $p(\boldsymbol{z} \,|\, \boldsymbol{\omega}) = \prod_{i=1}^{n} q(\boldsymbol{x}_i \,|\, \boldsymbol{\omega})$ and $\boldsymbol{\mathcal{Z}} = \boldsymbol{\mathcal{X}}^n$. Let $\boldsymbol{\theta}(\boldsymbol{\omega})$ be the vector of interest. Without loss of generality, the model may explicitly be expressed in terms of $\boldsymbol{\theta}$ so that $\mathcal{M}_{\boldsymbol{z}} = \{p(\boldsymbol{z} \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda}), \boldsymbol{z} \in \boldsymbol{\mathcal{Z}}, \boldsymbol{\theta} \in \boldsymbol{\Theta}, \boldsymbol{\lambda} \in \boldsymbol{\Lambda}\}$, where $\boldsymbol{\lambda}$ is some appropriately chosen nuisance parameter vector. Let $\pi(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \pi(\boldsymbol{\lambda} \,|\, \boldsymbol{\theta})\, \pi(\boldsymbol{\theta})$ be the assumed prior, and let $\pi(\boldsymbol{\theta} \,|\, \boldsymbol{z})$ be the corresponding marginal posterior distribution of $\boldsymbol{\theta}$. Appreciation of the inferential contents of $\pi(\boldsymbol{\theta} \,|\, \boldsymbol{z})$ may be enhanced by providing both point and region estimates of the vector of interest $\boldsymbol{\theta}$, and by declaring whether or not some context-suggested specific value $\boldsymbol{\theta}_0$ is compatible with the observed data $\boldsymbol{z}$ (precise hypothesis testing). A large number of Bayesian estimation and hypothesis testing procedures have been proposed in the literature. We argue that their choice is better made in decision theoretical terms.

Let $\ell\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\}$ describe, as a function of the (unknown) parameter values $(\boldsymbol{\theta}, \boldsymbol{\lambda})$ which have generated the available data, *the loss* to be suffered if, working with model $\mathcal{M}_{\boldsymbol{z}}$, the value $\boldsymbol{\theta}_0$ were used as a proxy for the unknown value of $\boldsymbol{\theta}$. Point estimation, region estimation and hypothesis testing procedures may all be appropriately described as specific decision problems using a common prior distribution and a common loss function of this type. The results, which are obviously all conditional on the assumed model $\mathcal{M}_{\boldsymbol{z}}$, may dramatically depend on the particular choices made for both the prior and the loss functions but, given the available data $\boldsymbol{z}$, they all only depend on those through the corresponding posterior expected loss,

$$\overline{\ell}(\boldsymbol{\theta}_0 \,|\, \boldsymbol{z}) = \int_{\boldsymbol{\Theta}} \int_{\boldsymbol{\Lambda}} \ell\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\}\, \pi(\boldsymbol{\theta}, \boldsymbol{\lambda} \,|\, \boldsymbol{z})\, d\boldsymbol{\theta} d\boldsymbol{\lambda}. \tag{1}$$

As a function of $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$, the expected loss $\overline{\ell}(\boldsymbol{\theta}_0 \,|\, \boldsymbol{z})$ provides a direct measure of the relative unacceptability of all possible values of the quantity of interest in the light of the information provided by the data.

In this paper, we will concentrate on precise hypothesis testing, with objective reference priors. For a more general perspective and many examples, see Bernardo (2011) and references therein.

1.1 Decision Theoretic Formulation

Consider a value $\boldsymbol{\theta}_0$ of the vector of interest which deserves special consideration, either because assuming $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ would noticeably simplify the model, or because there are additional context-specific arguments suggesting that $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. Intuitively, the value $\boldsymbol{\theta}_0$ should be judged to be *incompatible* with

the observed data $\boldsymbol{z}$ if the posterior expected loss $\overline{\ell}(\boldsymbol{\theta}_0 \,|\, \boldsymbol{z})$ of using $\boldsymbol{\theta}_0$ as a proxy for $\boldsymbol{\theta}$ is too large. This notion is now made precise.

Formally, testing the hypothesis $H_0 \equiv \{\boldsymbol{\theta} = \boldsymbol{\theta}_0\}$ may be described as a decision problem where the action space $\mathcal{A} = \{a_0, a_1\}$ contains only two elements: to accept (a_0) or to reject (a_1) the hypothesis under scrutiny. Foundations require specification of a loss function $\ell_h\{a_i, (\boldsymbol{\theta}, \boldsymbol{\lambda})\}$ measuring the consequences of accepting or rejecting H_0 as a function of the actual parameter values. By assumption, a_0 means to *act as if* H_0 were true, that is to work with the model $\mathcal{M}_0 = \{p(\boldsymbol{z} \,|\, \boldsymbol{\theta}_0, \boldsymbol{\lambda}_0), \boldsymbol{z} \in \boldsymbol{\mathcal{Z}}, \boldsymbol{\lambda}_0 \in \boldsymbol{\Lambda}\}$, while a_1 means to reject this simplification and to keep working with model $\mathcal{M}_{\boldsymbol{z}}$=$\{p(\boldsymbol{z} \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda}), \boldsymbol{z} \in \boldsymbol{\mathcal{Z}}, \boldsymbol{\theta} \in \boldsymbol{\Theta}, \boldsymbol{\lambda} \in \boldsymbol{\Lambda}\}$. Alternatively, an already established model $\mathcal{M}_0$ may have been embedded into a more general model $\mathcal{M}_{\boldsymbol{z}}$, constructed to include promising departures from $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, and it is required to verify whether presently available data $\boldsymbol{z}$ are still compatible with $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, or whether the extension to $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ is really necessary. Given the available data $\boldsymbol{z}$, the optimal action will be to reject the hypothesis considered if (and only if) the expected posterior loss of accepting (a_0) is larger than that of rejecting (a_1), so that $\int_{\boldsymbol{\Theta}} \int_{\boldsymbol{\Lambda}} [\ell_h\{a_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\} - \ell_h\{a_1, (\boldsymbol{\theta}, \boldsymbol{\lambda})\}] \, \pi(\boldsymbol{\theta}, \boldsymbol{\lambda} \,|\, \boldsymbol{z}) \, d\boldsymbol{\theta} d\boldsymbol{\lambda} > 0$. Hence, only the loss difference $\Delta\ell_h\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\} = \ell_h\{a_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\} - \ell_h\{a_1, (\boldsymbol{\theta}, \boldsymbol{\lambda})\}$, which measures the *advantage* of rejecting $H_0 \equiv \{\boldsymbol{\theta} = \boldsymbol{\theta}_0\}$ as a function of the parameter values, must be specified. The hypothesis H_0 should be rejected whenever the expected advantage of rejecting is positive. Without loss of generality, the function $\Delta\ell_h$ may be written in the form

$$\Delta\ell_h\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\} = \ell\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\} - \ell_0$$

where, as mentioned above (and precisely as in estimation problems), $\ell\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\}$ describes the non-negative loss to be suffered if $\boldsymbol{\theta}_0$ were used as a proxy for $\boldsymbol{\theta}$. Since $\ell\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}_0, \boldsymbol{\lambda})\} = 0$, so that $\Delta\ell_h\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}_0, \boldsymbol{\lambda})\} = -\ell_0$, the value $\ell_0 > 0$ describes (in the same loss units) the context-dependent non-negative advantage of accepting $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ when it is true. With this formulation, the optimal action is to reject $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ whenever the expected value of $\ell\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\} - \ell_0$ is positive, i.e., whenever $\overline{\ell}(\boldsymbol{\theta}_0 \,|\, \boldsymbol{z})$, the posterior expectation of $\ell\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\}$, is larger than ℓ_0. Thus, as intuition suggested, the solution to the precise hypothesis testing decision problem posed is found in terms of the value of the expected loss $\overline{\ell}(\boldsymbol{\theta}_0 \,|\, \boldsymbol{z})$ of using $\boldsymbol{\theta}_0$ as a proxy for the unknown value of $\boldsymbol{\theta}$.

Using the zero-one loss function, $\ell\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\} = 0$ if $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, and $\ell\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\} = 1$ otherwise, so that the loss advantage of rejecting $\boldsymbol{\theta}_0$ is a constant whenever $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ and zero otherwise, leads to rejecting H_0 if (and only if) $\Pr(\boldsymbol{\theta} = \boldsymbol{\theta}_0 \,|\, \boldsymbol{z}) < p_0$ for some context-dependent p_0. Notice that, using this particular loss function, if one is to avoid a systematic rejection of H_0 (whatever the data), the prior probability $\Pr(\boldsymbol{\theta} = \boldsymbol{\theta}_0)$ must be *strictly positive*. If $\boldsymbol{\theta}$ is a continuous parameter this requires the use of a non-regular "sharp" prior, concentrating a positive probability mass at $\boldsymbol{\theta}_0$. With no mention of the (rather naïve) loss structure which is implicit in the formulation, this type of solution was early advocated by Jeffreys (1961). Notice however, that this formulation implies the use of radically different priors for hypothesis testing than those used for estimation, and a different prior for each value to be tested. Moreover, this formulation is known to lead to the difficulties associated with Lindley's paradox (Lindley, 1957; Bartlett, 1957; Robert, 1993).

There are many real world situations where there is really a concentration of prior probability around particular value, and a sound Bayesian analysis should then certainly use this information. Under some conditions, those situations may well be described with a probability mass at a (measure zero) point. However, even in these cases, robustness concerns suggest that it may well be worth exploring the consequences of using a regular reference prior with the same data, if only to verify the possible dependence of the conclusions reached on the particular prior assumptions made.

Using the quadratic loss function leads to rejecting a $\boldsymbol{\theta}_0$ value whenever its Euclidean distance to $\mathrm{E}[\boldsymbol{\theta} \,|\, \boldsymbol{z}]$, the posterior expectation of $\boldsymbol{\theta}$, is sufficiently large. Observe that the use of continuous loss functions (such as the quadratic loss) permits the use in hypothesis testing of precisely the same priors

that are used in estimation, and the same prior for all values to be tested. In general, the Bayes test criterion is not invariant under one-to-one transformations. Thus, if $\phi(\boldsymbol{\theta})$ is a one-to-one transformation of $\boldsymbol{\theta}$, rejecting $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ does not generally imply rejecting $\phi(\boldsymbol{\theta}) = \phi(\boldsymbol{\theta}_0)$. Once more, invariant Bayes test procedures are available by using invariant loss functions.

The threshold constant ℓ_0, which is used to decide whether or not an expected loss is too large, is part of the specification of the decision problem, and should be context-dependent. However, as shown below, a judicious choice of the loss function leads to calibrated expected losses, where the relevant threshold constant has an immediate, operational interpretation.

2 The Intrinsic Divergence Loss

Conditional on model $\mathcal{M}_{\boldsymbol{z}}$=$\{p(\boldsymbol{z} \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda}), \boldsymbol{z} \in \boldsymbol{\mathcal{Z}}, \boldsymbol{\theta} \in \boldsymbol{\Theta}, \boldsymbol{\lambda} \in \boldsymbol{\Lambda}\}$, the required loss function $\ell\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\}$ should describe, in terms of the unknown parameter values $(\boldsymbol{\theta}, \boldsymbol{\lambda})$ which have generated the available data, the loss to be suffered if, working with model $\mathcal{M}_{\boldsymbol{z}}$, the value $\boldsymbol{\theta}_0$ were used as a proxy for $\boldsymbol{\theta}$. It may naïvely appear that what is needed is just some measure of the discrepancy between $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}$. However, since all parameterizations are arbitrary, what is really required is some measure of the discrepancy between the *models* labelled by $\boldsymbol{\theta}$ and by $\boldsymbol{\theta}_0$. By construction, such a discrepancy measure will be independent of the particular parameterization used. Robert (1996) coined the word *intrinsic* to refer to those model-based loss functions. They are always invariant under one-to-one reparameterizations.

A reasonable measure of the dissimilarity $\delta\{p_{\boldsymbol{z}}, q_{\boldsymbol{z}}\}$ between two probability densities $p(\boldsymbol{z})$ and $q(\boldsymbol{z})$ for a random vector $\boldsymbol{z} \in \boldsymbol{\mathcal{Z}}$ should surely be non-negative, zero if (and only if), $p(\boldsymbol{z}) = q(\boldsymbol{z})$ almost everywhere, and preferably symmetric. Moreover it should be invariant under one-to-one transformations of $\boldsymbol{z}$; indeed, if $\boldsymbol{y} = \boldsymbol{y}(\boldsymbol{z})$ is such a transformation and J is the appropriate Jacobian, $p_{\boldsymbol{y}} = p_{\boldsymbol{z}}/|J|$, and $q_{\boldsymbol{y}} = q_{\boldsymbol{z}}/|J|$ are expressions of precisely the same uncertainties and, therefore, one should certainly have $\delta\{p_{\boldsymbol{z}}, q_{\boldsymbol{z}}\} = \delta\{p_{\boldsymbol{y}}, q_{\boldsymbol{y}}\}$. Finally, it should also be possible to use δ to compare densities with strictly nested supports, since many approximations are precisely obtained by restricting the original support to some strict subspace. These desiderata are all satisfied by the *intrinsic discrepancy* (Bernardo and Rueda, 2002), a divergence measure which has both an information theoretical justification, and a simple operational interpretation in terms of average log-density ratios.

Definition 1 *The* intrinsic discrepancy $\delta\{p_1, p_2\}$ *between two probability distributions for the random vector* $\boldsymbol{z}$ *with densities* $p_1(\boldsymbol{z})$, $\boldsymbol{z} \in \boldsymbol{\mathcal{Z}}_1$, *and* $p_2(\boldsymbol{z})$, $\boldsymbol{z} \in \boldsymbol{\mathcal{Z}}_2$, *is*

$$\delta\{p_1, p_2\} = \min\left[\, \kappa\{p_1 \,|\, p_2\},\, \kappa\{p_2 \,|\, p_1\} \,\right] \tag{2}$$

where $\kappa\{p_j \,|\, p_i\} = \int_{\boldsymbol{\mathcal{Z}}_i} p_i(\boldsymbol{z}) \log[p_i(\boldsymbol{z})/p_j(\boldsymbol{z})]\, d\boldsymbol{z}$ *is the Kullback-Leibler* (KL) *directed logarithmic divergence of* p_j *from* p_i. *The intrinsic discrepancy between a probability distribution* p *and a family of distributions* $\mathcal{F} = \{q_i, i \in I\}$ *is the intrinsic discrepancy between* p *and the closest of them,*

$$\delta\{p, \mathcal{F}\} = \inf_{q \in \mathcal{F}} \delta\{p, q\}.$$

The intrinsic discrepancy $\delta\{p_1, p_2\}$ is the minimum average log density ratio of one density over the other, and has an operative interpretation as the minimum amount of information (in natural information units or *nits*) expected to be required to discriminate between p_1 and p_2. This may be used to define an appropriate loss function for the decision problem considered in this paper as the intrinsic discrepancy between the model, labelled by $(\boldsymbol{\theta}, \boldsymbol{\lambda})$, and the family $\mathcal{M}_0$ of models which satisfy the hypothesis to be tested:

Definition 2 *Consider* $\mathcal{M}_{\boldsymbol{z}} = \{p(\boldsymbol{z} \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda}), \boldsymbol{z} \in \boldsymbol{\mathcal{Z}}, \boldsymbol{\theta} \in \boldsymbol{\Theta}, \boldsymbol{\lambda} \in \boldsymbol{\Lambda}\}$. *The* intrinsic discrepancy loss *of using* $\boldsymbol{\theta}_0$ *as a proxy for* $\boldsymbol{\theta}$ *is the intrinsic discrepancy between the true model and the class of models*

with $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, $\mathcal{M}_0 = \{p(\boldsymbol{z} \,|\, \boldsymbol{\theta}_0, \boldsymbol{\lambda}_0), \boldsymbol{z} \in \boldsymbol{\mathcal{Z}}, \boldsymbol{\lambda}_0 \in \boldsymbol{\Lambda}\}$,

$$\ell_\delta\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda}) \,|\, \mathcal{M}_{\boldsymbol{z}}\} = \delta\{p_{\boldsymbol{z}}(\cdot \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda}), \mathcal{M}_0\} = \inf_{\boldsymbol{\lambda}_0 \in \boldsymbol{\Lambda}} \delta\{p_{\boldsymbol{z}}(\cdot \,|\, \boldsymbol{\theta}_0, \boldsymbol{\lambda}_0), p_{\boldsymbol{z}}(\cdot \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda})\}. \tag{3}$$

Notice the complete generality of Definition 2; this may be used with either discrete or continuous data models (in the discrete case, the integrals in Definition 1 will obviously be sums), and with either discrete or continuous parameter spaces of any dimensionality.

The intrinsic discrepancy loss has many attractive invariance properties. For any one-to-one reparameterization of the form $\boldsymbol{\phi} = \boldsymbol{\phi}(\boldsymbol{\theta})$ and $\boldsymbol{\psi} = \boldsymbol{\psi}(\boldsymbol{\theta}, \boldsymbol{\lambda})$,

$$\ell_\delta\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda}) \,|\, \mathcal{M}_{\boldsymbol{z}}\} = \ell_\delta\{\boldsymbol{\phi}_0, (\boldsymbol{\phi}, \boldsymbol{\psi}) \,|\, \mathcal{M}_{\boldsymbol{z}}\},$$

so that the use of this loss function will lead to estimation and hypothesis testing procedures which are *invariant* under those transformations. Moreover, if $\boldsymbol{t} = \boldsymbol{t}(\boldsymbol{z})$ is a sufficient statistic for model $\mathcal{M}_{\boldsymbol{z}}$, one may equivalently work with the marginal model $\mathcal{M}_{\boldsymbol{t}} = \{p(\boldsymbol{t} \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda}), \boldsymbol{t} \in \boldsymbol{\mathcal{T}}, \boldsymbol{\theta} \in \boldsymbol{\Theta}, \boldsymbol{\lambda} \in \boldsymbol{\Lambda}\}$ since, in that case,

$$\ell_\delta\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda}) \,|\, \mathcal{M}_{\boldsymbol{z}}\} = \ell_\delta\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda}) \,|\, \mathcal{M}_{\boldsymbol{t}}\}.$$

Computations are often simplified by using the additive property of the intrinsic discrepancy loss : if data consist of a random sample $\boldsymbol{z} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ from some underlying model $\mathcal{M}_{\boldsymbol{x}}$, so that $\boldsymbol{\mathcal{Z}} = \boldsymbol{\mathcal{X}}^n$, and $p(\boldsymbol{z} \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda}) = \prod_{i=1}^n p(\boldsymbol{x}_i \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda})$, then

$$\ell_\delta\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda}) \,|\, \mathcal{M}_{\boldsymbol{z}}\} = n\, \ell_\delta\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda}) \,|\, \mathcal{M}_{\boldsymbol{x}}\}.$$

An interesting interpretation of the intrinsic discrepancy loss follows directly from Definitions 1 and 2. Indeed, $\ell_\delta\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda}) \,|\, \mathcal{M}_{\boldsymbol{z}}\}$ is just the minimum log-likelihood ratio which may be expected under repeated sampling between the true model, identified by $(\boldsymbol{\theta}, \boldsymbol{\lambda})$, and the class of models which have $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. Thus, *the intrinsic discrepancy loss formalizes the use of the minimum average log-likelihood ratio under sampling as a general loss function*.

In particular, a suggested value $\boldsymbol{\theta}_0$ for the vector of interest should be judged to be incompatible with the observed data $\boldsymbol{z}$ if $\overline{\ell}_\delta(\boldsymbol{\theta}_0 \,|\, \boldsymbol{z})$, the posterior expectation of $\ell_\delta\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda}) \,|\, \mathcal{M}_{\boldsymbol{z}}\}$, is larger than a suitably chosen constant ℓ_0. For instance, if for some arbitrary k, $\ell_0 = \log[10^k]$, then $\boldsymbol{\theta}_0$ would be rejected whenever, given the observed data, the minimum sampling average likelihood ratio against $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, may be expected to be larger than about 10^k. Conventional choices for ℓ_0 are $\{\log 100, \log 1000, \log 10000\} \approx \{4.6, 6.9, 9.2\}$.

Under regularity conditions, the intrinsic discrepancy loss has an alternative expression which is generally much simpler to compute:

Theorem 1 (Juárez, 2004, Sec. 2.4) *If the support of* $p(\boldsymbol{z} \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda})$ *is convex for all* $(\boldsymbol{\theta}, \boldsymbol{\lambda})$*, then the intrinsic discrepancy loss may also be written as*

$$\ell_\delta\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda}) \,|\, \mathcal{M}_{\boldsymbol{z}}\} = \min\left[\inf_{\boldsymbol{\lambda}_0 \in \boldsymbol{\Lambda}} \kappa\{\boldsymbol{\theta}_0, \boldsymbol{\lambda}_0 \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda}\},\ \inf_{\boldsymbol{\lambda}_0 \in \boldsymbol{\Lambda}} \kappa\{\boldsymbol{\theta}, \boldsymbol{\lambda} \,|\, \boldsymbol{\theta}_0, \boldsymbol{\lambda}_0\}\right], \tag{4}$$

where $\kappa\{\boldsymbol{\theta}_j, \boldsymbol{\lambda}_j \,|\, \boldsymbol{\theta}_i, \boldsymbol{\lambda}_i\}$ *is the KL-divergence of* $p_{\boldsymbol{z}}(\cdot \,|\, \boldsymbol{\theta}_j, \boldsymbol{\lambda}_j)$ *from* $p_{\boldsymbol{z}}(\cdot \,|\, \boldsymbol{\theta}_i, \boldsymbol{\lambda}_i)$.

When there is no danger of confusion, $\mathcal{M}_{\boldsymbol{z}}$ may be dropped from the notation and $\ell_\delta\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda}) \,|\, \mathcal{M}_{\boldsymbol{z}}\}$ may be written $\ell_\delta\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\}$, but the dependence on the model of intrinsic losses should always be kept in mind.

In the important case of a multivariate normal model with known covariance matrix, the intrinsic discrepancy loss is proportional to the Mahalanobis distance:

Example 1 (Multivariate normal model). Let $\boldsymbol{z} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ be a random sample from a k-variate normal distribution $\mathrm{N}(\boldsymbol{x} \,|\, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ with known covariance matrix $\boldsymbol{\Sigma}$. The KL divergence of $\mathrm{N}(\boldsymbol{x} \,|\, \boldsymbol{\mu}_j, \boldsymbol{\Sigma})$ from $\mathrm{N}(\boldsymbol{x} \,|\, \boldsymbol{\mu}_i, \boldsymbol{\Sigma})$ is $\kappa\{\boldsymbol{\mu}_j \,|\, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}\} = \frac{1}{2}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^t \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$. Since this is symmetric, and the intrinsic discrepancy is additive,

$$\delta\{\boldsymbol{\mu}_0, \boldsymbol{\mu} \,|\, \boldsymbol{\Sigma}\} = \frac{n}{2}\,(\boldsymbol{\mu}_0 - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}),$$

which is $n/2$ times the Mahalanobis distance between $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}$.

2.1 Approximations

Under regularity conditions, the result of Example 1 may be combined with conventional asymptotic results to obtain large sample approximations to intrinsic discrepancy losses:

Theorem 2 (Bernardo, 2011) *Let data $\boldsymbol{z} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ consist of a random sample from $p(\boldsymbol{x} \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda})$, let $F(\boldsymbol{\theta}, \boldsymbol{\lambda})$ be the corresponding Fisher matrix, and let $V(\boldsymbol{\theta}, \boldsymbol{\lambda}) = F^{-1}(\boldsymbol{\theta}, \boldsymbol{\lambda})$ be its inverse. Then, for large n and under conditions for asymptotic normality,*

$$\ell\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda}) \,|\, \mathcal{M}_{\boldsymbol{z}}\} \approx \frac{n}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^t V^{-1}_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{\theta}, \boldsymbol{\lambda})(\boldsymbol{\theta} - \boldsymbol{\theta}_0),$$

where $V_{\boldsymbol{\theta}\boldsymbol{\theta}}$ is the submatrix of $V(\boldsymbol{\theta}, \boldsymbol{\lambda})$ which corresponds to the vector of interest $\boldsymbol{\theta}$.

The invariance of the intrinsic discrepancy loss under reparameterization may be exploited to improve the approximation above, by simply choosing a parameterization where the asymptotic convergence to normality is faster. The following result is a one-parameter example of this technique, which makes use of the variance stabilization transformation.

Theorem 3 (Bernardo, 2005b) *Let $\boldsymbol{z} = \{x_1, \ldots, x_n\}$ be a random sample of size n from model $p(x \,|\, \theta)$, and let $\tilde{\theta}_n = \tilde{\theta}_n(\boldsymbol{z})$ be an asymptotically sufficient consistent estimator of θ, whose sampling distribution is asymptotically normal with standard deviation $s(\theta)/\sqrt{n}$. Define $\phi(\theta) = \int^{\theta} s(y)^{-1} dy$. Then,*

$$\ell\{\theta_0, \theta \,|\, \mathcal{M}_{\boldsymbol{z}}\} = \frac{n}{2}\,[\phi(\theta_0) - \phi(\theta)]^2 + o(1).$$

3 Reference Analysis

Foundations indicate that the prior distribution should describe available prior knowledge. In many situations however, either the available prior information on the quantity of interest is too vague or too complex to warrant the effort required to formalize it, or it is too subjective to be useful in scientific communication. An “objective" procedure is therefore often required, where the prior function is intended to describe a situation where there is no relevant information about the quantity of interest. Objectivity is an emotionally charged word, and it should be explicitly qualified whenever it is used. No statistical analysis is really objective, since both the experimental design and the model assumed have very strong subjective inputs. However, frequentist procedures are often branded as “objective" just because their conclusions are only conditional on the model assumed and the data obtained. Bayesian methods where the prior function is directly derived from the assumed model are objective in is this limited, but precise sense. For lively discussions of this, and related issues, see Bernardo (1997), Berger (2006), and ensuing discussions.

There is a vast literature devoted to the formulation of objective priors; relevant pointers are included in Bernardo and Smith (1994, Sec. 5.6), Kass and Wasserman (1996), Datta and Mukerjee (2004), Bernardo (2005a), Berger (2006), Ghosh, Delampady and Samanta (2006), and references therein. Reference analysis, introduced by Bernardo (1979) and further developed by Berger and Bernardo (1989,

1992a,b,c), Sun and Berger (1998) and Berger, Bernardo and Sun (2009, 2011a,b), has been one of the most popular approaches for developing objective priors.

We will not repeat here arguments for reference analysis, but it may be worth synthesizing the basic definition and briefly reviewing some recent developments.

Note first that the same mathematical concepts which lie behind the definition of the intrinsic discrepancy provide the intuitive basis for the definition of reference priors. Indeed, for the one parameter model $\mathcal{M} = \{p(\boldsymbol{z} \,|\, \theta), \boldsymbol{z} \in \mathcal{Z}, \theta \in \Theta \subset \Re\}$, the intrinsic discrepancy $I\{p_\theta \,|\, \mathcal{M}\} = \delta\{p(\boldsymbol{z}, \theta), p(\boldsymbol{z})\, p(\theta)\}$ between the joint prior $p(\boldsymbol{z}, \theta)$ and the product of their marginals $p(\boldsymbol{z})\, p(\theta)$ is a functional of the prior $p(\theta)$ which measures the association between the data and the parameter and hence, the amount of information that, given prior $p(\theta)$, data $\boldsymbol{z}$ may be expected to provide about θ. If one considers k independent observations from $\mathcal{M}$ then, as k increases, $I\{p_\theta \,|\, \mathcal{M}^k\}$ will approach the *missing information* about θ which repeated sampling from $\mathcal{M}$ could provide. If $\pi_k(\theta)$ denotes the prior which maximizes $I\{p_\theta \,|\, \mathcal{M}^k\}$, the sequence $\{\pi_k(\theta)\}_{i=1}^k$ will converge to that prior function which maximizes the missing information about θ, and this is defined to be the reference prior $\pi(\theta \,|\, \mathcal{M})$.

Theorem 4 (Berger, Bernardo and Sun, 2009). *Let $\boldsymbol{z}^{(k)} = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_k\}$ denote k conditionally independent observations from $\mathcal{M}_{\boldsymbol{z}}$. Then, the reference prior is defined as an appropriate limit of*

$$\pi_k(\theta) \propto \exp\left\{ \mathrm{E}_{\boldsymbol{z}^{(k)} \,|\, \theta}\big[\log p_h(\theta \,|\, \boldsymbol{z}^{(k)}) \big] \right\} \tag{5}$$

where $p_h(\theta \,|\, \boldsymbol{z}^{(k)}) \propto \prod_{i=1}^k p(\boldsymbol{z}_i \,|\, \theta)\, h(\theta)$ is the posterior which corresponds to any arbitrarily chosen prior function $h(\theta)$ which makes the posterior proper for any $\boldsymbol{z}^{(k)}$.

Theorem 4 implies that the reference prior at a particular point θ is proportional to the *logarithmic average* of the posterior density which this point would have under repeated sampling, if this θ value were the true parameter value. The parameter values which could be expected to get relatively large asymptotic posterior densities if they were true, will then precisely be those with relatively large reference prior densities.

The result in Theorem 4 makes very simple the numerical derivation of a one-parameter reference prior. One first chooses some formal prior $h(\theta)$, maybe one for which exact or approximate posterior computation is easy, and a relatively large number of replications k. For each particular θ value whose reference prior is desired, one generates a collection $\{\boldsymbol{z}_1^{(k)}, \ldots, \boldsymbol{z}_s^{(k)}\}$ of s replications $\boldsymbol{z}_i^{(k)} = \{\boldsymbol{z}_{i1}, \ldots, \boldsymbol{z}_{ik}\}$ of size k from the original model $p(\boldsymbol{z} \,|\, \theta)$, computes the corresponding s posterior densities at θ, $\{p_h(\theta \,|\, \boldsymbol{z}_j^{(k)})\}_{j=1}^s$, and approximates the reference prior at this point by its logarithmic average,

$$\pi(\theta) \approx \exp\Big\{ \frac{1}{s} \sum\nolimits_{j=1}^{s} \log p_h(\theta \,|\, \boldsymbol{z}_j^{(k)}) \Big\}. \tag{6}$$

Under regularity conditions explicit formulae for the reference priors are readily available. In particular, if the posterior distribution of θ given a random sample of size n from $p(\boldsymbol{x} \,|\, \theta)$ is asymptotically normal with standard deviation $s(\tilde{\theta}_n)/\sqrt{n}$, where $\tilde{\theta}_n$ is a consistent estimator of θ, then the reference prior is $\pi(\theta) = s(\theta)^{-1}$. This includes as a particular case the famous Jeffreys-Perks prior (Jeffreys, 1946, independently formulated by Perks, 1947)

$$\pi(\theta) \propto i(\theta)^{1/2}, \quad i(\theta) = \mathrm{E}_{\boldsymbol{x} \,|\, \theta}[-\partial^2 \log p(\boldsymbol{z} \,|\, \theta)/\partial\theta^2]. \tag{7}$$

Similarly, if $p(\boldsymbol{x} \,|\, \theta)$ is a non regular model with a support $S(\theta)$ which depends on the paramenter in the form $S(\theta) = \{x;\, a_1(\theta) < x < a_2(\theta)\}$, where the $a_i(\theta)$'s are monotone functions of θ and $S(\theta)$ is either increasing or decreasing then, under regularity conditions (Ghosal and Samanta, 1997), the reference prior is

$$\pi(\theta) \propto \mathrm{E}_{\boldsymbol{x} \,|\, \theta}[|\partial \log p(\boldsymbol{z} \,|\, \theta)/\partial\theta|]. \tag{8}$$

In multiparameter problems, reference priors depend of the quantity of interest, a necessary feature in the construction of objective priors, if one is to prevent unacceptable behaviour in the posterior, such as marginalization paradoxes (Dawid, Stone and Zidek, 1973) or strong inconsistencies (Stone, 1976).

If the model has more than one parameter, the required joint reference prior is derived sequentially. Thus, if the model is $p(\boldsymbol{z} \,|\, \theta, \lambda)$ and θ is the quantity of interest, one works conditionally on θ and uses the one-parameter algorithm to derive the *conditional reference prior* $\pi(\lambda \,|\, \theta)$. If this is proper, it is used to obtain the *integrated model* $p(\boldsymbol{z} \,|\, \theta) = \int_\Lambda p(\boldsymbol{z} \,|\, \theta, \lambda)\, \pi(\lambda \,|\, \theta)\, d\lambda$, to which the one-parameter algorithm is applied again to obtain the *marginal reference prior* $\pi(\theta)$. The *joint reference prior* to compute the reference posterior for θ is then defined to be $\pi(\lambda \,|\, \theta)\, \pi(\theta)$. If $\pi(\lambda \,|\, \theta)$ is not proper, one proceeds similarly within a compact approximation to the parameter space (where all reference priors will be proper) and then derives the corresponding limiting result.

In general, reference priors are sequentially derived with respect to an ordered parameterization. Thus, given a model $\mathcal{M}_{\boldsymbol{z}} = \{p(\boldsymbol{z} \,|\, \boldsymbol{\omega}), \boldsymbol{z} \in \boldsymbol{\mathcal{Z}}, \boldsymbol{\omega} \in \boldsymbol{\Omega}\}$ with m parameters, the reference prior with respect to a particular ordered parameterization $\boldsymbol{\phi}(\boldsymbol{\omega}) = \{\phi_1, \ldots, \phi_m\}$ (where the ϕ_i's are ordered by inferential importance) is sequentially obtained as $\pi(\boldsymbol{\phi}) = \pi(\phi_m \,|\, \phi_{m-1}, \ldots, \phi_1) \times \cdots \times \pi(\phi_2 \,|\, \phi_1)\, \pi(\phi_1)$. Unless all reference priors turn out to be proper, the model must be endowed with an appropriate compact approximation to the parameter space $\{\boldsymbol{\Omega}_j\}_{j=1}^{\infty} \subset \boldsymbol{\Omega}$, which should remain the same for all reference priors obtained within the same model. Berger and Bernardo (1992c) describe the relevant algorithm for regular multiparameter models where asymptotic normality may be established. In typical applications, $\theta = \phi_1$ will be the quantity of interest, and the joint reference prior $\pi(\boldsymbol{\phi})$, which is often denoted $\pi_\theta(\boldsymbol{\phi})$ to emphasize the role of θ, is a just a technical device to produce the desired one-dimensional marginal reference posterior $\pi(\theta \,|\, \boldsymbol{z})$ of the quantity of interest.

4 Objective Bayesian Hypothesis Testing

With the loss function chosen to be the intrinsic discrepancy loss, all that is required to define an objective Bayesian testing procedure is to specify an objective prior distribution. It will not come as a surprise that we recommend the use of a reference prior. Thus, one must obtain the posterior expectation of the intrinsic discrepancy loss with respect to the appropriate joint reference posterior

$$d(\boldsymbol{\theta}_0 \,|\, \boldsymbol{z}) = \int_{\boldsymbol{\Theta}} \int_{\boldsymbol{\Lambda}} \ell_\delta\{\boldsymbol{\theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda}) \,|\, \mathcal{M}_{\boldsymbol{z}}\}\, \pi(\boldsymbol{\theta}, \boldsymbol{\lambda} \,|\, \boldsymbol{z})\, d\boldsymbol{\theta} d\boldsymbol{\lambda}. \tag{9}$$

and decide whether or not this is big enough to reject that $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. The function $d(\boldsymbol{\theta}_0 \,|\, \boldsymbol{z})$ is the relevant *intrinsic* test statistic, a direct measure of the incompatibility of $\boldsymbol{\theta}_0$ with the data $\boldsymbol{z}$ in terms of the expected average log-likelihood ratio against the null.

In one parameter problems, the reference prior is unique and the solution is therefore conceptually immediate. The following toy example is intended to illustrate the general procedure:

Example 2 (Poisson data). Let $\boldsymbol{z} = \{x_1, \ldots, x_n\}$ be a random sample from a Poisson model, so that $p(x \,|\, \lambda) = \mathrm{Po}(x \,|\, \lambda) = e^{-\lambda}\lambda^x / x!$. This is a regular model, and using (7), the reference prior is immediately found to be $\pi(\lambda) = \lambda^{-1/2}$. This leads to the gamma reference posterior $\pi(\lambda \,|\, \boldsymbol{z}) = \pi(\lambda \,|\, t, n) = \mathrm{Ga}(\lambda \,|\, t + 1/2, n) \propto e^{-n\,\lambda}\lambda^{t-1/2}$, with $t = \sum_{j=1}^n x_j$.

Using Definition 2 and the additive property of the intrinsic discrepancy, the intrinsic discrepancy loss of using λ_0 as a proxy for λ with a random sample of size n from a Poisson distribution with parameter λ is

$$\delta\{\lambda_0, \lambda \,|\, \mathcal{M}_{\boldsymbol{z}}\} = n\, \delta\{\lambda_0, \lambda \,|\, \mathcal{M}_x\} = n\, \min\left\{ \mathrm{E}_{x \,|\, \lambda}\left[\log \frac{\mathrm{Po}(x \,|\, \lambda)}{\mathrm{Po}(x \,|\, \lambda_0)} \right], \mathrm{E}_{x \,|\, \lambda_0}\left[\log \frac{\mathrm{Po}(x \,|\, \lambda_0)}{\mathrm{Po}(x \,|\, \lambda)} \right] \right\}$$

which immediately yields

$$\delta\{\lambda_0, \lambda \,|\, \mathcal{M}_{\boldsymbol{z}}\} = \begin{cases} n\,(\lambda - \lambda_0 + \lambda_0 \log \frac{\lambda_0}{\lambda}) & \text{if} \quad \lambda_0 \le \lambda, \\ n\,(\lambda_0 - \lambda + \lambda \log \frac{\lambda}{\lambda_0}) & \text{if} \quad \lambda_0 \ge \lambda. \end{cases} \tag{10}$$

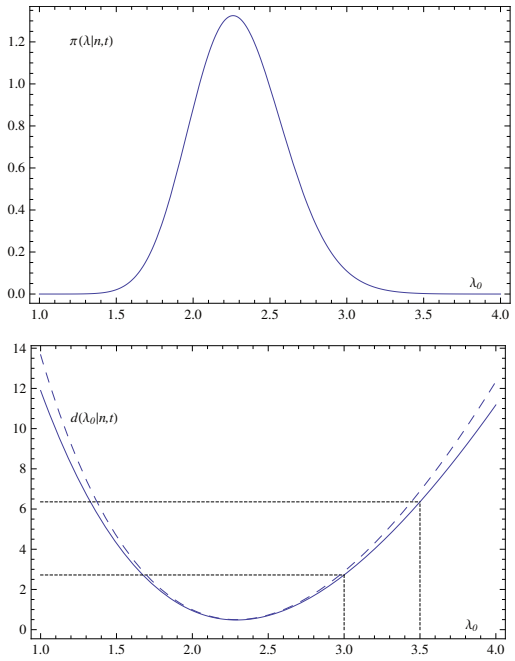


Figure 1: Posterior reference analysis of the parameter of a Poisson model.

a non-negative concave function of λ and λ_0, with minimum equal to zero when $\lambda = \lambda_0$.

The intrinsic statistic $d(\lambda_0 \,|\, \boldsymbol{z})$ is the corresponding reference posterior expectation,

$$d(\lambda_0 \,|\, \boldsymbol{z}) = \int_0^\infty \delta\{\lambda_0, \lambda \,|\, \mathcal{M}_{\boldsymbol{z}}\} \, \mathrm{Ga}(\lambda \,|\, t + 1/2, n) \, d\lambda. \tag{11}$$

This has no simple analytical expression, but is easily computed by numerical integration. Moreover, using Theorem 3 and the fact that the sampling distribution of the sufficient and consistent mle, $\hat{\lambda} = \bar{x}$ is asymptotically normal with standard deviation $\sqrt{\lambda}/\sqrt{n}$, one finds

$$\begin{aligned} d(\lambda_0 \,|\, \boldsymbol{z}) &\approx \int_0^\infty \frac{n}{2} \left(2\sqrt{\lambda_0} - 2\sqrt{\lambda}\right)^2 \mathrm{Ga}(\lambda \,|\, t + 1/2, n) \, d\lambda \\ &= 1 + 2\,t + 2n\,\lambda_0 - 4\sqrt{n\,\lambda_0} \, \frac{\Gamma(t+1)}{\Gamma(t+1/2)} \qquad (12) \\ &\approx 1 + 2\,t + 2n\,\lambda_0 - 4\sqrt{n\,\lambda_0} \left(\sqrt{t} + \frac{1}{8\sqrt{t}}\right). \qquad (13) \end{aligned}$$

To illustrate the type of results obtained, a sample of size $n = 25$ was simulated from a Poisson distribution with parameter $\lambda = 2$ resulting in $t = 57$. The corresponding reference posterior density is plotted in the top panel of Figure 1. The expected intrinsic discrepancy loss $d(\lambda_0 \,|\, t, n)$ (both computed from (11) (continuous line) and analytically approximated with (13) (dashed line)) are plotted in bottom panel of Figure 1. It may be appreciated that, even with this rather small sample size, the approximation is quite good.

Suppose that the value $\lambda_0 = 3$ is to be tested. The corresponding intrinsic statistic is $d(3 \,|\, \boldsymbol{z}) = 2.72 \approx \log(15)$; thus the average likelihood ratio against the $H_0 \equiv \{\lambda = 3\}$ may be expected to be about 15, not really strong evidence against this value, even if this may be seen to be in the right tail of the reference posterior of λ_0. On the other hand, if the value to test is $\lambda_0 = 3.5$, the corresponding intrinsic statistic is $d(3.5 \,|\, \boldsymbol{z}) = 6.36 \approx \log(576)$; hence the average likelihood ratio against this value may be expected to be about 576 and the hypothesis $H_0 \equiv \{\lambda = 3.5\}$ should be rejected in most scenarios. Notice that, in marked difference to conventional testing using Bayes factors, the same prior has been used to test these two possible parameter values (as it would obviously be for any other value).

The following example illustrates the use of the methods described to derive a new solution to a classical precise testing problem.

Example 3 (Equality of Normal means). Let $\boldsymbol{z} = \{\boldsymbol{x}, \boldsymbol{y}\}$ be two independent random samples, $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ from $\mathrm{N}(x \,|\, \mu_x, \sigma\}$, and $\boldsymbol{y} = \{y_1, \ldots, y_m\}$ from $\mathrm{N}(x \,|\, \mu_y, \sigma\}$, and suppose that one in interested in

comparing the two means. In particular, one may be interested in testing whether or not the precise hypothesis $H_0 \equiv \{\mu_x = \mu_y\}$ is compatible with available data $\boldsymbol{z}$. Using the additive property of the intrinsic discrepancy loss and the fact that the KL divergence between two normals distrubutions with the same variance is simply $\kappa\{\mu_j, \sigma \,|\, \mu_i, \sigma\} = (\mu_i - \mu_j)^2/(2/\sigma^2)$ to derive the logarithmic divergence of $p(\boldsymbol{z} \,|\, \mu_0, \mu_0, \sigma_0)$ from $p(\boldsymbol{z} \,|\, \mu_x, \mu_y, \sigma)$, and then minimizing over both μ_0 and σ_0 yields $\inf_{\mu_0 \in \Re,\, \sigma_0 > 0} \kappa\{\mu_0, \mu_0, \sigma_0 \,|\, \mu_x, \mu_y, \sigma\} = k_{nm}\ \theta^2$, where $k_{nm} = 2nm/(m+n)$ is the harmonic mean of the two sample sizes, and $\theta = (\mu_x - \mu_y)/\sigma$ is the standardized difference between the two means. On the other hand, $\inf_{\mu_0 \in \Re,\, \sigma_0 > 0}\ \kappa\{\mu_x, \mu_y, \sigma \,|\, \mu_0, \mu_0, \sigma_0\}$ yields $[(m+n)/2]\, \log[1 + (k_{nm}/(2(m+n)))\, \theta^2]$, which is always smaller. Hence, the intrinsic discrepancy loss of accepting H_0 is

$$\ell_\delta\{H_0, (\mu_x, \mu_y, \sigma)\} = \ell_\delta\{H_0, \theta \,|\, \mathcal{M}\} = \frac{n+m}{2}\ \log\Big[1 + \frac{k_{nm}}{2(n+m)}\ \theta^2\Big],$$

which reduces to $n \log[1 + \theta^2/4]$ when $n = m$. Here, the parameter of interest is θ. Bernardo and Pérez (2007) find that the marginal reference posterior of θ only depends on the data through the sample sizes and $t = t(\boldsymbol{z}) = (\bar{x} - \bar{y})/(s/\sqrt{2/k_{nm}})$, where s is the m.l.e. of σ. Therefore, the required marginal reference posterior of θ is $\pi(\theta \,|\, \boldsymbol{z}) = \pi(\theta \,|\, t, m, n) \propto p(t \,|\, \theta)\, \pi(\theta)$ where $p(t \,|\, \theta)$ is the noncentral Student sampling distribution of t, and $\pi(\theta) = (1 + (k_{nm}/(4(m+n))\, \theta^2)^{-1/2}$ is the marginal reference prior for θ. The posterior $\pi(\theta \,|\, t, m, n)$ may be used to provide point and interval estimates of θ, the standardized difference between the two means, and hence inferential statements about their relative positions.

The relevant expected loss, $d(H_0 \,|\, t, n, m) = \int_{-\infty}^{\infty} \ell_\delta\{H_0, \theta \,|\, \mathcal{M}\}\, \pi(\theta \,|\, t, n, m)\, d\theta$, may be used to test H_0. This has no simple analytical expression, but its value may easily be obtained by one-dimensional numerical integration. A good large sample approximation is

$$d(H_0 \,|\, t, n, m) \approx \frac{n+m}{2} \log\Big[1 + \frac{1}{n+m}(1 + t^2)\Big].$$

The sampling distribution of $d(H_0 \,|\, t, n, m)$ is asymptotically $(1/2)[1 + \chi^2_1(\lambda)]$, where $\chi^2_1(\lambda)$ is a non-central chi-squared distribution with one degree of freedom and non-centrality parameter $\lambda = k_{nm}\theta^2/2$. If follows that the expected value under sampling of $d(H_0 \,|\, t, n, m)$ is equal to one when $\mu_x = \mu_y$, and increases linearly with the harmonic mean of the samples when this is not true. Thus, the testing procedure is consistent.

For many more sophisticated examples of precise hypothesis testing, see Bernardo (2011) and references therein.

References

Bartlett, M. (1957). A comment on D. V. Lindley's statistical paradox. *Biometrika* **44**, 533–534.

Berger, J. O. (2006). The case for objective Bayesian analysis. *Bayesian Analysis* **1**, 385–402 and 457–464 (with discussion).

Berger, J. O. and Bernardo, J. M. (1989). Estimating a product of means: Bayesian analysis with reference priors. *J. Amer. Statist. Assoc.* **84**, 200–207.

Berger, J. O. and Bernardo, J. M. (1992a). Ordered group reference priors with applications to a multinomial problem. *Biometrika* **79**, 25–37.

Berger, J. O. and Bernardo, J. M. (1992b). Reference priors in a variance components problem. *Bayesian Analysis in Statistics and Econometrics* (P. K. Goel and N. S. Yyengar, eds.) Berlin: Springer, 323–340.

Berger, J. O. and Bernardo, J. M. (1992c). On the development of reference priors. *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) Oxford: University Press, 35–60 (with discussion).

Berger, J. O., Bernardo, J. M. and Sun, D. (2009). The formal definition of reference priors. *Ann. Statist.* **37**, 905–938.

Berger, J. O., Bernardo, J. M. and Sun, D. (2011a). Reference priors for discrete parameters. *J. Amer. Statist. Assoc.* (under revision).

Berger, J. O., Bernardo, J. M. and Sun, D. (2011b). Overall reference priors. *Tech. Rep.*, Duke University, USA.

Bernardo, J. M. (1979). Reference posterior distributions for Bayesian inference. *J. Roy. Statist. Soc. B* **41**, 113–147 (with discussion). Reprinted in *Bayesian Inference* (N. G. Polson and G. C. Tiao, eds.) Brookfield, VT: Edward Elgar, 1995, 229–263.

Bernardo, J. M. (1997). Noninformative priors do not exist *J. Statist. Planning and Inference* **65**, 159–189 (with discussion).

Bernardo, J. M. (2005a). Reference analysis. *Bayesian Thinking: Modeling and Computation, Handbook of Statistics* **25** (Dey, D. K. and Rao, C. R., eds.) Amsterdam: Elsevier, 17–90.

Bernardo, J. M. (2005b). Intrinsic credible regions: An objective Bayesian approach to interval estimation. *Test* **14**, 317–384 (with discussion).

Bernardo, J. M. (2011). Integrated objective Bayesian estimation and hypothesis testing. *Bayesian Statistics 9* (J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West, eds.) Oxford University Press, 1–68, (with discussion).

Bernardo, J. M. and Rueda, R. (2002). Bayesian hypothesis testing: A reference approach. *Internat. Statist. Rev.* **70**, 351–372.

Bernardo, J. M. and Pérez, S. (2007). Comparing normal means: New methods for an old problem. *Bayesian Analysis* **2**, 45–58.

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory*. Chichester: Wiley.

Datta, G. S. and Mukerjee, R. (2004). *Probability Matching Priors: Higher Order Asymptotics*. Berlin: Springer.

Dawid, A. P., Stone, M. and Zidek, J. V. (1973). Marginalization paradoxes in Bayesian and structural inference. *J. Roy. Statist. Soc. B* **35**, 189–233 (with discussion).

Ghosh, J. K., Delampady, M. and Samanta, T. (2006). *An Introduction to Bayesian Analysis: Theory and Methods*. Berlin: Springer.

Ghosal, S. and Samanta , T. (1997). Expansion of Bayes risk for entropy loss and reference prior in nonregular cases. *Statistics and Decisions* **15**, 129–140.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proc. Royal Soc.* **186**, 453–461.

Jeffreys, H. (1961). *Theory of Probability* (3rd edition). Oxford: University Press.

Juárez, M. A. (2004). *Métodos Bayesianos Objetivos de Estimación y Contraste de Hipótesis*. Ph.D. Thesis, Universitat de València, Spain.

Kass, R. E. and Wasserman, L. (1996). The selection of prior distributions by formal rules. *J. Amer. Statist. Assoc.* **91**, 1343–1370.

Lindley, D. V. (1957). A statistical paradox. *Biometrika* **44**, 187–192.

Perks, W. (1947). Some observations on inverse probability, including a new indifference rule. *J. Inst. Actuaries* **73**, 285–334 (with discussion).

Robert, C. P. (1993). A note on Jeffreys-Lindley paradox. *Statistica Sinica* **3**, 601–608.

Robert, C. P. (1996). Intrinsic loss functions. *Theory and Decision* **40**, 192–214.

Stone, M. (1976). Strong inconsistency from uniform priors. *J. Amer. Statist. Assoc.* **71**, 114–125 (with discussion).

Discussion with José Bernardo on Bayesian reference analysis

Transcribed and edited by Luc Demortier (Rockefeller University)

Abstract
The discussion session that followed José Bernardo's talk was devoted to Bayesian reference analysis. Questions were asked about the use of sharp priors in Bayesian hypothesis tests, the relationship between objective Bayesian and frequentist testing, the construction, computation, and interpretation of reference posteriors, the combination of measurements, and sensitivity analysis. As the discussion unfolded, comments were made about more general issues such as the hierarchy of problem formulations, and somewhat esoteric topics such as the ratio of normal means.

Luc Demortier

Let's first take questions about José Bernardo's talk, and then later we'll open the floor to more general questions about reference priors and reference analysis.

Kyle Cranmer (New York University)

I have a question about invariance under reparametrization. I can specify a model parametrized, say, in θ, or in α. There is the procedure that you follow and you have all the various invariance properties. But what I am confused about is that, if you need to specify this ordering to construct the reference prior, what does it even mean to talk about invariance under reparametrization? Isn't that meaningless, especially if the transformation from θ to α is "weird"? So how do you talk about invariance, when you need ordering?

José Bernardo (University of Valencia)

Two answers. First of all, the real differences are hierarchical, in that it is important to select what the parameter of interest is. There are slight differences in the results depending on the ordering of the other parameters, but it is really the specified parameter of interest which matters, and the whole thing is defined in such a way that it is invariant under one-to-one transformations of this parameter of interest.

Apart from that I agree that often you have several parameters of interest, you can decide how many. Then you can try to get a global approximation, a reference prior which will not be *the* reference prior for each of these things as the parameter of interest, but will be such that the marginals of the joint posterior that you will get will not be far off of the corresponding reference posteriors. We have found in many examples that this actually works, the differences are so minute that when you have ten observations you don't see them. So in practical terms it's not that important. Of course if you have very few parameters then it does matter, but then it should matter. For instance if you are in a normal situation, with parameters μ and σ, and you are not interested in either μ or σ, but in μ/σ, so that the parameter of interest is the signal to noise ratio, then it does matter. In that case it doesn't make sense to use the global approximation because you can get the analytical answer for the quantity you want. However, in many-parameter problems with many quantities of interest, you'd better do something like a global approximation because otherwise the whole thing becomes unmanageable.

Glen Cowan (Royal Holloway)

When you look at what you call the objective Bayesian hypothesis testing and you write down this intrinsic test statistic, this d of H_0 given the data, is there an easy way of seeing how a test based on that statistic maps onto a classical test in terms of its power?

José Bernardo

Yes, this may be done. In fact, under asymptotic conditions, it's very easy, because if you have enough regularity to get a posterior distribution which is asymptotically normal, then what happens is that you are in an situation much like testing $\mu = \mu_0$ in a normal distribution. And in this particular example what you get is that the intrinsic statistic d is actually $(1 + t^2)/2$, where t is the number of standard deviations out. In asymptotic conditions, this will be approximately true in any problem, so if $t = 3$, then $d = 5$. Thus $d = 5$ is exactly equivalent to three standard deviations in the normal conditions. Of course this will be only an approximation if the data are not normal, but it will be a good approximation if the posterior of the quantity of interest is asymptotically normal, and you will get an immediate feeling for how it works. Actually you do get that d, the intrinsic statistic, is very often a (sometimes approximate) one-to-one transformation of some already known statistic. But the Bayesian analysis gives you an interpretation which is immediate and unique in terms of expected log-likelihood ratios, and you don't have to find the sampling distribution of anything. With the same argument, the often suggested five standard deviations gives $d = 13$, so if you get an intrinsic statistic around thirteen, it is as if you were five standard deviations off in the normal case.

Frederik Beaujean (Max Planck Institute for Physics)

I want to ask about your theorem 1 where you showed the example of how one can numerically construct the reference prior in one-dimensional problems. You said for large k this is good. Do you have any idea when k is large enough? If I have a few data points, is k large or not large?

José Bernardo

Well, I think this must be problem-dependent, but k around 1000 will often suffice. I have played with this in many different applications, like election forecasting and industrial quality control, and then I have always found that for the sample sizes that people use, this kind of approximation to the reference prior is more than enough. However, to be sure you would really have to do the numbers and see, by simulation probably, how far you would be. It should be possible to prove general things for conditions that guarantee that the approximation is good but, as far as I know, that has not been done.

Glen Cowan

So this is a point that went by quickly and I just wanted to make sure I understand this. You are saying that putting a sharp prior is bad?

José Bernardo

No, not necessarily bad, I am saying it is different, and that using a sharp prior might be dangerous

in specific circumstances, that you should check. Putting a sharp prior is just one specific Bayesian solution, which is *not* non-informative, because by definition you are selecting a prior which concentrates on one point, and that's certainly not non-informative. Sharp priors *are* informative, by definition. To consider whether or not you want to use a sharp prior or a "flattish" one, just think in one dimension. There is obviously a huge difference between having a kind of flattish prior over the real line and a prior concentrated near one point. What I am saying is that, if you are interested in testing whether or not $\theta = 3$, and you believe, you really believe (and you want to make this part of the analysis) that the true value of θ is either 3 or close to 3, then of course you should use a prior that reflects this. And then you could do all that I have done, without any change. Now, this is complicated in some cases. A simple approximation is to replace this kind of presumably continuous prior by a point mass at $\theta = 3$ and something else around. And that is an approximation. You always have to be careful, because it is known to be an approximation, and it is known that under some conditions this approximation breaks down. So you have to check that you are not in those conditions. You have to check both that you are interested in a prior distribution which is concentrated on something, and you really want answers that depend on that, plus that the approximation is going to be all right mathematically, which it usually will be. The ESP example is a real example, but it is special because you have a hundred million observations. In general, the differences are not that large. But I insist that if you want a hypothesis testing procedure that does not depend on a prior distribution concentrated on the null value, then you have objective Bayesian alternatives, by just using a continuous loss and a continuous prior.

Glen Cowan

It seemed that, in the ESP example, when you put a point mass at 50%, then in some sense you got the answer that you didn't want, and I guess what confused me is you said that saying you should have a point mass at 50% is in some sense an approximation, whereas it seems to me that, no, that's exactly what you mean by ESP not existing. But I guess what you are saying is that in a real experiment, that you have perhaps a lack of ESP, but in addition some small bias, and so the more exact prior therefore includes a little bit of variation around 0.5 in order to account for the inevitable bias.

José Bernardo

In the ESP example, if rather than putting a probability mass at 0.5 (50% or any other), you use a continuous prior distribution centered on 0.5, then, unless you specify a hugely (really hugely) concentrated prior, then you will get precisely the same answer I did, because this prior will be actually wiped out by the data anyway. It will give practically the same posterior, and therefore you will see that the data suggest that the true value is *not* exactly 0.5, but about 0.5002 or whatever. Now, whether this is because of ESP, the mind being able to move about 0.02% of the particles, or more likely to my taste, because there is a small bias in the machine, well, it's not statistics that is going to determine this.

Diego Casadei (New York University)

When we choose the Bayes factor to make a decision, we have seen from Jim Berger's talk this morning what happens, and you also made a caveat: there are conditions in which this may give surprises. To avoid those problems, what is the alternative? Shall we consider the ratio of posteriors between the two hypotheses?

José Bernardo

Bayes factors by themselves are not a choice. Their use is a consequence of the fact that if one uses a sharp type of prior, the posterior probability of the null depends on the data only through the Bayes factor. This is just a mathematical fact. But if you do not do this sharp prior approximation, this is no longer true. So it's not that Bayes factors are good or not, it's just that they do not appear as the relevant quantities to compute. What appears in place of the Bayes factor is the kind of expression that I have mentioned: you get a continuous posterior distribution. The simplification that the posterior probability of the null depends on the data only through the Bayes factor occurs only if you have a sharp prior. And that has the implications that if you don't believe this sharp prior to be sensible for one reason or another, then of course you should not be using the corresponding result. A lot of people tend to use only Bayes factors (rather than posterior probabilities), because of course then you don't have to argue with all the prior probabilities of the hypotheses. However, what Bayesian statistics tells you, is that you should really use posterior probabilities. Now, because of these mathematical equivalencies it is tempting to forget about it. This would be too long to discuss here, but if you insist on using only Bayes factors, as opposed to using posterior probabilities, you are going to get a number of problems, because you are easily getting outside foundations. For instance, the kind of approximate Bayes factors that you are forced to introduce may sometimes not be transitive, so you have that the Bayes factor of H_1 over H_2 times the Bayes factor of H_2 over H_3 is not the Bayes factor of H_1 over H_3, so these are not real Bayes factors. All these things may happen because all the time you are using approximations. There is nothing wrong with approximations, except that you have to realize that they are approximations, and therefore to be careful. That's all I am saying. And the advantage of using the sort of procedure that I have described is that you don't have to do that at all. I mean, you can always use the standard, continuous reference prior, precisely the same as in estimation, and I find that attractive.

Luc Demortier

We have about half an hour left before the coffee break, so maybe we should switch to a discussion of reference priors. Louis suggested that we ask professional statisticians first if they have any comments about reference priors. If that's the case, let's hear them first and then go to questions.

Jim Berger (Duke University)

I think I'd rather answer questions.

Luc Demortier

So let's open the floor to everybody then.

Harrison Prosper (Florida State University)

I hope this is not too technical, but in one of your most recent papers [1], with Jim Berger, you go through a formal definition of reference priors, and you arrive at the form that you illustrated on your slides. There is a sentence after that formula, to the effect that one no longer needs to use this compact set argument to make sure that, every step along the way, the reference prior is normalized. Did I read that correctly?

José Bernardo

If I remember correctly (I guess you refer to the recent Annals paper) what we meant is that, if you have only one parameter, it just happens that once you have created the structure, the particular choice of the compact sets doesn't matter, so the answer is unique. Unfortunately this is not true with more than one parameter. So even if you have just two parameters, we have examples where you can see that the answer that you get might depend on the particular sequence of compactification that you use, and following one of the questions that Luc has posed this morning, we don't have an answer as yet, as to what the standard procedure would be to have more or less automatic choice of the compacts. We have answers for Luc's other two questions, but not for the one about compacts. Except in one dimension. In one dimension you need the compacts to make the mathematics work, but then, the specific compact thing doesn't happen, so whatever compact set you choose, you are going to get the same answer.

Harrison Prosper

So presumably that means that we can use the numerical algorithm to calculate the reference prior, without having to worry about compact sets.

José Bernardo

In one dimension you don't have any problem. But if you do it in more than one dimension, then because it is a conditional argument, the problem will come up.

Harrison Prosper

Yes, the reason that this important is because of the sort of thing we are looking at right now, problems in which you have, say, one parameter of interest, and you may have many many nuisance parameters. And so the question is whether in that circumstance, supposing you wanted to calculate a conditional reference prior probability for θ given a whole bunch of nuisance parameters, whether in that case one has to use the compact set argument or one can just immediately calculate, for that particular parameter value of θ, using the algorithm you have in your paper.

José Bernardo

Well, this is an important research challenge, and I am not totally sure what may be done, but my attitude in that kind of problem would be to integrate out all the nuisance parameters using some approximate reference prior on a particular set of compacts, and then do a strict reference analysis with the parameter of interest with the resulting integrated model. It might be very complicated, in fact it *will* typically be very complicated, but because you have a numerical procedure you can obtain the reference posterior from this (approximate) integrated model.

Jim Berger

Harrison, tell me if I am wrong, but I think you are thinking of a situation where you have evidence-based priors for the nuisance parameters. So they actually have priors for all the nuisance parameters.

José Bernardo

Real priors?

Jim Berger

Real priors. And so his question is simply, would we trust the fact that the one-dimensional numerical algorithm works, so then we look at the problem of the one parameter of interest conditioned on the nuisance parameters, and would we need compact sets there? And I think not. I mean, because all of the examples where we needed compact sets were examples where we were doing more than one parameter at a time, and it was the interaction among them. So, barring finding an example tomorrow where it doesn't work, I'd be quite happy with evidence-based priors for all the nuisance parameters and the numerical algorithm for the parameter of interest.

Louis Lyons (Imperial College)

So, can I ask a question then? Say we are convinced that we want to use reference priors in our analyses. The only paper I knew of up till this present conference was by Luc and Harrison and Supriya. [2] If I want to do a reference prior analysis, is there somebody I can send an e-mail to, tell them what my problem is, and get a reference prior back? Or do I have to do it myself?

José Bernardo

I think the answer is that, if your problem belongs to a set of textbook examples, then yes I can give you the reference prior, or Jim, or other people, or even in the internet you can probably find a reference to a paper. If the model is not standard, then you have to do it.

Günter Zech (University of Siegen)

I have a non-expert question. If you have two experiments measuring the same parameter, and they have different acceptances, then they have different reference priors, right? When I measure the same parameter with different priors, how do you combine or compare these two results?

José Bernardo

That's a very interesting question, and one that has many sides to it. First of all I think you have to make sure that from a statistical point of view, the parameter is the same. From a Bayesian point of view, a parameter is the limit, a frequentist limit actually, of a function of the observations. So for a Bayesian, the parameter p in a binomial situation has a precise definition, namely it's the limit of the relative frequencies when n goes to infinity. Now, if you have two different models, you have to make sure that if you call θ, the same θ in two models, this θ is really the same thing. If you take one limit and you take the other limit, you get the same thing. And that's not trivial, you have actually to do it. If you have done that, then you have the same parameter.

Even then, you will get two different posteriors, even though it's the same parameter. Now these posteriors, these reference posteriors, are the answer to a "what if" question, namely what would I say about θ if my prior beliefs were as — let me get this sentence right — if my prior beliefs were those under

which the data from this particular experiment would give me the highest possible information, that's the formal definition. Now two different experiments may give you the highest possible information for two different priors. And because they are conditioned answers, they are mathematically compatible, just different answers. Whether you would use one, or the other, or a mixture, you have these two models, you still can have, say a finite mixture of the two of them, either multiplicative or additive, and you have yet a third model and a slightly different answer. And all those answers are conditional answers to the particular model.

There is no such thing as *the* objective answer, it's a conditional answer, to a particular model, and you want to be precise about the assumption that your prior is such that the information provided by the data would be maximum. And somehow often there are underlying assumptions there, for instance the best known case is the binomial, where you have a fixed number of tosses or a fixed number of successes, binomial or inverted binomial. Now, in the standard binomial case, the reference prior, which is Jeffreys' prior, is Beta(1/2,1/2), and in the other one it's Beta(0,1/2), which is different. But the point is that in the inverted binomial situation you are assuming that the probability of success is strictly positive, because otherwise you would never get the required number of successes, and in the other one you are assuming that it might be zero. This slight difference is a radical assumption, which is reflected in the prior and therefore in the posterior.

Of course, if you have a lot of data the thing is not going to matter, but for very small datasets it *is* going to matter. If you do not know what the mechanism has been, binomial or inverted binomial, or "I am just tired" (you don't really know), neither of these answers is objective. I think English people have this saying that there no such thing as a free lunch, so there is no such thing as a free answer: you have to put in assumptions. Reference analysis is very specifically tailored to give an answer to a very precise, conditional question: what could I say about the parameter if my prior was that prior that maximizes the information from the data drawn from a specified model. And for this question we have an answer which I think is useful.

As for the original question, if you have two sets of data x and y from different experiments measuring the same parameter, it should be possible to specify a single experiment from which (x, y) may be assumed to have been drawn, and then derive the corresponding reference posterior for the common parameter.

Günter Zech

If you measure for instance the lifetime of the Lambda particle, and two experiments do it, and they use two different reference priors, it's a mess afterwards, to compare these data.

José Bernardo

Well, I see it as an example of sensitivity analysis. You have to do such a sensitivity analysis with respect to change in the model, with respect to change in the prior, with respect to everything, and in a sense, the reference posteriors that you would get for all kinds of different models that you can think of, would precisely result in a robust answer, in that my posterior must be in the convex set defined by all these posteriors. And if you are not able to select more precisely which of the experiments you prefer, well, you have a range of posteriors. I don't think there is anything wrong with that.

Luc Demortier

I think you just answered the first question that was on my list. I don't know if you want to say anything more about that. So the question was, what is the correct probabilistic interpretation of

reference posteriors. A lot of my colleagues in high energy physics have this question, and they always come back to it. And related to that is, when you calculate a reference posterior, is it OK to just report that result or should you make it part of a sensitivity analysis, and if so, how do you choose other priors? Because reference priors have a special status, so is it necessary to go look for other priors, and if so, what other priors?

José Bernardo

OK first things first. The idea that you need to make reference analysis part of a global analysis, I don't think you have to, but obviously if you have the time and the resources to do it, it's always better. With respect to interpretation, I think that that is more or less clear. The interpretation is that a reference posterior is a posterior, an expression of beliefs in the parameter of interest, given the data. But as a posterior, it depends on the prior. Because of the way the prior has been constructed, it is the solution to what should I believe if I did not have any relevant information, in the precise sense defined by that prior that would maximize the information from the data. And there is of course a second answer, which is essentially asymptotic, namely that it is also true that if you give a credible interval with posterior probability 0.95, you will cover the true value, under most circumstances, with probability close to 0.95, and in some examples with exact 0.95 probability. So the two interpretations are out there. Indeed, in most cases credible intervals are approximate confidence intervals, but in some problems there are not, and for a good reason. If you insist in getting a frequentist interval for the ratio of means problem, from a frequentist point of view you get a disaster, you get the whole real line with probability 0.95. Of course you do not want to reproduce that, and indeed the Bayesian reference posterior does not reproduce that. [3] By and large, it is always true that you should be able to bet on credible intervals, if your own real priors are kind of non-informative with respect to the parameter of interest. Besides, they mostly have a rough sort of interpretation in frequentist terms; but this is not a general result, it does require conditions.

Glen Cowan

I have a comment for Günter Zech, and then a naïve question, and that is that if you had two experiments with independent data, so they are each characterized by a certain likelihood function, but they are measuring the same underlying parameter, then is it not true that there would be a reference prior for experiment A, a reference prior for experiment B, but in addition if you were simply to consider both experiments together, they are characterized by a single likelihood function which is the product of the two, so there is a unique reference prior that characterizes in some sense the combined experiment. Would that not solve your problem, Günter?

Günter Zech

You would have to publish the acceptances of the two experiments because the reference prior is not simply a function of the likelihoods.

Glen Cowan

Ah, how you publish is another problem! I think the thing is, if you really want to get together and compare results and combine results, that cannot be done if each guy shows up with his own posterior distribution, you have to take a step back. I guess we knew that already.

David Cox (Nuffield College, Oxford)

I'd like to make a couple of comments and just a historical point. I find the ideas of reference priors highly attractive, but... And the but is largely an issue of interpretation. It's the posterior I would get if this was my prior, but I don't want it, it's not my prior normally. And in particular if it's improper... That's the issue of interpretation.

The more technical question concerns the issue that Harrison raised, I think really, that if the number of nuisance parameters is large, what's going on? The asymptotics can be taken to be that as the number of observations tends to infinity, so does the number of nuisance parameters. Now I know you have considered one of Charles Stein's examples, where you get a nice answer from the reference priors, a correct answer about non-central chisquared and so forth. But I am unclear whether your results apply more generally than that. If we have a problem in which the number of nuisance parameters is quite large in some sense, do things tend to go wrong?

Another issue where I do very strongly disagree with the answers you get, which you mentioned about a minute or so ago, is about the ratio of two normal means, where I think the traditional answer is the right one, and the answer you give is not, from a purely common sense point of view.

The historical point is, just by chance I mentioned Renyi this morning. Renyi did give an axiomatization of probability in which improper priors where allowed, in the spirit of Kolmogorov, and Kolmogorov did approve this in some sense. It might possibly be interesting to go back to that. That was, I shudder to think, sixty years ago that he gave this.

Jim Linnemann (Michigan State University)

I'd like to provoke a little more conversation amongst the statisticians, because for me coming here is certainly about learning about their views on these issues. What we physicists have always wanted from statisticians, is to give us a unique answer. We have gotten used to the fact that we are not going to get a unique answer out of discussions of what Bayesian techniques can do, and what frequentist techniques can do. But maybe, at least the (objective) Bayesians could give us a unique answer. There are two kinds of uniqueness. One is mathematical uniqueness, for a sufficiently well-posed problem. The other is practical uniqueness: something sufficiently convincing that it sweeps the field. So my question is: there are many attractive features to this procedure, but why isn't everyone using it? And if all statisticians aren't convinced, why should we use it?

Louis Lyons

So, any Bayesian answers?

Jim Berger

Another sort of question, or answer of uniqueness, is in the cases we have been talking about where the reference prior is not unique, how different are the answers? And it will be negligibly different. I mean in the cases that we talked about, where there are two different experiments, if both people use one reference prior or both use another, the answers are not going to be distinguishable, in a confidence limit sense. So the answer will be, not exactly unique, but more or less the scientific conclusion will be unique. So if the choice of the reference prior doesn't matter that much, why do we do them? Well, it's because there are other priors, like vague, constant priors and what not, that we know are dangerous, that they may work well, but we know that there are scenarios where they just collapse. We don't know scenarios, well we know of only one, where a reference prior doesn't work, in the sense of giving a sensible answer,

that's usually very close to a frequentist answer, except in the problems where it's impossible to do both, like the mu over sigma problem. And I would argue with David that it's not necessarily possible to say that the unknown Poisson mean is between 0 and infinity with probability 0.95. I mean the frequentist answer in that case, to get a genuine frequentist answer, sometimes there are theorems that say you have to say the whole real set with probability one minus alpha. Oh, you have other frequentist answers? We'll talk about that offline. But there are problems where the frequentist and the Bayesian simply can't agree, because of conditioning issues, and in those problems there is, you know, a serious debate as to what you should do. But in the ones where the frequentist and Bayesian can reach agreement, it seems like reference priors get to that agreement better than anything.

David Cox

I would say that the answer to the crucial question about uniqueness is, how deep a formulation are you willing to give of a problem? I mean, problems come in many different kinds, and some have only weak specification. The ideal situation is that you have your priors based on evidence, you have your model, you have your likelihood, and then the Bayesian solution, clearly, is the right one. And nobody, as far as I know, has ever argued with that. The issue is, how far along that route of specification are you willing to go? And the reference priors are a very beautiful attempt, and successful, perhaps to a large extent successful, of evading part of that question by saying, well no we can't look down a very specific, probabilistic statement of the nature of the evidence external to the data. So we do something else instead.

Frequentist statisticians of course don't disregard external information, they simply say, we can't formulate it probabilistically, so we have to incorporate it qualitatively, with a confidence interval or significance test or whatever. And the weakest form of such a procedure is the simple significance test, which I talked mostly about this morning, where the formulation is extremely weak, it's just a null hypothesis and an indication of which direction you are looking in, nothing else is formulated probabilistically. And if that's all as far as you can go quantitatively, that's as far as you can go. And it's far from an ideal situation. Nobody, I think, really likes it too much. And then, you may have an idea of an alternative, you may have a detailed model for the alternative, confidence intervals, or posterior intervals, you may have your prior distribution and so forth, there is this hierarchy.

R.A. Fisher, who really invented most of the more traditional, the non-Bayesian side of the subject, repeatedly emphasized, especially in his last book, that there are hierarchies of formulation, including — he very explicitly, and he used occasionally — Bayesian formulations, when he thought they were appropriate in genetical problems. And he emphasized, there really might well be forms that are not already discovered. And speaking purely personally, I first learned about statistics from Jeffreys, and Jeffreys' work is very much extensive about that. I find that appealing, but I have my reservations, as I tried to indicate.

Nikolai Krasnikov (Institute for Nuclear Research, Moscow)

For some priors Bayesian results coincide with frequentist results, for instance upper limits for Poisson distributions. So maybe they use these priors as the best priors, which coincide with frequentist results.

David Cox

I have said more than I intended to already, but on this issue of the ratio of the means, the point is that, supposing we have an answer 0.5 with a standard error of 5 in the numerator, and 0.5 with a

standard error of 5 in the denominator, the frequentist answer is, any value of the ratio is consistent with the data. Now Neyman, who strongly emphasized taking the confidence interval interpretation very very literally, never resolved the issue. And if you start to talk about 95% confidence intervals, then you are in the trouble that Jim emphasizes. But I don't regard that as the frequentist answer. The frequentist answer is that, at any level of significance, up to some maximum, or down to some minimum, any value might be consistent with the data. And in another context, it may be that any value outside certain intervals is consistent with the data. And the data tell you which of three possible answers you get: a confidence interval for a ratio, the exterior of an interval for a ratio, or that the whole real line is consistent with the data, up to whatever level of probability seems appropriate. The data tell you that. If you force yourself to make an interval statement, when interval statements are inappropriate, then you may get quite the wrong answer.

Jim Berger

To come back to what a discussion like this ends up saying, one has to start being precise about what frequentism means. I have gotten the feeling that when this body is talking about confidence sets in a frequentist way, it interprets it in a literal Neyman way, where, if I have a 95% confidence set, by gosh that thing has to cover the true parameter 95% of the time no matter what the true parameter is. And so the discussion about this ratio of means problem that we are having is that there is this theorem that says that the only way a frequentist can do that is by saying some of the time for some data, the whole real line. David points out that perhaps this is perfectly reasonable to say, but my only complaint about it is saying the parameter is in the whole real line, and I have 95% confidence in that statement. Because obviously I am sure that the parameter is in the whole real line. So it's the attachment of the 95% confidence statement, which is demanded by strict frequentist, Neyman-type thinking, that is the issue in my mind.

Luc Demortier

Well, if there are no more questions, there was a third question on my list this morning, it's a quick one, whether you think it is possible to combine ABC (Approximate Bayesian Computation) methods with the reference prior algorithm.

José Bernardo

I do not have a real answer because I have never tried to. My hunch is that you can always do something, either exactly or approximately, but I don't know enough about the subject to make a serious comment. But Jim?

Jim Berger

I don't think so, because the ABC methods are when you don't have an explicit form for the likelihood, and given that the computation of reference priors is a delicate asymptotic computation, that depends very much on the form of the likelihood, I don't think that any kind of ABC approximation could be combined with that. So I think the answer is no.

Kyle Cranmer

I am just trying to think of something a little bit more physics motivated but still about reference

priors. If we go to something like searches for the Higgs, and we wanted to, say, set limits on the Higgs cross section, or the Higgs mass or something like that, there we may or may not have relationships between various parameters. For instance the Higgs mass is the only free parameter in the standard model, and we could try to come up with a reference prior for it. But we don't have to work hard to go to a different class of theoretical physics models where the relationship between how the Higgs might decay in various different ways gets broken, and it's no longer this one-dimensional model but it's now two or three or four of five. A reasonable question for us is to imagine that someone handed us all of the machinery for using reference analysis, and we are doing something like the Higgs, we need for instance the normal standard model Higgs and then maybe a Higgs where the cross section is independent of the mass, and maybe one where we have multiple Higgs bosons, each with different masses and cross sections, and different production modes and different branching ratios, and so there we could quickly get like 16, 20 parameters of interest in that model. We will somehow need to figure out how to navigate through that. So that's really a question for the physicists but still along the lines that if we had this kind of machinery available, maybe that prompts some thought for discussion.

Harrison Prosper

This is a very interesting, important problem, actually, how to deal with these multi-parameter models in this framework. In fact, just to give you a concrete example that my colleague and I are working on. We use the standard Poisson model, but this time there are two parameters of interest, which enter linearly in the mean. So you have some constant times θ_1, plus another constant times θ_2, plus a nuisance parameter, the background. And so one of the questions I have is that right now we were talking about this one-dimensional parameter problem, but if you have two parameters that seem to enter the problem in a symmetric way, can we treat those two parameters simultaneously, that is, to apply the reference algorithm on two parameters at the same time, or does one still have to do the sequential algorithm?

José Bernardo

I really think you should do it sequentially, not because of the technicalities, but because of the results to be obtained. I'll give you an example: under regularity conditions, you can easily extend the argument by which the reference prior becomes a Jeffreys prior, to the argument that in a multivariate situation you get a multivariate Jeffreys prior; but we know that this doesn't work, in fact it works very badly. The reason is that you are maximizing the amount of information simultaneously on all the parameters. It's just a fact that if you want to be able to get good properties of the marginal reference posterior for the parameter of interest, you just cannot do it globally, you have to do it sequentially, and in order. If you reverse the order, as some people have tried, you don't get the right answer either. The Stein's paradox, the problem that David Cox mentioned before, is a very dramatic example of that. When you have many parameters, if you try to do anything globally, with all of them, you are going to get very bad results. The answer I think is, it would work technically, but you would not get the right answer. So I believe you have to do it sequentially.

Jim Berger

This is a sort of comment on this question of reference priors when you have many parameters. Just when I sat down with Kyle, he reminded me that with his Asimov datasets [4] he can compute Fisher information matrices. And the numerical reference prior algorithms are very, very general, but everything simplifies a great deal if one is in a regular situation, where the Fisher information exists and makes sense. Then, in one of our earlier papers [5], José and I have a much simpler algorithm, based

only on the Fisher information matrix, for computing the sequential, iterative reference prior, and I think it's worth taking a look at that algorithm, combined with Kyle and collaborators' numerical computation of the Fisher information matrix, that might end up being a much easier way to implement the numerical algorithm.

References

[1] J. O. Berger, J. M. Bernardo, and D. Sun, "The formal definition of reference priors," Ann. Statist. **37**, 905 (2009); `http://www.uv.es/~bernardo/2009Annals.pdf`.

[2] L. Demortier, H. B. Prosper, and S. Jain, "Reference priors for high energy physics," Phys. Rev. D **82**, 034002 (2010); arXiv:1002.1111v2 [stat.AP] (2010).

[3] See section 5.2 in J. M. Bernardo, "Reference posterior distributions for Bayesian inference," J. R. Statist. Soc. B **41**, 113 (1979); `http://www.uv.es/~bernardo/1979JRSSB.pdf`.

[4] G. Cowan *et al.*, "Asymptotic formulae for likelihood-based tests of new physics," arXiv:1007.1727v2 [physics.data-an] (2010).

[5] J. O. Berger and J. M. Bernardo, "Ordered group reference priors with application to the multinomial problem," Biometrika **79**, 25 (1992); `http://www.uv.es/~bernardo/1992Biometrika.pdf`.

Model Inference with Reference Priors

M. Pierini[1], *H. Prosper*[2], *S. Sekmen*[2], *M. Spiropulu*[1,3]
(1) CERN, Geneva, Switzerland
(2) Florida State University, Tallahassee (FL), USA
(3) Caltech, Pasadena (CA), USA

Abstract
We describe the application of model inference based on reference priors to two concrete examples in high energy physics: the determination of the CKM matrix parameters $\bar{\rho}$ and $\bar{\eta}$ and the determination of the parameters m_0 and $m_{1/2}$ in a simplified version of the CMSSM SUSY model. We show how a 1-dimensional reference posterior can be mapped to the n-dimensional (n-D) parameter space of the given class of models, under a minimal set of conditions on the n-D function. This reference-based function can be used as a prior for the next iteration of inference, using Bayes' theorem recursively.

1 Introduction

It is typical in high energy physics (HEP) to deal with classes of models, e.g. new physics extensions of the Standard Model (SM), differing by the values of a set of (typically continuous) unknown parameters.

Given a set of experimental measurements, one would like to define the region of the model parameter space that is in agreement with the data. This is what we refer to as *Model Inference*. The following ingredients are needed:

- a theoretical tool that predicts the expected values of the measured observables, given a point in the model parameter space;
- a multi-dimensional likelihood, built from the available measurements;
- and a statistical procedure that evaluates the level of agreement between the data and the predictions.

While the first and second steps are not controversial, the third step is often polemical and is subject to some degree of arbitrariness. Two main approaches are typically followed: Bayesian, which computes the posterior probability of the expected values of the model parameters given the likelihood and a prior probability, and frequentist, which provides probability statements about possible values of the *measurements* given the assumed values of the model parameters.

Historically, most high energy physicists have preferred frequentist statistics because (they say) it allows one to extract statistical information from data without the need for subjective input. In this sense, these physicists are victim of the utopian idea of an analyst-free analysis, in which the "data speak for themselves", independently of the personal opinion and judgement of the physicists who perform the analysis. However, we are rudely awakened from this utopian dream on a daily basis as anybody who has had to evaluate a systematic uncertainty can confirm [1]. Beyond this simple fact, we also tend to underestimate how strongly the subjective beliefs of the analyst enters the earlier stages of an analysis, as for instance when we *define* the form of the likelihood. Physicists quote results as $m \pm \sigma$, where m and σ summarize the result of, perhaps, a likelihood-ratio-based analysis, which already implies assumptions about the form of the likelihood. When estimating the systematic uncertainty, we typically sum the different contributions in quadrature, implying that the systematic errors are uncorrelated and,

[1] About 10% of the `hep-ex` papers on `INSPIRE` match the search for the word *assume*, which is quite far from the analyst-free paradigm of our dreams.

more importantly, that they may be treated as if they are statistical. This may be true of systematic uncertainties that arise ultimately from other statistics; but many systematic uncertainties are “assigned" based on judgement or official policy.

We push this even further when we perform phenomenological analyses. While connecting the parameters of a model to the experimental observables, we often need to know a set of additional quantities (theoretical nuisance parameters) which are not measurable, but which may be known with some uncertainty through a theoretical calculation. This is the case, for instance, for the non-perturbative QCD parameters determined using lattice QCD calculations. In order to take into account the uncertainty on the predictions correctly, a Bayesian analyst would introduce a prior probability density function (pdf) for the theoretical nuisance parameters based on the best judgement of the theorist. While this is considered *dangerously subjective* by many high energy physicists, the same physicists consider it safe to modify the likelihood to take account of the theoretical uncertainty on predictions. This breaks the objective-frequentist-physicists paradigm twice: i) the functional form used to account for theoretical uncertainty is no less subjective than the prior of a Bayesian analysis and ii) the likelihood loses its deep and precise meaning of that function obtained by inserting the observations into the probability density function describing possible observations. Nobody ever did (and it is likely that nobody ever will) measure the theoretical nuisance parameters — indeed, many such parameters such as the factorization and renormalization scales are pure artifacts of our current reliance on perturbation theory in theoretical calculations. As a matter of fact, a physicist performing data analysis is forced to make assumptions. And there is nothing wrong with that as long as the assumptions are clearly stated. The problems come when the assumptions are hidden in the procedure and not transparent to the people not directly involved in the analysis.

The contrasting attitudes described above can be summarized in terms of the following two perceived problems:

- For some high energy physicists, introducing a prior is unacceptable because it brings subjectivity into science. *“The origin of the problem lies in the very first Bayesian assumption, namely that unknown model parameters are to be understood as mathematical objects distributed according to PDFs, which are assumed to be known: the priors. Obviously, the choice of the priors cannot be irrelevant; hence, the Bayesian treatment is doomed to lead to results which depend on the decisions made, necessarily on an unscientific basis, by the authors of a given analysis, for the choice of these extraordinary PDFs.”* [1].
- For some statisticians, a meaningful statistical analysis is not possible in the absence of an analysis procedure that allows one to incorporate a priori knowledge in a coherent way. *“The frequentist approach to hypothesis testing does not permit researchers to place probabilities of being correct on the competing hypotheses. This is because of the limitations on mathematical probabilities used by frequentists. For the frequentists, probabilities can only be defined for random variables, and hypotheses are not variables (they are not observables)... This limitation for frequentists is a real drawback because the applied researcher would really like to be able to place a degree of belief on the hypothesis. He or she would like to see how the weight of evidence modifies his/her degree of belief (probability) on the hypothesis being true.”* [2].

The use of reference priors [3] is emerging as a concrete way to solve the two problems. While a detailed discussion of the reference priors is beyond the scope of this paper, we highlight here their most appealing properties.

The main concern against the use of a Bayesian analysis in HEP is related to a priori ignorance, more than a priori knowledge. Whenever a priori knowledge is available (e.g. the measurement of the luminosity, which is used to translate an observed signal yield into a cross section measurement), there is a general consensus that an *evidence-based* prior should be used. The real issue is how we should parameterize "ignorance". The use of a flat prior, a HEP standard, is not quite the right answer. Reference

priors can be seen as a model of ignorance in the sense that, on average, they maximize the influence of the likelihood relative to the prior; hence they are a solution to this problem. More precisely, for a given likelihood, the reference prior is the prior function that *on average* maximizes the asymptotic Kullback-Leibler divergence [4] between the prior and the posterior, hence enhancing the role of the likelihood (the data) over the prior. This is exactly the kind of behavior that we would like for a model of ignorance. And this is what we assume the flat prior does for us, when we use it. Unfortunately, the flatness of the prior is not invariant under reparameterization. Unlike the flat prior, reference priors give reparameterization-invariant results in the cases typically considered in HEP (e.g. one-to-one transformations for which the Jacobian is not singular [5]). The use of reference priors in HEP has been recently proposed in Ref. [6], where the application in the case of a counting experiment is discussed. This has been applied to real LHC data, in one of the CMS Supersymmetry (SUSY) searches [7].

In the following, we apply the procedure described in Ref. [6] to two specific cases: i) the determination of the parameters $\bar{\rho}$ and $\bar{\eta}$ (at fixed A and λ) of a *simplified* CKM matrix and ii) the determination of the parameters in the case of a SUSY model [2]. In both cases, as an illustration, we limit the discussion to the determination of two parameters. The generalization to $n > 2$ dimensions is computationally more demanding, but conceptually equivalent. In both cases, we start from one experimental measurement, for which the likelihood can be analytically modeled without too much arbitrariness. We briefly describe the derivation of the reference posterior, following Ref. [6]. We then map the 1-D posterior into a n-D ($n = 2$ in our examples) function of the model parameters, introducing the *look-alike* (LL) prescription. This function, based on a reference prior, can then be used as the prior in a recursive application of Bayes' theorem to include other measurements.

2 The reference posterior for a 1-D analysis

When looking for a signal, produced by the process under study, we are confronted with a Poisson count of a signal on top of a background coming from other physics processes. The likelihood for the signal, in the absence of a background, is described by a Poisson function. In the presence of a background the likelihood asymptotically converges to a Gaussian density. Under these conditions, the reference prior is Jeffrey's prior for a Poisson likelihood, $\pi(\theta) \sim 1/\sqrt{\theta}$.

This is the case for the exclusive measurement of V_{ub} from $B \to \pi\ell\nu$ decays. What one measures is the branching ratio $BR(B \to \pi\ell\nu)$, which is the related to the the absolute value of the CKM matrix element V_{ub} as:

$$|V_{ub}|^2 = \frac{BR(B \to \pi\ell\nu)}{\Gamma_B F(B \to \pi)}, \tag{1}$$

where evidence-based priors are available both for the width of the B meson Γ_B (from other measurements) and the $B \to \pi$ form factor $F(B \to \pi)$ (from theory). One can determine the reference posterior for the BR using $\pi(BR) \sim 1/\sqrt{BR}$.

For SUSY searches, one looks for a signal yield s in a signal-sensitive *box*, defined by a selection using signal-vs-background separating variables. One observes a yield $n = s + \mu$, where μ is the background surviving the signal-enhancing selection. The expected background $\bar{\mu}$ is estimated from a sideband region where no signal is expected, where the observed yield in the sideband is y and the scaling factor b is such that $b\mu$ is the expected background yield in the sideband. In formulae:

- the likelihood is $p(n|s,\mu) = (s+\mu)^n e^{-(s+\mu)}/n!$,
- the prior for μ is $\pi(\mu) = b(b\mu)^{y-1/2}e^{-b\mu}/\Gamma(y+1/2)$ and
- the prior on s is $\pi(s) = \pi(s|\mu) \propto 1/\sqrt{s+\mu}$,

where $\Gamma(x)$ is the gamma function.

[2] For simplicity, we take a simplified CMSSM with m_0 and $m_{1/2}$, fixing $A_0 = 0$, $\tan\beta = 10$, and positive μ.

Once the 1-D reference-posterior is derived as described above, we translate this into an n-D function of the model parameters. While rigorous algorithms exist to build the n-D reference prior [3], we follow here a computationally simpler *heuristic* construction, described below, which we call the *look-alike prescription*.

3 Look-alike prescription

Let us consider a class of models identified by a set of n parameters θ (e.g. the parameters of a SUSY model). Given a measurement of a physics observables (e.g. the counting experiment of a SUSY search at the LHC), one culd map the 1-D reference posterior to the n-D space of the θ parameters demanding that the n-D pdf for θ satisfies two requirements:

- all models predicting the same values for the parameter x (V_{ub} and s in our two examples) associated with the posterior density $P(x)$ are equi-probable and
- the n-D function should be such that it maps back to a 1-D function $P'(x)$ identical to the $P(x)$ with which we started.

Given the mapping $\theta :\longrightarrow x$ predicted by the physics model, these requirements are sufficient to map $P(x)$ to $\pi(\theta)$. We first write the n-D function as $\pi(\theta) = K(x(\theta)) \times P(x(\theta))$. The computation of $K(x(\theta))$ goes as follows:

$$\begin{aligned} P'(\tilde{x}) &= \int d\theta\, P(x(\theta))\, K(x(\theta))\, \delta(\tilde{x} - x(\theta)), \\ &= P(\tilde{x}) K(\tilde{x}) \int d\theta\, \delta(\tilde{x} - x(\theta)) = P(\tilde{x}), \end{aligned} \tag{2}$$

where the last equality follows from the second condition. This implies that

$$K(\theta) = \frac{1}{\int d\theta\, \delta(\tilde{x} - x(\theta))}, \tag{3}$$

which is the surface of the region spanned by the look-alike (LL) models, that is, models giving the same value $\tilde{x}$ [3].

4 Two Examples

4.1 Example 1

The case of V_{ub} is useful because it allows us to explain how this works in practice. All the models such that $\bar{\rho}^2 + \bar{\eta}^2 = k$ predict the same value of $|V_{ub}|$. This makes them LL models, by our definition. The LL domain is a circle centered at 0 with radius $\sqrt{k}$. The n-D function is therefore,

$$\pi(\bar{\rho}, \bar{\eta}) = \frac{P(V_{ub}(\bar{\rho}^2 + \bar{\eta}^2))}{2\pi\sqrt{\bar{\rho}^2 + \bar{\eta}^2}}, \tag{4}$$

where $P(V_{ub})$ is the reference posterior for $|V_{ub}|$. The function $\pi(\bar{\rho}, \bar{\eta})$ is then used as the prior to fit the CKM matrix [8] including the measurement of the CKM phase γ. This step gives the allowed region for $\bar{\rho}$ and $\bar{\eta}$ shown in the left plot of Fig. 1), which is to be compared to a similar plot obtained using flat priors for $\bar{\rho}$ and $\bar{\eta}$ (right plot of Fig. 1). The results of these two calculations are consistent. However, the reference posterior for $|V_{ub}|$ provides a more solid foundation for determining the prior to associate with the CKM parameters.

[3] The challenge of generalizing this approach to a generic n-D problem is the calculation of this surface term.

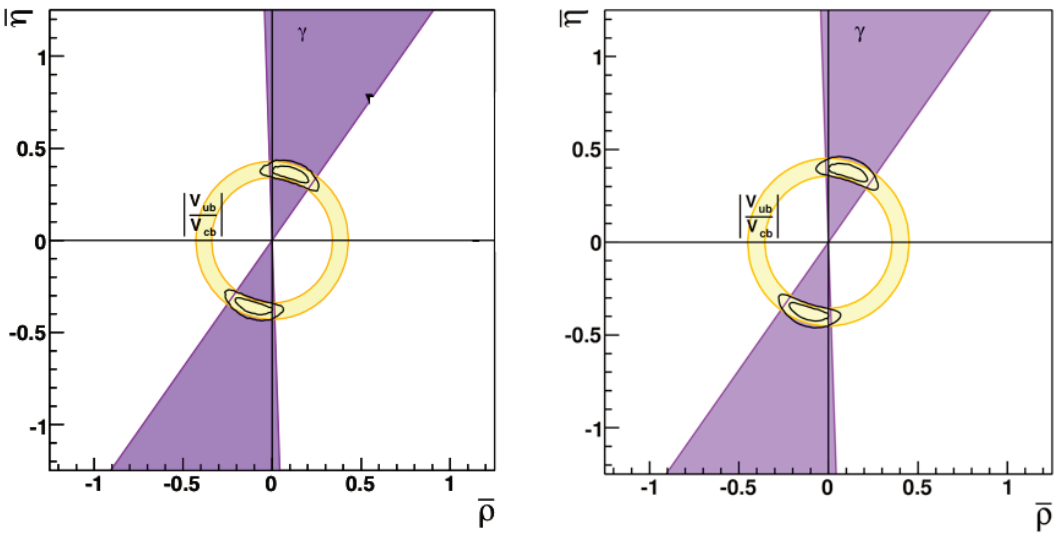


Fig. 1: Result for the 2-D allowed region for the CKM parameters $\bar{\rho}$ and $\bar{\eta}$, obtained using the reference posterior for V_{ub} and the LL prescription (left), or using flat priors for $\bar{\rho}$ and $\bar{\eta}$ (right). The information from V_{ub} is combined to the constraint from γ to derive a 68% and a95% credibility region for $\bar{\rho}$ and $\bar{\eta}$.

4.2 Example 2

Our second case study, which uses a simplified version of the mSUGRA model, is more complicated since Eq. 4 cannot be solved analytically in the case of a generic search for new physics. In this case, the LL domain is given by all the models predicting the same expected signal yield s. The expected signal yield as a function of the model parameters can be written as $s(m_0, m_{1/2}) = \epsilon(m_0, m_{1/2})\sigma(m_0, m_{1/2})\mathcal{L}$, where only the luminosity $\mathcal{L}$ is a constant, while both the cross section σ and the efficiency ϵ of the applied selection depends on the features of the model (e.g. the masses of the SUSY particles), and hence on the model parameters. The function $\sigma(m_0, m_{1/2})$ can be computed from the SUSY Lagrangian, while $\epsilon(m_0, m_{1/2})$ has a non-trivial dependence on the models, through several effects connected to the detector response. For instance, a model with large (small) mass differences would give a large (small) value of ϵ, since harder (softer) spectra for the visible particles produced in the SUSY decay chain will have larger (smaller) chance to survive the kinematic cuts. In general, the connection between the features of the model and the detector performance produces non-analytical iso-yield contours for the LL domains. This is illustrated in Fig. 2, where $\epsilon(m_0, m_{1/2})$ and $\sigma(m_0, m_{1/2})$ are shown in the case of a hypothetical SUSY search [9].

On the other hand, all the iso-yield contours have infinite length, resulting in constant $K(m_0, m_{1/2})$ if one considers the full domain for m_0 and $m_{1/2}$, and approximately constant if one uses a large-enough domain in practice [4]. We can then take $K(m_0, m_{1/2})$ as a constant and show how the method works. It has to be clearly stated that this is an approximation, and that the computation of the surface term of Eq. 4 is the main challenge in the applicability of the proposed method in its exact form (see Ref. [9] for details).

For illustration, we take the CMS low mass (LM) point [10] ($m_0 = 60$, $m_{1/2} = 250$) as the *true* state of nature and we simulate the case of an experiment giving a result exactly at the expectation, for low (1 pb^{-1}), moderate (100 pb^{-1}), and large (500 pb^{-1}) statistics. Figure 3 shows the 2-D function

[4] In case the measurement points to particular region of the plane, i.e. when there is hint of a signal, one could use the Savage prescription and cut the plot where the likelihood drops to negligible values. In absence of a signal hint, the situation is complicated by the fact that the likelihood peaks at infinite values of m_0 and $m_{1/2}$, where the SUSY particles are so heavy that they decouple from the SM ones, effectively recovering the SM limit.

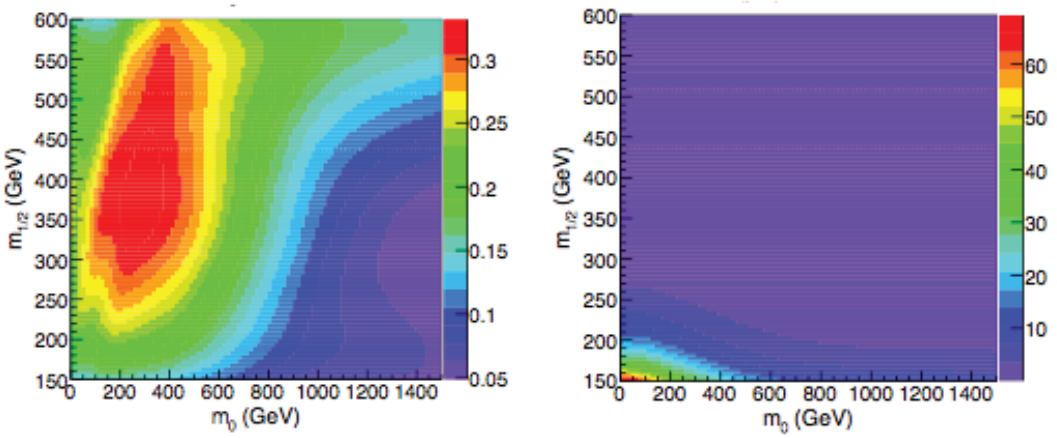


Fig. 2: $\epsilon(m_0, m_{1/2})$ and $\sigma(m_0, m_{1/2})$ functions in the case of a hypothetical SUSY search [9].

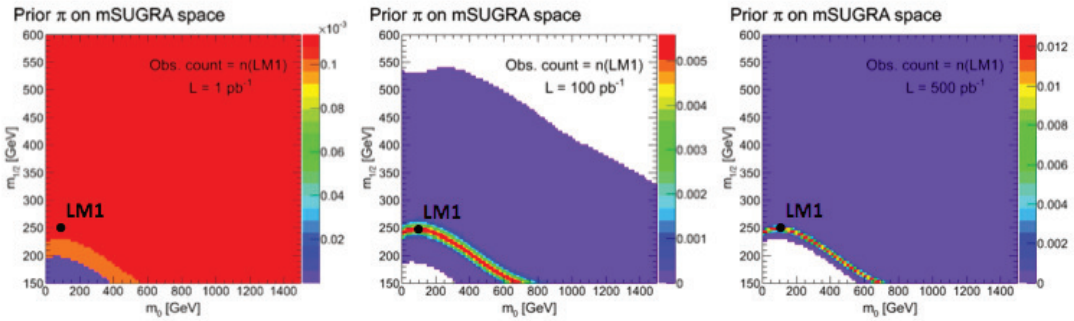


Fig. 3: Result for the 2-D function mapped from the 1-D reference posterior in the case of 1 pb^{-1} (left), 100 pb^{-1} (center), and 500 pb^{-1} (right) integrated luminosity. We assume a *measurement* perfectly in agreement with the expectation from the true model, corresponding to $m_0 = 60$, $m_{1/2} = 250$.

obtained by the LL prescription. With increasing sample size, the function shows a peak corresponding to the *true* value and to all its (degenerate) LL models, showing the consistency of the procedure.

5 Conclusions

We described the use of the reference prior in 1-D cases (typical of a HEP measurement) and how this can be used to define an n-D function of the model, induced by the 1-D reference posterior, which may then be used as a prior for further applications (e.g. to fit to model parameters). The connection between the 1-D posterior on a measurable quantity s (e.g. a signal yield on top of a background b) and an n-D function of a set of interesting parameters (e.g. the parameters of a SUSY model) is established through the look-alike prescription, which defines a heuristic procedure on the basis of two minimal conditions: i) the models predicting the same expected value for the interesting variable s are equi-probable and ii) the n-D function should map back to the 1-D reference posterior for s, from which we started. This requires the calculation of a surface term (see Eq. 4), which can be performed numerically [9]. While in specific cases this choice of a prior might be in conflict with a subjective assessment that could favor one region of the parameter space over another, it should be stressed that this *Bayesian* approach is likely to give the best frequentist performance because of the good frequentist properties of reference priors.

We provided two simplified 2-D examples to illustrate the method, for which computational complications are absent or marginal. Work is in progress to extend this procedure to more realistic cases [9].

6 Acknowledgment

The authors would like to thank J. Berger for helpful suggestions and an interesting discussion.

References

[1] J. Charles, A. Hocker, H. Lacker, F. R. Le Diberder and S. T'Jampens, hep-ph/0607246.

[2] S. James Press, Subjective and objective Bayesian statistics: principles, models, and applications, John Wiley & Sons, 2003

[3] See for instance, J. Berger and J. Bernardo (1992) 'Bayesian Statistics 4', J.M. Bernardo *et. al.* (Eds.), 35-60, Oxford University Press, Oxford (and references therein).

[4] S. Kullback and R. A. Leibler, Annals of Mathematical Statistics **22**, 79 (1951).

[5] G. Sankar Datta and M. Ghosh, The Annals of Statistics, **24**, 141 (1996).

[6] L. Demortier, S. Jain and H. B. Prosper, Phys. Rev. **D82** , 034002 (2010); [arXiv:1002.1111 [stat.AP]].

[7] The CMS Collaboration, Physics Analysis Summary SUS-10-009.

[8] M. Bona *et al.* [UTfit Collaboration], JHEP **0610** , 081 (2006); [hep-ph/0606167].

[9] M. Pierini, H. B. Prosper, S. Sekmen and M. Spiropulu, [arXiv:1108.0523 [physics.data-an]].

[10] The CMS Collaboration, J. Phys. G **34** (2007).

Banff Challenge 2

Thomas R. Junk
Fermi National Accelerator Laboratory

Abstract
Experimental particle physics collaborations constantly seek newer and better ideas for improving the sensitivity of their searches for new particles and phenomena. Statistical techniques are the last step in interpreting the results of an experiment; they are used to make discoveries (hypothesis testing), and to measure parameters (point estimation). They are also used in the first step – experiment and analysis design. Banff Challenge 2 asks participants to test their methods of discovering hidden signals in simulated datasets and of measuring the properties of these signals. The Challenge problems are described, and the performances of the submitted entries is summarized, for datasets with and without simulated signals present.

1 Introduction

Experimental particle physicists are becoming more aware as time goes on of statistical techniques that have been developed in the context of other fields over the years, and are interested in new research as well, in order to maximize the usefulness of their experiments. To that end, an ongoing dialogue between physicists and statisticians has been very fruitful, providing useful benefits to both parties. Particle physicists gain knowledge of established and new techniques, and statisticians can explore their techniques with particle physics data. Unfortunately, the direct use of experimental collider data requires permission from a usually very large collaboration, and a substantial investment in understanding (and possibly improving) the modeling of imperfect detectors and imperfectly known physics processes. In order to simplify the process and allow as many people to participate as possible, well-defined "Challenge" problems have been created so that simulated data may be freely exchanged and results compared. This model of collaboration between statisticians and particle physicists was very successful in the experience of the first Banff Challenge [1], which explored the case of setting one-sided bounds on new physics processes in a Poisson counting experiment in which the background rate is constrained by an auxiliary Poisson counting experiment. We seek with Banff Challenge 2 to test methods of discovery and measurement.

2 Banff Challenge 2 Problems

Two problems were posed. Each has different features that illustrate some of the challenges faced by experimentalists when analyzing the data. In High-Energy Physics (HEP) language, the first problem consists of seeking a mass bump on top of an exponentially falling background. In the language of statisticians, the data are generated by a marked Poisson process – x is a quantity measured on each selected collision event, and is called a "mark". The signal and background distributions are given parametrically in this problem to simplify the treatment and to make the problem more accessible. The position of a localized excess is not specified in advance, although only one bump at a time is allowed to be present in this problem. Because the bump may be anywhere within the range of x provided, the issue of multiple testing, also called the "Look-Elsewhere Effect" (LEE), arises [2]. Participants were asked to measure the peak position and rate. The second problem asks participants to address a common analysis situation. The signal and background predictions against which the data are tested are provided by a Monte Carlo simulation and not an analytic parameterization. No LEE is present in the second problem. In both problems, both the null and the test hypotheses are compound hypotheses – the background rates are subject to systematic uncertainty. In real HEP problems, the rates and shapes of the signals and

multiple sources of background are uncertain, and these parameters are constrained by auxiliary data and/or approximate theoretical predictions.

In both problems, the task is to identify those simulated datasets that have signals injected in them and to measure the parameters of the signals. The Type-I error rate, that is, the rate at which evidence is claimed in the case that a signal is absent, should be no larger than 1%. Typically this is what is meant in a HEP experiment by "significance", that evidence is claimed by a method with a specified Type-I error rate. Participants then should optimize the power of their tests, that is, to claim evidence for a signal on a many simulated datasets that actually do contain a signal, while keeping the Type-I error rate within its bound. The correct evidence rate is $1 - \beta$, where β is the Type-II error rate, and this rate depends on the details of the signal injected, such as its strength and its position in the distribution of the marks. Participants were in addition asked to calculate their correct evidence rates for a small number of signal hypothesis choices, which were also among the choices tested in the blind samples. The estimation of the power of the test by the participant models an important ingredient in a HEP experiment. Physicists who propose building an experiment or conduct a specific analysis with an existing apparatus must justify their efforts to their colleagues and to their funders. They must provide a convincing argument that their experiment can provide a result that is interesting – that it is possible to find evidence for the sought-after new particle or process, or to exclude it if it is not present. Such estimates are used in decisions that allocate resources among experiments, and it is important that these estimates are neither underestimates nor overestimates. Banff Challenge 2 provides a blind way to check the methods used to arrive at power and coverage estimates.

The criteria for "winning" the competition among participants were not spelled out explicitly when the challenge problems were posed. It was made clear that a Type-I error rate of 1% or less is important to satisfy, and the best power satisfying the Type-I error rate requirement is a natural ordering to rank methods. Nonetheless, since we would like to test also the ability to estimate power properly, and because HEP experimentalists usually only have the estimated powers to use to rank methods instead of blind tests, it becomes natural to rank the methods based on their estimated powers, provided that the measured powers are not measurably below the estimated powers.

The problems are discussed in more detail below. The Challenge problems and simulated datasets may be found on the author's web site [3]. A summary of the submissions is available as well [4].

2.1 Problem 1

For this problem, the simulated data samples are drawn from the following density functions. Statisticians prefer the term "intensity" in order to indicate that they are not normalized to unit area. The background intensity function is

$$B(x) = Ae^{-Cx} \tag{1}$$

where x is the mark of the event. The domain of x is restricted to be between 0 and 1. We choose the values $A = 10000 \pm 1000$ and $C = 10.0 \pm 0$. The background rate parameter A is drawn from a truncated Gaussian distribution of width 1000, truncated so that $A \geq 0$. The signal intensity function is

$$S(x) = De^{-(x-E)^2/2\sigma^2} \tag{2}$$

The problem statement specifies that $D \geq 0$ and that $\sigma = 0.03$ and $0 < E < 1$ in the generation of simulated datasets. Two example distributions are shown in Figure 1, one for a signal injected near the upper end of the range of E, and one generated near the lower end of the range.

The Challenge datasets for Problem 1 were generated randomly according to the distribution $B(x) + S(x)$. There are 24 different subsets of simulated pseudoexperiments, corresponding to different choices of D and E, and these are listed in Table 1. The numerical choices were governed by the desire to have a correct-discovery rate that can be measured accurately with a limited number of repetitions, and thus should not be too close to 0% or 100%, and that we would like to test regimes in

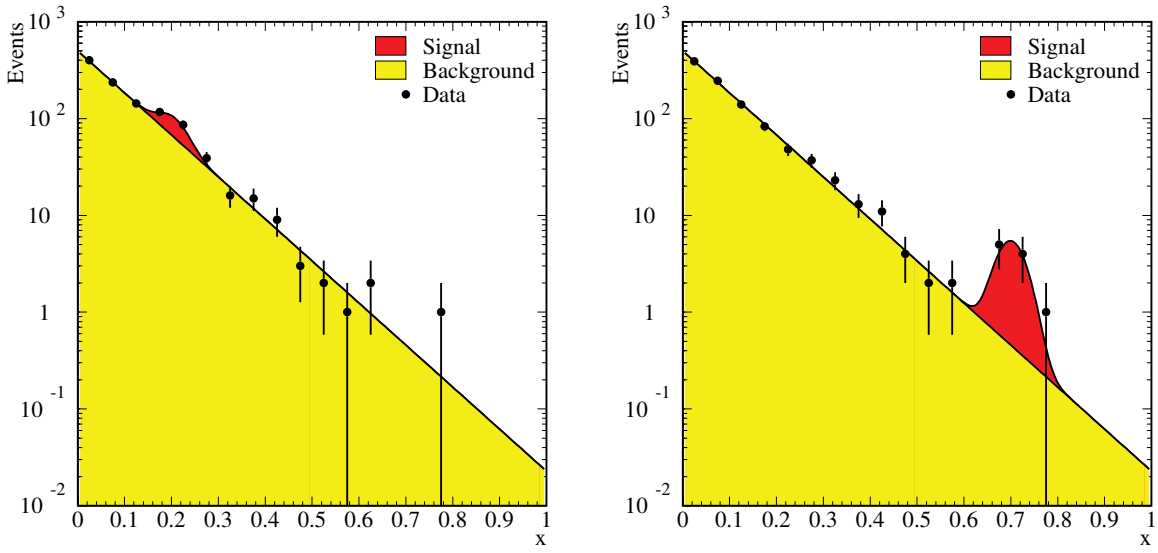


Fig. 1: Two example distributions for Problem 1. The simulated data are displayed in binned histograms, indicated by points with error bars drawn as $\sqrt{n_i}$ where n_i is count of simulated data events with marks that fall in that bin. Analytic functions for the signal and background intensities are also shown. Left: a signal injected with $E = 0.2$, and Right: a signal injected with $E = 0.7$.

which large numbers of data events are needed for discovery and regimes in which a single mark or two can make all the difference. Signals with large values of D and small values of E have an approximately Gaussian distribution of the signal yield, while signals with small values of D and large values of E are very sensitive to the Poisson nature of the data in sparsely populated areas of the distribution.

The parameters D and E are parameters of interest and are not affected by unknown values of nuisance parameters, of which there is only one in this problem, a simplification compared to a real experiment. Similarly, the location of a peak does not always correspond to the true value of the mass of a new particle, although the significance of a peak should not be affected by the uncertainty in the relationship between the measured peak position and the underlying process that makes events in the peak. Similarly, since the significance of a peak that is found depends on the comparison of the data with the prediction of the null hypothesis, uncertainties in the probability of the detector to detect a signal event and for the analysis technique to select it should have little impact on the significance of a peak that is found, although these effects do affect the expected sensitivity, signal rate measurements, and limits.

The background parameter A was chosen for each simulated dataset from its prior distribution, a Gaussian centered on 10000 with a width of 1000. An integer n_b was then drawn from a Poisson distribution whose mean is the total background integral from $x = 0$ to 1 using the randomly selected value of A. Then n_b marks x were generated from the exponential distribution $B(x)$. A similar procedure was followed for generating marks for the signal component, according to $S(x)$. The marks were then shuffled and written out to the Challenge dataset file. Simulated datasets from the 24 categories were also shuffled so that no clue to the injected values would be provided by either the ordering of the datasets or of the marks within a dataset.

Three standard signal models were chosen for the purpose of asking participants to estimate their correct-discovery fraction, and correspond to categories 4, 9, and 21 in Table 1. The correct-discovery fractions were measured on the corresponding challenge datasets and compared with the participants'

estimates.

The presence of a nuisance parameter in the null and test hypotheses complicates the definition of the Type-I error rate. One approach is to evaluate the Type-I error rate as a function of the true value of the unknown nuisance parameter(s). Another approach is to evaluate the Type-I error rate in the prior-predictive ensemble whose generation is described above. A third is to quote the largest Type-I error rate for a fixed range of values of the nuisance parameters. The ideal that a method should cover for all values of the nuisance parameter requires a specification of what is meant by "all". The approach here is to quote the error rate and the correct-discovery rates using the prior-predictive ensemble, although this is not the only valid definition. A method which has a Type-I error rate which is larger than the stated value, which is usually written in a high-energy physics publication as a confidence level or a significance level, is said to undercover and is unlikely to pass collaboration review.

A feature of Challenge Problems 1 and 2 is that signal rate intervals were requested only in the case that evidence is claimed, and the problem statement asks for zero to be reported if evidence is not claimed. These instructions reflect a flip-flopping procedure which is very commonly used in HEP. If a collaboration measures the mass of a new particle but does not claim evidence for the new particle, the result may be easily misconstrued. Coverage for signal rate and peak position measurements were only computed on the subset of simulated datasets on which evidence is claimed.

Not quoting the measured signal yield in simulated datasets for which evidence is not claimed biases upwards the measured signal yields and the intervals containing them. A simple example is the null hypothesis – the true signal rate is zero in null hypothesis simulated datasets, but in 1% of them, a method that is performing well should claim evidence for a signal. Even if the set of intervals for the signal rate cover properly for a method, selecting this sample of them will in general not have proper coverage. This is true to a lesser extent for test hypotheses with true signals present.

A final feature of Problem 1 is that at most one signal is present, at a single value of E. In a real experiment in which the signal is *a priori* unknown, there may be more than one signal present. Since most methods fit for the background rate in the process of testing for the signal, a second signal (or more) will change the background fit. One may legitimately ask whether all of the events are signal events from a broad spectrum of multiple signals, and this is where some theoretical input and auxiliary information from other experiments is needed to constrain the background prediction. For this problem, we treat the presence of at most one signal as auxiliary *a priori* information. The Challenge datasets were generated with no more than one signal in each.

2.2 Problem 2

Unlike Problem 1, Problem 2 parameterizes the predictions of the signal and background yields using finite samples of Monte Carlo. In a real HEP experiment, samples of collider data from control regions are sometimes used instead. From a statistical standpoint, these are very similar and are treated identically. Often there is an extrapolation uncertainty associated with using a different sample of data which pass different selection requirements, and which are used to predict the background in the sample passing the signal requirements. Monte Carlos are similarly fraught with uncertainty in their predictions, and these uncertainties are parameterized with nuisance parameters which are simplified in this Challenge problem to two – one for each background governing its rate.

The simulated datasets and the Monte Carlo samples were generated from smooth distributions for the marks. The distribution of the marks for Background 1 is given by

$$x = \min(1.0, 1.4y^{2.74}e^{-y/3}), \tag{3}$$

where y is uniformly distributed on the interval $(0, 1]$. Background 2 was generated with a uniform distribution. The signal distribution was generated using

$$x = z^{0.21}, \tag{4}$$

Table 1: Problem 1 Challenge dataset categories, listing the input values of E and D, the signal peak position and the signal rate parameters, respectively. The first category is the null hypothesis. For the categories marked with a "*", the participants were asked to compute their expected correct-discovery rates. The column headed $n_{\rm rep}$ lists how many simulated datasets were supplied for each of the categories.

Category	$E_{\rm input}$ (location)	$D_{\rm input}$ (intensity)	$n_{\rm rep}$
1	—	0.00	15400
2	0.50	83.78	200
3	0.38	265.96	200
4*	0.10	1010.65	200
5	0.10	478.73	200
6	0.66	66.49	200
7	0.78	39.89	200
8	0.10	744.69	200
9*	0.50	136.97	200
10	0.90	15.29	200
11	0.50	190.16	200
12	0.14	664.90	200
13	0.50	163.57	200
14	0.38	531.92	200
15	0.14	1196.83	200
16	0.50	110.37	200
17	0.10	1276.62	200
18	0.90	20.61	200
19	0.66	132.98	200
20	0.90	12.63	200
21*	0.90	17.95	200
22	0.90	23.27	200
23	0.78	79.79	200
24	0.10	1542.58	200

where z is uniformly distributed on the interval $(0, 1]$. The Challenge problem hid these underlying parameterizations, and gave samples of 5000 simulated marks for each of the three processes – signal, Background 1, and Background 2. The *a priori* yields given to the participants, which would be obtained from auxiliary experiments or theoretical predictions, are 900 ± 90 events for Background 1 and 100 ± 100 events for Background 2. The large fractional uncertainty on Background 2 leaves open many possibilities of how to interpret this prediction. It is a frequent occurrence in HEP experiments to have at least one background component that has a large fractional uncertainty evaluated for its prediction. Such ill-constrained predictions have less of an effect on experimental results if they constitute a small amount of background, where small is in relation to the other background components or to the signal that is tested. The signal rate is left unspecified, and varies from one simulated dataset to another.

In each of the Challenge datasets, a rate was chosen for Background 1, Background 2, and the signal, based on the hypothesis under test. Truncated Gaussian distributions were sampled for the Background 1 and Background 2 rates. The seven signal hypothesis categories are listed in Table 2. A Poisson random number was chosen using the randomly chosen rates, and then marks were generated using the prescriptions described above. The resulting lists of marks were then shuffled. The list of which simulated dataset was drawn from which signal test case was also shuffled.

Category 2 in Table 2 was chosen as the standard reference signal model for which the participants

Table 2: Problem 2 dataset categories – signal rates and how many repetitions of each were represented in the Challenge datasets. For the category marked with a "*", the participants were asked to compute their expected correct-discovery rates. The column headed n_{rep} lists how many simulated datasets were supplied for each of the categories.

Category #	Input Signal	n_{rep}
1	0.00	17600
2*	75.00	400
3	50.00	400
4	25.00	400
5	100.00	400
6	150.00	400
7	125.00	400

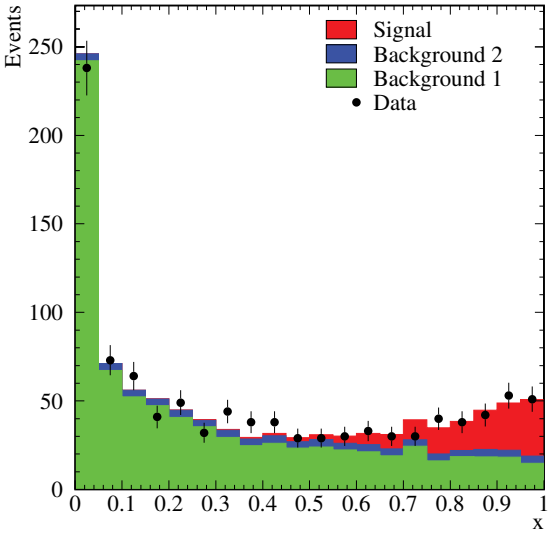


Fig. 2: An example distributions for Problem 2. The simulated data are displayed in a binned histogram, indicated by points with error bars drawn as $\sqrt{n_i}$ where n_i is count of simulated data events with marks that fall in that bin. The background and signal intensities are shown also with binned histograms formed from the Monte Carlo mark distributions for their respective samples.

were to estimate their correct-discovery fractions.

Figure 2 shows an example distribution of marks for a simulated dataset with a prominent injected signal.

3 Submissions

Submitters were asked to describe briefly their methods, and their descriptions are reproduced below, alphabetized by the author's last name. There is some variety in the notation used.

3.1 Mark Allen

For Problem 1, Mark Allen provided a solution based on an unbinned maximum-likelihood fit, $\Delta \log \mathcal{L}$ as the test statistic for computing p values. In order to find the global maximum of the likelihood most often, several fits are performed with different starting conditions. The p values are computed by comparing a dataset's test statistic with a distribution of a large number of simulated background-only datasets. Since a signal can be found anywhere in the distribution on any of the simulated background-only datasets, the LEE is taken into account.

3.2 Stefano Andreon

For Problem 1, Stefano Andreon provided a solution based on a Bayesian computation with uniform priors on A and D, with a zero value for the prior for negative (unphysical) values, and an uniform prior, between 0 and 1, on E. Stefano computes $p(D = 0|\mathrm{data})$, up to a multiplicative factor, and selects simulated datasets for discovery claims if $p(D = 0|\mathrm{data}) < 3 \times 10^{-3}$ for the first solution, and $p(D = 0|\mathrm{data}) < 4 \times 10^{-3}$ for the second. The Type-I error rate is higher for the second set, but the power is also larger. An estimate of the power of the tests was not supplied.

3.3 Frederik Beaujean

For Problem 1, Frederik Beaujean and the Bayesian Analysis Toolkit (BAT) team provided a solution based on BAT's fast Poisson p value estimation, corrected for the number of degrees of freedom. The value of A that maximizes the posterior probability in the background-only case is used. If the p-value is less than 0.01, a Bayesian analysis is conducted, and a discovery is claimed if $P(B|\mathrm{Data}) < 0.001$. The LEE is taken into account by assuming a prior that favors the background model. A rather small fraction of the simulated datasets with injected signals had a discovery claim using this technique.

3.4 Matt Bellis and Doug Applegate

Matt and Doug's solution to Problem 2 involves a fit to each dataset using a nearest-neighbors algorithm to estimate the PDFs of the two background and one signal MC components, and a bootstrapping procedure to marginalize over the correlated uncertainties inherent in this approach. Matt and Doug use toy Monte Carlo studies to estimate the Type-I error rate and the power of this procedure.

The nearest-neighbor estimation of the PDF for each process is designed to evaluate the PDF at values of the mark x where there are data events, but not for arbitrary values of x. This reduces the problem from estimating a function to estimating a set of numbers $P^k(\vec{x}) = \{P^k(x_i)\}$ for the kth contributing process (one of the two backgrounds or the signal). For each data point x_i and channel k, Matt and Doug calculate a probability density by counting the number of nearby Monte Carlo samples N_s^k within a range r_s, then divide by r_s and the total number of Monte Carlo samples N_{tot}^k. The estimate of the PDF will have noise from the finite size of the Monte Carlo sample. In addition, the values $P(x_i)$ will be correlated since the same points are used for neighboring density estimations. To account for these effects, Matt and Doug produce bootstrap realizations of the Monte Carlo and calculate PDFs for each random draw. The ensemble of $\{P_j^k(\vec{x})\}$, where j indexes bootstraps, are random draws from the PDF of $P^k(\vec{x})$ that includes the uncertainty from both finite number and correlation.

Matt and Doug use the MINUIT minimizer [5] to find the best-fit fractions for the three processes by minimizing the negative log likelihood, which includes the PDF information above for each process, as well as Gaussian constraints on the rates from the problem specification. Matt and Doug elect to compute the ratio of the probabilities of each model at the best fit parameters, *i.e.* the delta log-likelihood. They calibrate the delta log-likelihood statistic by computing its distribution in zero-signal toy Monte Carlo simulations and use that to compute p-values and sensitivities.

3.5 Georgios Choudalakis

The BUMPHUNTER [6] is a hypothesis test sensitive to local excesses of data with respect to the null hypothesis. It is configurable, which means it can also be sensitive to deficits, and can optionally require agreement in the sidebands around the local discrepancies it evaluates. It is not assuming any specific shape for the potential discrepancy it is looking for, and by default it does not assume any specific width either, therefore it is highly model-independent. The trials factor associated with checking for discrepancies of various widths at various positions is taken into account using pseudo-experiments. The algorithm is fast enough to make the use of pseudo-experiments practical in most cases, without excluding the possibility of analytic approximations if needed. Understanding the BUMPHUNTER as a hypertest [6] allows for straight-forward generalizations, such as looking for discrepant tails in distributions (TAILHUNTER [6] [7]) or other features, and combining multiple datasets in a single hypothesis hypertest while accounting for the LEE.

3.6 Eilam Gross and Ofer Vitells

Eilam and Ofer provided a solution to Problem 1 based on a two-fit log likelihood ratio similar to those used by other participants. The LEE is addressed using a procedure described in [8]. It was found that the submitted p-value distribution extended well above 1.0, although this is not a problem for discovery. This method of addressing the LEE works well for small p-values which are required for discovery, but rather conservatively magnifies large p-values. Confidence intervals for D and E are computed using the likelihood ratio test $\Delta 2 \log \lambda = 1$, additionally setting the lower bound on the signal rate to be zero when $P(q_0 \leq q_0^{observed} | H_0) = 68\%$, where $q_0 = -2 \log \lambda(0)$ if the best-fit signal rate is positive, and zero otherwise, and

$$\lambda(N_s) = \frac{\mathcal{L}(N_s, \hat{\hat{N}}_b, \hat{\hat{M}})}{\mathcal{L}(\hat{N}_s, \hat{N}_b, \hat{M})},$$

where N_s and N_b are signal and background yields, and the hats indicate best-fit parameters. Eilam and Ofer provided a solution to Problem 2 using a likelihood ratio test statistic similar to that of Problem 1, except in this case the likelihood ratio is binned, and there is no LEE.

3.7 Tom Junk

For Problem 1, Tom provided a solution based on an unbinned profile likelihood test statistic. Two fits are done, both using MINUIT [5], one in the test hypothesis, and one for the null hypothesis. Simulated datasets were generated using the prior-predictive ensemble. The LEE is incorporated by testing all datasets in the same way, allowing a peak to be found anywhere in the ranges $0 < E < 1$ and $0 < D$. Tom reports the values of D and E returned by the MINUIT fit.

Tom provided a solution to Problem 2 using a binned likelihood technique. Aside from the binning, and the lack of a peak position parameter, the method used is very similar to the solution used for Problem 1. An additional feature is the limited sample size of the Monte Carlo used to predict backgrounds. This adds an extra nuisance parameter for each bin for each sample – signal, background 1, and background 2. Tom fluctuates all of the nuisance parameters in each of his simulated data samples used to characterize the test statistic. This differs from the prior used to generate the datasets in that the characterizing datasets are binned, and the priors in each bin are taken as Gaussian approximations to the distributions of the bin-by-bin parameters. A possibly better choice is to use a Gamma prior in each bin for the bin-by-bin uncertainties, which is the Bayesian result using the finite Monte Carlo and a uniform prior in the unknown true background and signal rates. This however biases the prediction upward in each bin. Tom fit the two background rates, but did not fit the separate bin-by-bin uncertainties, to get values of the $-2 \ln Q$ test statistic for the simulated datasets and the challenge datasets, where Q is the ratio of profile likelihoods under the test and null hypotheses.

For the signal rate intervals, Tom performed a Bayesian calculation, integrating the likelihood function times a uniform prior in the signal rate over the uncertain parameters (this time, the two background rates and the bin-by-bin uncertainties). The 68% credibility interval is computed as the shortest interval containing 68% of the integral of the posterior.

Since Tom had access to the correct answers for each simulated dataset, Tom's solutions are not eligible to "win" the competition.

3.8 Valentin Niess

Valentin's analyses of the two problems rely on frequentist hypothesis testing tools, but they differ with respect to the test statistic that is considered. The first algorithm counts the number of events within a subinterval of the possible range of marks $\Gamma = [0; 1]$ chosen in order to maximize the separation of signal from noise. The optimal half-width of the bracketing interval was found to be $\Delta = 1.4\sigma$ in this case, where $\sigma = 0.03$ is the signal width in E. The background contamination in the subsample is simultaneously estimated from the sidebands. The search for the best value of E is repeated over N_{bin} brackets overlapping over the range $[0; 1]$ by steps of $\delta = \sigma/2$. The LEE is taken into account by correcting the p value by an effective trial factor, given as: $N_{eff} = |\Gamma|/\sqrt{2\delta\Delta}$, where $|\Gamma|$ is the length of the interval Γ.

The second algorithm proceeds with the Kolmogorov-Smirnov (KS) statistic, parameterizing the signal and background cumulative distributions with power-law functions of the marks. The KS test statistic is minimized numerically over the uncertain values of the signal and background rates.

3.9 Wolfgang Rolke

Wolfgang's solution to both problems is based on the likelihood ratio test statistic

$$\lambda(\mathbf{x}) = 2\left(\max\{\log L(\theta|\mathbf{x}) : \theta\} - \max\{\log L(\theta|\mathbf{x}) : \theta \in \mathbf{\Theta}^0\}\right)$$

where $L(\theta|\mathbf{x})$ is the likelihood function.

According to standard theorems in statistics $\lambda(\mathbf{X})$ often has a χ^2 distribution in which the number of degrees of freedom is the difference between the number of free parameters under the test hypothesis and the number of free parameters under the null hypothesis. This turns out to be true for Problem 2 but not for Problem 1, in which case the null distribution can be found via simulation.

For Problem 1 the main difficulty is finding the maximum likelihood estimator (MLE) because the likelihood surface has a large number of local minima. To find the MLE, Wolfgang used a two-step procedure: first a fine grid search over values of the signal location E from -0.015 to 1 in steps of 0.005. At each value of E the corresponding value of signal size α that maximizes the log-likelihood is found. In a second step, Wolfgang starts at the best point found above and uses the Newton-Raphson method to find the overall MLE.

In Problem 2 the difficulty is in estimating the densities for the backgrounds and the signal from the available Monte Carlo data. Wolfgang explored the following solutions:

a) parametric fitting: for all three data sets Beta densities (with different parameters) yielded fits that passed a number of goodness-of-fit tests.

b) non-parametric: the densities are estimated using non-parametric kernel estimators.

c) semi-parametric: a combination of a) and b).

3.10 Stefan Schmitt

Stefan Schmitt analyzed both Challenge problems using the method of fractional event counting [9]. This method defines a test statistic $X = \sum N_i w_i$, where N_i is the observed number of events in bin i and

w_i is the fractional event weight. The weights w_i depend on the unknown nuisance parameters, namely the signal rate (problem 1 and 2) and the signal position (problem 1 only). The w_i are constructed from the expected signal and background contributions together with the size of the variations expected from systematic uncertainties. This calculation is done in a way which maximizes the sensitivity of X to the presence of a signal and at the same time minimizes the sensitivity of X to systematic variations [9].

Once the w_i are defined, the calculation of X is inexpensive in terms of computing power. Probabilities are thus calculated using Monte Carlo methods. In particular, the p-value is defined as the fraction of background Monte Carlo experiments which have a test statistic X larger than the one found for the experimental data. All nuisance parameters but the unknown signal properties are integrated over when generating the Monte Carlo experiments. For calculating the p-value and deciding on the presence or absence of a signal, the signal rate in the weight calculation is fixed to a value r_0. The rate r_0 is chosen such that the expected Type-II error for a signal with rate r_0 is approximately 50%.

For the case of Problem 1, the signal position E is not known. A scan is performed as a series of tests with variable E, where E is increased in the range 0 to 1 in steps finer than the signal resolution. The minimum p-value found in this scan is corrected for the LEE by repeating the scan on a sufficiently large number of independent Monte Carlo experiments.

3.11 Stanford Challenge Team

The SCT provided a solution to Problem 1 based on a log-likelihood ratio test statistic performing two fits to each dataset. The distribution of the test statistic is predicted using simulation. The LEE is handled by allowing any value of E to be fit in the simulated null hypothesis datasets used to calibrate the critical value. The parameters D and E were obtained using a maximum-likelihood fit. The SCT used the nonparametric bootstrap to estimate the variability of the results.

The SCT provided a solution to Problem 2 using a likelihood ratio test similar to that used in Problem 1, comparing a three-component fit to a two-component fit (three including the signal, and two backgrounds are fit in either hypothesis). The distributions of the marks for the two background components and the signal component were approximated with Beta distributions.

4 Performance Summary – Problem 1

Challenge participants generally did quite well discovering the signals that were hidden in the simulated datasets. Table 3 lists the measured Type-I error rates and the correct-discovery sensitivities for the submissions for Problem 1. Two calculations of the correct-discovery rates are listed. The "claimed" rate is estimated by the participant, and the "measured" rate is that obtained from the challenge datasets.

No one ignored the LEE – the Type-I error rates were nearly all under the desired 1%. Stefano Andreon's submission provided a Bayesian test statistic with two suggested cuts on it, both of which gave Type-I error rates in excess of the desired 1%. Adjustment of the cut can certainly produce the desired error rate, although at the price of fewer correct evidence outcomes. Stefano Andreon's submission has rather high discovery rates measured with the challenge datasets, and allowing tuning of the cut on the test statistic it is estimated that the performance of the Bayesian method is similar to that of the other methods.

There appears to be an upper limit on the correct-discovery fractions for each of the three standard signal hypotheses for which the participants were asked to evaluate their sensitivities, approximately 40%, 50%, and 20% (rounding up slightly), indicating that the choice of signal rates and positions was optimized well in order to make them measurable with fewer repetitions. No one participant had the highest claimed sensitivity for all three points and also a Type-I error rate under 1%. Figure 3 lists the fractions of signal-containing simulated datasets for which each participant claims evidence, for all 24 signal hypotheses (including the hypothesis of no signal). The author would like to thank Ofer Vitells for producing this figure and the following two.

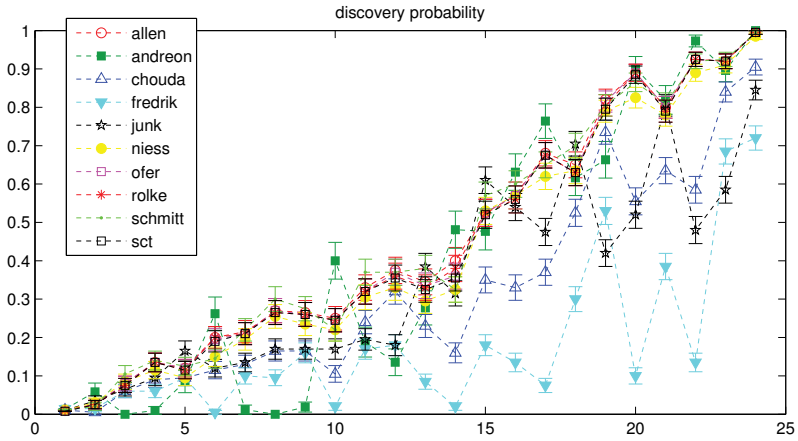


Fig. 3: Fractions of simulated datasets for which each participant claims evidence, for the 24 signal hypotheses of Problem 1 The categories are sorted on the horizontal scale according to the average correct-assignment fraction. The author would like to thank Ofer Vitells for preparing this figure.

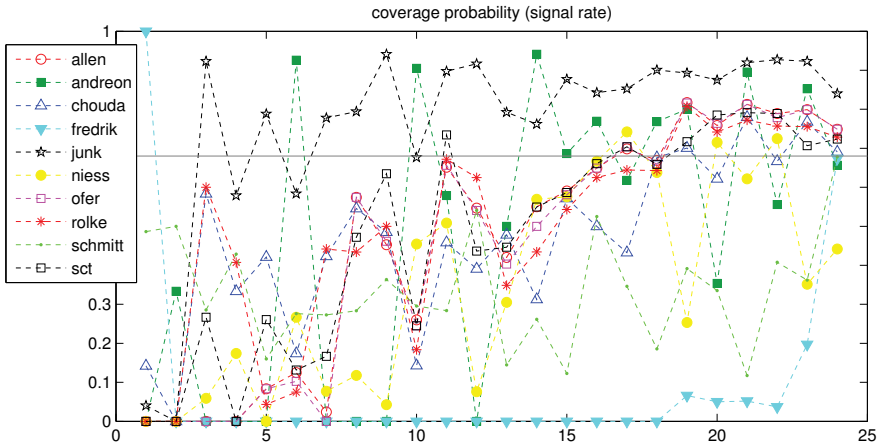


Fig. 4: Fractions of simulated datasets in which the 68% CL intervals quoted by the participants contain the true value of D, the signal strength parameter. The 24 signal categories are sorted in order of increasing average correct assignment probability. The author would like to thank Ofer Vitells for preparing this figure.

The performance for extracting the parameters D (which controls the signal strength) and E (which controls the signal position) were not as good. The summary note [4] provides listings for each participant for each signal model, the fraction of the datasets that claim evidence that also have intervals for D that contain the true value, and separately also for E. The average lengths of these intervals are also listed. Distributions of the fitted values of E, and the upper and lower edges of intervals of D are also shown in that note. Figure 4 in this article shows the fraction of the simulated datasets in which the signal rate parameter D is within the quoted intervals, which ideally should be 68% or greater. Figure 5 shows the fraction of the simulated datasets in which the signal position parameter E is within the quoted intervals, which also ideally should be 68% or greater. The true values of the signal rates and locations for each simulated dataset are provided on the web page [4] as an aid in investigating coverage issues with the intervals supplied.

Table 3: Listing of the claimed and measured correct-discovery rates for the three scenarios of Problem 1. Stefan Schmitt states that his unbinned sensitivities are rather similar to his binned sensitivities. The Bayesian technique proposed by Stefano Andreon did not have estimated correct-discovery rates provided.

Contributor	Type-I Error Rate	$D = 1010, E = 0.1$		$D = 137, E = 0.5$		$D = 18, E = 0.9$	
	Measured	Claimed	Measured	Claimed	Measured	Claimed	Measured
Tom Junk	0.0097 ± 0.0008	0.256	0.3150 ± 0.0328	0.543	0.6100 ± 0.0345	0.108	0.1350 ± 0.0242
Wolfgang Rolke	0.0103 ± 0.0008	0.356	0.3800 ± 0.0343	0.457	0.5250 ± 0.0353	0.184	0.2150 ± 0.0290
Stanford Challenge Team (SCT)	0.0077 ± 0.0007	0.3483	0.3550 ± 0.0338	0.4335	0.5200 ± 0.0353	0.0175	0.2100 ± 0.0288
Eilam Gross & Ofer Vitells	0.0082 ± 0.0007	0.35	0.3600 ± 0.0339	0.46	0.5250 ± 0.0353	0.19	0.2100 ± 0.0288
Valentin Niess	0.0111 ± 0.0008	0.34	0.3250 ± 0.0331	0.46	0.5300 ± 0.0353	0.17	0.1950 ± 0.0280
Georgios Choudalakis	0.0110 ± 0.0008	0.213	0.1600 ± 0.0259	0.290	0.3500 ± 0.0337	0.107	0.1300 ± 0.0238
Mark Allen	0.0106 ± 0.0008	0.385	0.4000 ± 0.0346	0.486	0.5250 ± 0.0353	0.187	0.2100 ± 0.0288
Frederik Beaujean (BAT)	0.0000 ± 0.0000		0.0000 ± 0.0000		0.0300 ± 0.0121		0.0050 ± 0.0050
Stefan Schmitt							
Unbinned	0.0112 ± 0.0009		0.4500 ± 0.0352		0.5450 ± 0.0352		0.1850 ± 0.0275
Binned	0.0110 ± 0.0008	0.37	0.3850 ± 0.0344	0.53	0.5450 ± 0.0352	0.17	0.2200 ± 0.0293
Stefano Andreon							
$p < 3 \times 10^{-3}$	0.0126 ± 0.0013		0.4811 ± 0.0485		0.4766 ± 0.0483		0.0120 ± 0.0120
$p < 4 \times 10^{-3}$	0.0191 ± 0.0016		0.5189 ± 0.0485		0.4766 ± 0.0483		0.0120 ± 0.0120

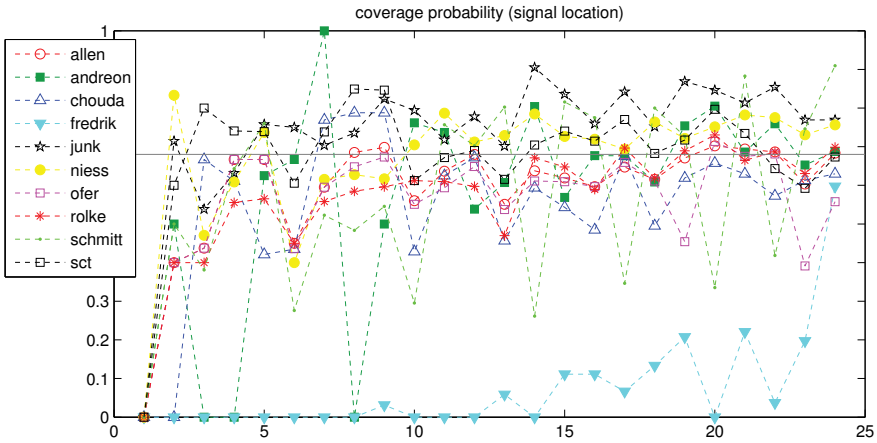


Fig. 5: Fractions of simulated datasets in which the 68% CL intervals quoted by the participants contain the true value of E, the signal location parameter. The 24 signal categories are sorted in order of increasing average correct assignment probability. The author would like to thank Ofer Vitells for preparing this figure.

5 Performance Summary – Problem 2

The solutions to Problem 2 had a broader spectrum of performance than for Problem 1. This is largely due to the ambiguity in specifying the model of the signal and background when only a Monte Carlo model is available. In practice, all an experimentalist has on hand for many signal and background predictions are finite Monte Carlo samples simulated to model these processes. If the size of the Monte Carlo samples is inadequate, larger samples can usually be generated. The data sample helps constrain systematic mismodeling features of the Monte Carlo simulation by comparing observed and expected rates and distributions for events which fail the main selection requirements but pass others designed to select events that can test the features of the Monte Carlo. For Challenge Problem 2, the average total number of background events is of order 1000, and so corresponding Monte Carlo model samples of size 5000 are in a typical ratio to actual data.

Some of the submissions parameterized the distributions of the marks for the signals and backgrounds using analytic functions, and others used binned likelihoods. It was found that the methods that parameterized the shapes as beta functions tended to undercover on the Type-I error rate. The beta function parameterization likely mis-predicts the density of the marks near zero. Participants that ran into this issue re-performed their analysis of the simulated datasets using their method and the true spectrum once the results were distributed, and achieved an error rate of 1%. Coverages for the signal rate fits are provided on the problem web page [4].

6 Summary

The solutions to the Banff Challenge 2 problems provided by the participants span a range of different approaches. Most of the hypothesis tests are based on ratios of profile likelihoods, with Monte Carlo simulation of the distribution of the test statistic. Minor variations between submissions arise from the choice of binning or unbinned fits, and the strategy used to find a global minimum among many local minima in the first problem, and in the parameterization and handling of the distributions of the marks in the second problem. Alternate approaches involved counting events inside signal windows while fitting backgrounds in the sidebands, counting fractional events, and using the Kolmogorov-Smirnov test statistic, and Bayesian methods. Bayesian methods do not naturally focus on error rates, which are frequentist concepts, making the problem setup somewhat clumsy for Bayesian analysis.

The Look-Elsewhere Effect is an issue in Problem 1 (but not in Problem 2) since the presence of

Table 4: Listing of the Type-I error rates, and the claimed and measured correct-discovery rates for the signal scenario Problem 2 for which the participants were asked to estimate their discovery power. Stefan Schmitt states that the power of his 50-bin test is similar to that of his 25-bin test.

Contributor	Type-I Error Rate	Signal = 75 Events	
	Measured	Claimed	Measured
Tom Junk	0.0068 ± 0.0006	0.865	0.870 ± 0.017
Wolfgang Rolke	0.0256 ± 0.0012	0.88	0.850 ± 0.018
Stanford Challenge Team	0.0389 ± 0.0015	0.84	0.9100 ± 0.0143
Eilam Gross & Ofer Vitells	0.0107 ± 0.0008	0.815	0.7725 ± 0.0210
Valentin Niess	0.0085 ± 0.0007	0.761 ± 0.001	0.7125 ± 0.0226
Stefan Schmitt			
25 Bins	0.0047 ± 0.0005	0.85	0.8200 ± 0.0192
50 Bins	0.0047 ± 0.0005		0.8250 ± 0.0190
Doug Applegate & Matt Bellis	0.0168 ± 0.0010	0.95	0.8950 ± 0.0153

a signal introduces an additional parameter – the location of the peak E in the test hypothesis which is not present in the null hypothesis. All participants handled this effect rather well – there are no signs of noticeable undercoverage in the Type-I error rate measurements. One of the methods of accounting for the LEE had the effect of producing p values in excess of unity however.

A typical HEP experiment uses a flip-flopping approach to decide when to quote a two-sided interval and when to quote a one-sided upper limit. The part of the Challenge specification asking for two-sided intervals when evidence was claimed and otherwise not did not allow a unified approach. This request biased the intervals on the rate parameter upwards, most noticeably in the simulated datasets drawn from the null hypothesis. Quoting a two-sided interval for the production rate of a new particle for which evidence is not claimed can be misconstrued by the broader community, even though doing so would help the coverage properties of the methods. Nonetheless, most of the methods provided solutions that undercovered for the signal rate and location parameters for Problem 1.

It is in Problem 2 that significant undercoverage in the main result, quoting evidence or not, meaning a higher-than-expected Type-I error rate, was seen in several submissions. Participants both underestimated their Type-I error rates and overestimated their discovery power. Because the distributions of the marks were not given to the participants, instead relying on simulated Monte Carlo samples of them, participants either binned the data or calculated unbinned likelihoods using parameterizations that appear to fit the distributions of the marks in the simulated Monte Carlo samples. If these parameterizations do not match the true distribution (and they are guesses since the true distribution is hidden), they could, and did, result in poor estimates of the Type-I and Type-II error rates. It could also be that the *a priori* uncertainty of 100% on the rate of Background 2 causes ambiguities to arise in the approach to follow that is reflected measurably in the results, particularly since Background 2 looks more like the signal than Background 1 does.

In general, the methods did very well – an impressive array of approaches, conscientiously applied, gave similar performances and for the most part met the specifications set forth in the Challenge. The Challenge's goal of giving practice on a realistic set of problems is well met. There are many different possible metrics for success on the Challenge problems, just as there are in a real HEP experiment, and no participant's solution came out on top in every possible metric. In a real HEP experiment, the statistical methods used for discovery and exclusion must be approved by the collaboration and reviewed by journal editors and referees. This Challenge provides useful practice in developing, applying, and characterizing

techniques which can be used to test for new phenomena.

Acknowledgements

The author would like to thank the Banff Challenge 2 team, Louis Lyons, Richard Lockhart, Jim Linnemann, and Wade Fisher. This work was performed at Fermi National Accelerator Laboratory, operated by Fermi Research Alliance, LLC under Contract No. DE-AC02-07CH11359 with the United States Department of Energy.

References

[1] J. Heinrich, CERN Report CERN-2008-001, p. 125 (2008). http://cdsweb.cern.ch/record/1021125?ln=en

[2] L. Lyons, Annals of Applied Statistics **2**, 887 (2008);
L. Demortier, Proceedings of PHYSTAT-LHC 2007, CERN-2008-001.

[3] T. Junk, http://www-cdf.fnal.gov/~trj/

[4] T. Junk, http://www-cdf.fnal.gov/~trj/bc2sub/bc2sub.html

[5] F. James, CERN Program Library Long Writeup D506, Version 94.1 (1998). http://wwwasdoc.web.cern.ch/wwwasdoc/minuit/minmain.html.

[6] G. Choudalakis, [arXiv:1101.0390 [physics.data-an]].

[7] ATLAS Collaboration, Phys. Rev. Lett. **105**, 161801 (2010). [arXiv:1008.2461 [hep-ex]].

[8] E. Gross and O. Vitells, "Trial factors for the look elsewhere effect in high energy physics", Eur. Phys. J. C **70**, 525 (2010).

[9] P. Bock, JHEP **0701** (2007) 080 [arXiv:hep-ex/0405072].

Experience from Searches at the Tevatron

Harrison B. Prosper
Department of Physics. Florida State University, Tallahassee, USA

Abstract
I describe, by way of examples, the experience physicists have gained during two decades of searching for physics, both expected and new, at the Fermilab Tevatron.

1 Introduction

2011 marks the end of the Tevatron program [1] and the rapid rise of the Large Hadron Collider (LHC) [2] as the preeminent accelerator in the world. On 22 April, 2011, the LHC reached a luminosity of $4.67 \times 10^{32}\,\mathrm{cm}^{-2}\mathrm{s}^{-1}$, beating the record set by the Tevatron in 2010. Towards the end of the same month, the LHC produced in one week more than double the integrated luminosity of the datasets that yielded the top quark discovery [3]. The era of the Large Hadron Collider is definitely here; 2011 may be remembered not only as a significant year of transition in high energy physics but perhaps also as the year in which the Standard Model (SM) was finally dethroned.

We have reached this crossroad in large measure because of the achievements of physicists at the Tevatron and other accelerator centres around the world. The goals of the Tevatron program were principally to test the SM and to search for significant deviations from it. Alas, none were found. Rather, numerous predictions of the SM have been confirmed, including

1. the shape of jet transverse momentum spectra,
2. the existence of a 6^{th} quark, the top,
3. the existence of reactions in which top quarks are produced singly,
4. the existence of reactions yielding di-bosons (WW, ZZ, WZ, Wγ, Zγ),
5. and properties of B mesons.

These achievements, along with several precision measurements, have established the Standard Model as one of humanity's crowning intellectual achievements [4]. The quantitative agreement between the predictions of the theory and observations is stunning, witness Fig. 1, which shows a comparison of the SM predictions for the jet transverse momentum (p_T) spectra — of jets produced in 1.96 TeV proton antiproton collisions — with the unfolded [5] measurements of the D0 Collaboration. The unfolded results agree with the SM predictions over a dynamic range of 10 orders of magnitude. When searching for new physics, it is not surprising that we take the SM, the null hypothesis, very seriously!

Physicists at the LHC are engaged in an intense search for deviations from the SM, continuing the eclectic approach to searches established at the Tevatron. The Tevatron era is drawing to a close, while that of the LHC is ramping up. Given the theme of this meeting, it is an opportune moment to take stock of the statistical procedures we have used in searches at the Tevatron. This experience may inform what we do at the LHC. One purpose of these proceedings is to encourage closer reflection on what we mean when we say we have found something with "high statistical significance". In this paper, I describe the use of statistical procedures at the Tevatron, in the context of searches, using four case studies: a search for a rare decay of a particle, the search for single top, the search for B_s^0 oscillations, and the search for the Higgs boson.

2 Case Studies

I have chosen to describe four somewhat disparate topics in order to illustrate both the similarities and differences in the statistical approaches that have been pursued at the Tevatron. In reviewing the many

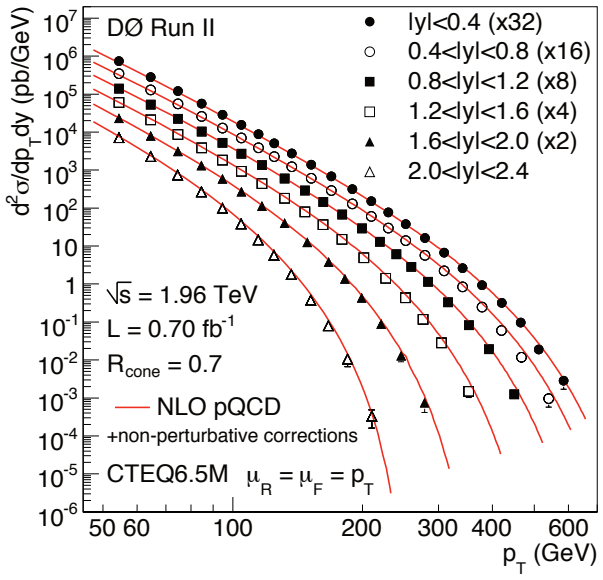


Fig. 1: Comparison of the observed spectrum of jets (points with error bars) with the predictions of the Standard Model (curves) [6].

searches that have been conducted during the period 1991—2011, one notices an interesting sociological evolution. At the start of that period, statistical procedures tended to be described algorithmically with essentially no mention of what statistical procedure was being used nor what quantity was being calculated. Towards the end of that period, however—and one would like to think that this is due in part to the influence of the PHYSTAT series of conferences—words such as *frequentist*, *Bayesian*, *coverage*, *p-value*, *nuisance parameters*, *profile likelihood*, *prior*, etc., began to appear in a few high-profile physics publications. Since these words are now an accepted part of the lexicon of analysis, I shall use them freely in describing the case studies, whether or not such jargon was used in the cited publications.

Another interesting aspect of the statistical work at the Tevatron, and typical of the field, is that almost all hypotheses tested have been *nested* in that the null hypothesis is a special case of the alternative. The canonical example is the search for a signal s above some background μ. The null hypothesis of no signal, $s = 0$, is nested within the alternative hypotheses that the expected event count is $s + \mu$.

2.0.1 Particle Physics Data

From a statistical viewpoint, high energy physicists perform near-perfect Bernoulli trials, tens of millions of times every second. A trial in the context of high energy physics is a collision between particles—protons against antiprotons at the Tevatron and protons against protons or heavy ions against heavy ions at LHC, while a success is some desired outcome. A success could be say the creation of a Higgs boson one of whose decay products (perhaps a muon) has a momentum that falls within a given momentum bin. Each collision yields about 1MB of data. However, of the tens of millions of collisions that occur per second, it is feasible to record only a few hundred per second. The trick, of course, is to ensure that the ones recorded are potentially the most interesting. The data from each collision, that is, *event*, are compressed by a factor of 10^3–10^4 during a process called *event reconstruction*, the goal of which is to infer from the raw data the characteristics of the particles that emanated from the collision point. The cartoon in Fig. 2 illustrates how, ideally, different species of particles are manifested in the particle detectors. It is from the known patterns of particle/detector interactions that the identity of particles can be inferred.

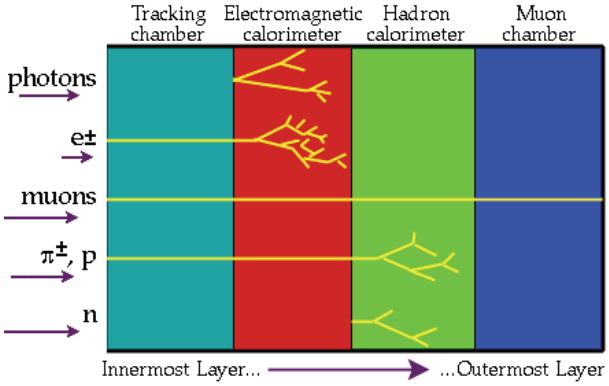


Fig. 2: This cartoon illustrates how, ideally, each species of particle interacts with different components of the detector. In practice, the manner in which particles interact with the detector components is not as clear-cut as this cartoon suggests; ambiguities can arise that lead to particle mis-identification—a jet, for example, could be misconstrued as an electron. (Courtesy CDF Collaboration.)

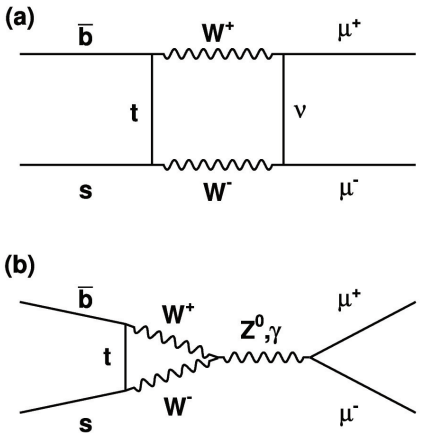


Fig. 3: These diagrams [7] depict a prediction of the SM: the annihilation of a $\bar{b}$ quark and an s quark to a current with zero net charge (a neutral current) that materializes into a pair of oppositely charged muons. Diagram (a) is referred to as a *box* diagram for obvious reasons and diagram (b) is called a *penguin* diagram for reasons that require some imagination.

2.1 Search for a Rare Decay

The search for rare processes, such as the search by the D0 Collaboration described here, is a potentially fruitful way to look for new physics. In many theories of possible new physics, the rates for processes that are rare in the SM are typically predicted to be much higher. Therefore, the observation of a decay rate that differs significantly from the SM prediction would be unambiguous evidence of new physics.

The goal of the search by D0 [7] was to test the SM prediction,

$$\mathcal{B} = \frac{B_s^0 \to \mu^+\mu^-}{B_s^0 \to \text{ anything}} = (3.6 \pm 0.3) \times 10^{-9}. \quad (1)$$

The decay $B_s^0 \to \mu^+\mu^-$ is an example of a process in which there is an apparent neutral current (that is, a current with a net charge of zero) between quarks of different flavor, here the $\bar{b}$ and s quarks. This is an example of a so-called flavor changing neutral current (FCNC) interaction, which are rare in the SM. The lowest order Feynman diagrams describing $B_s^0 \to \mu^+\mu^-$ are shown in Fig. 3. Table 1 shows the results obtained by D0. These data are described by the 2-count likelihood model

Table 1: D0 results for $B_s^0 \to \mu^+\mu^-$: observed event counts, estimated background counts and the scale factors that relate the branching fraction $\mathcal{B}$ to the signals, via $\mathcal{B} = f_i\, s_i$ with $i = a, b$. The subscripts pertain to the two Tevatron run periods, RunIIa and RunIIb.

Run period	observed count (events)	estimated background count (events)	estimated scale factors ($\times 10^{-9}$)
RunIIa	$n_a = 256$	264 ± 13	4.90 ± 1.00
RunIIb	$n_b = 823$	827 ± 23	1.84 ± 0.36

$$p(n|s,\mu) = \text{Poisson}(n_a|s_a + \mu_a)\,\text{Poisson}(n_b|s_b + \mu_b), \tag{2}$$

where n, s and μ are the observed counts, expected signal and expected background counts, respectively. The branching fraction $\mathcal{B}$ is related to the expected signals through scale factors f_i, where $\mathcal{B} = f_i\, s_i$ with $i = a, b$. The likelihood in Eq. (2) therefore contains one parameter of interest, namely the branching fraction $\mathcal{B}$, and the four nuisance parameters f_a, f_b, μ_a, and μ_b. Information about the nuisance parameters is encoded in an *evidence-based* prior $\pi(f_a, f_b, \mu_a, \mu_b)$, modeled as the product of four normal distributions with the means and standard deviations listed in Table 1, one set for each nuisance parameter. (Given the size of the uncertainties for the scale factors, listed in Table 1, the priors for these parameters were, in fact, truncated Gaussians.)

The likelihood in Eq. (2) was marginalized with respect to the nuisance parameters, f_a, f_b, μ_a, and μ_b to yield the marginal likelihood $p(n|\mathcal{B})$. From this, the limit $\mathcal{B} < 5.1 \times 10^{-8}$ at 95% C.L. was derived using the CL_s method [8]. In the CL_s method one defines the tail probability

$$p_1(\mathcal{B}) = \Pr[t < t_0|H_1(\mathcal{B})], \tag{3}$$

for some suitable statistic t, for a given (alternative) hypothesis H_1 about the branching fraction $\mathcal{B}$. One then rejects all values of $\mathcal{B}$ for which $p_1(\mathcal{B}) < \gamma p_1(0)$ and *defines* a $(1-\gamma)$ C.L. upper limit as the smallest rejected value of $\mathcal{B}$. The statistic used by D0 is the logarithm of the *Bayes factor* $p(n|\mathcal{B})/p(n|0)$. (It is a Bayes factor rather than a likelihood ratio because the marginal likelihoods entail integrations over priors.)

2.2 Search for Single Top

The goal of this search is to test the SM prediction that the process

$$p + \bar{p} \to t + X, \tag{4}$$

exists in which the set of particles denoted by X does *not* contain a top quark. (The top quark was discovered [3] through the reaction $p + \bar{p} \to t\bar{t}$.) The SM predicts how often the reaction in Eq. (4) should occur, which is quantified in terms of the cross section $\sigma(p + \bar{p} \to t + X) = 3.46 \pm 0.18$ pb (assuming a top quark mass of 170 GeV). At the Tevatron, this cross section corresponds to a production rate of about 1 in 10 billion collisions, which is just under half the rate for the pair production of top quarks. It would seem therefore that the search for single top ought not be that much harder than was the search for top quark pairs ($t\bar{t}$). In fact, owing to the greater similarity between the signal and background events, the search for single top proved to be considerably more challenging. This is illustrated in Fig. 4, which shows the event sample composition before b-tagging (that is, before selecting events with identified b-quark jets), but after the first level of cuts. The signal to background ratio at this stage was a daunting $1 : 260$.

It was clear from the outset, that only the most sophisticated methods of analysis were likely to yield a successful outcome in a reasonable amount of time. Indeed, the first evidence of the existence of single top reactions [9] and their subsequent definitive observation by CDF [10] and D0 [11] both made

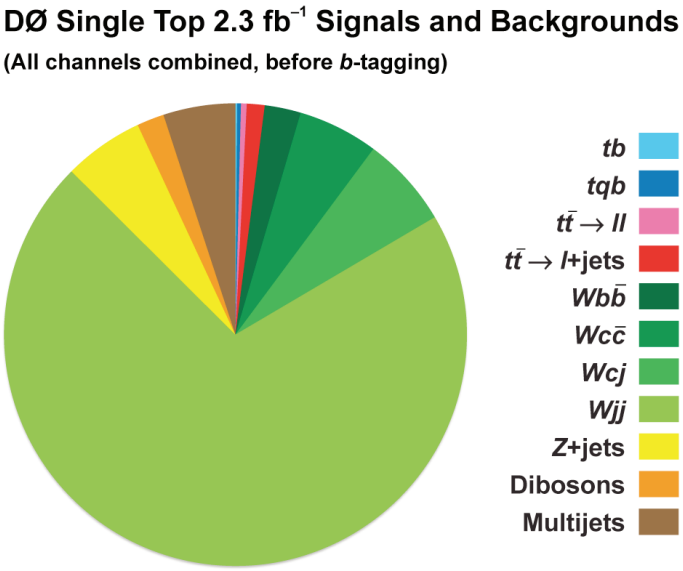


Fig. 4: Predicted composition of the D0 data that were the basis of the single top discovery. CDF predicted a similar composition. The single top signal is the thin wedge at the top of the pie-chart. (Courtesy D0 Collaboration.)

extensive use of multivariate discrimination methods such as boosted decision trees (BDT), Bayesian neural networks (BNN), and *ab initio* semi-analytical calculations of the signal and background probability densities, the so-called Matrix Element (ME) method. This was the first time in high energy physics that a major discovery was based on such methods. The extensive use of Bayesian methods, by D0, was another first.

After reducing the multivariate data x to a discriminant function $D(x)$, the data were binned into M bins in the variable D (see Fig. 5). The M counts are described by a likelihood function similar in structure to that used in the rare decay search (see Section 2.1),

$$p(n|\sigma,\epsilon,\mu) = \prod_{i=1}^{M} \text{Poisson}(n_i|\epsilon_i\,\sigma + \mu_i), \tag{5}$$

where σ, the single top cross section, is the parameter of interest and the $2M$ nuisance parameters ϵ_i and μ_i, respectively, are the expected effective integrated luminosities (integrated luminosity $\times$ signal efficiency $\times$ signal acceptance) and the expected background counts, respectively, while n_i are the observed bin counts. Information about the nuisance parameters was encoded in an evidence-based prior $\pi(\epsilon,\mu)$ modeled as a multivariate normal distribution that took account of the known correlations between the nuisance parameters. The overall prior $\pi(\sigma,\epsilon,\mu)$ was factorized as follows $\pi(\sigma,\epsilon,\mu) = \pi(\epsilon,\mu|\sigma)\,\pi(\sigma) = \pi(\epsilon,\mu)\,\pi(\sigma)$ and $\pi(\sigma)$ was taken to be a *flat* prior.

The posterior density resulting from the integration over the nuisance parameters is shown in Fig. 6. The D0 analysts considered Bayes factors, $p(n|\sigma)/p(n|0)$, but chose, in the end, to follow tradition and estimate the significance of the single top observations using a prior-predictive p-value, $p_0 = \Pr[t > t_0|H_0]$, computed using a null hypothesis (H_0) in which the expected background is marginalized with respect to the background prior. The statistic t (which of course could have been any suitable function of the data) was taken to be the mode of the posterior density, $p(\sigma|n)$. The basic intuition is that larger values of the cross section σ cast greater doubt on the null, that is, on the background-only hypothesis.

The distribution of t, shown in Fig. 7, was simulated using 67.8 million pseudo datasets, generated with background only, which, for the measured cross section of 3.94 pb, yielded a prior predictive p-value

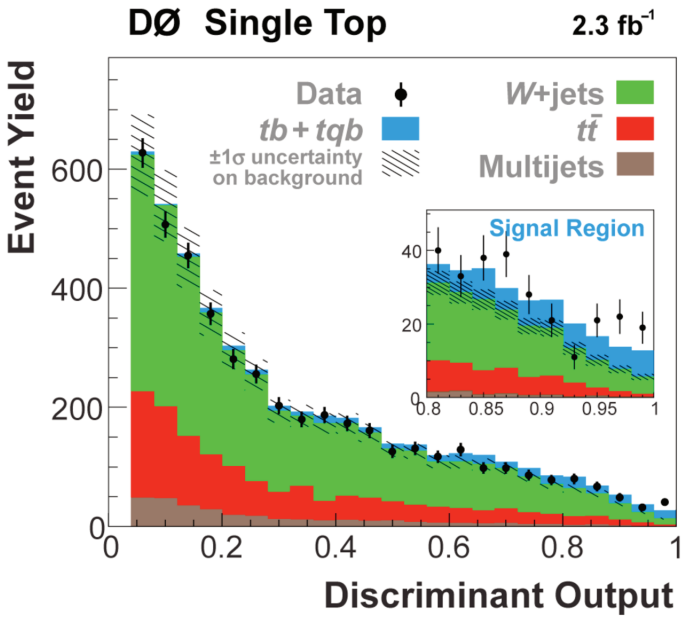


Fig. 5: Distribution of the Bayesian neural network discriminant that combines the discriminants from three separate, but correlated, D0 analyses (BDT, BNN, and ME) [11] into a single overall discriminant $D(x)$. The single top signal is the thin line that becomes slightly more visible as one moves towards the signal region, $D(x) \to 1$. (Courtesy D0 Collaboration.)

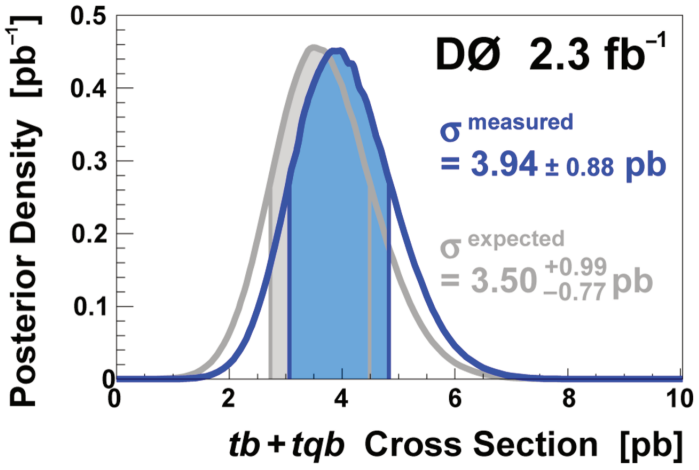


Fig. 6: The posterior density $p(\sigma|n)$ is plotted as a function of the total cross section ($\sigma(p + \bar{p} \to tb) + \sigma(p + \bar{p} \to tqb)$) assuming the SM prediction for the ratio of the tb and tqb cross sections. The expected cross section is computed using an *Asimov* dataset (see Cowan, these Proceedings), that is, an artificial dataset in which the "observed" counts are set equal to the sum of the expected background and signal counts, assuming the SM prediction for the signal. (Courtesy D0 Collaboration.)

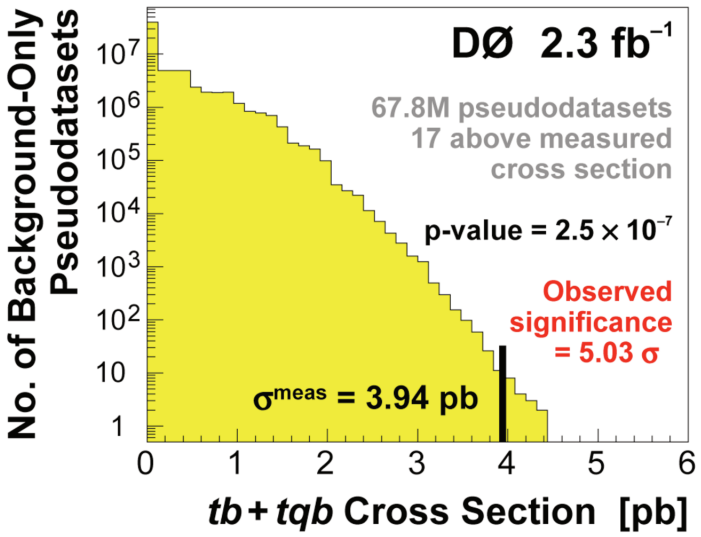


Fig. 7: Distribution of the statistic $t = \text{mode}[p(\sigma|n)]$ over a background-only ensemble of pseudo datasets. (Courtesy D0 Collaboration.)

of 2.5×10^{-7}. Again, by tradition, this was converted to a normal standard deviation scale as an observed significance of 5 standard deviations, the long-accepted threshold in high energy physics for claiming a discovery.

2.3 Search for B_s^0 Oscillations

The goal of this search by the CDF Collaboration [12] was to test the SM prediction that the B_s^0 and $\bar{B}_s^0$ mesons form an oscillating pair in which each member of the pair changes into its partner at an (extremely) high frequency predicted by the SM. The oscillations are governed by the time-dependent probability densities,

$$\begin{aligned} p_{B_s^0 \to \bar{B}_s^0}(t|A, \Delta m) &= \frac{1}{2\tau} e^{-t/\tau}[1 - A\cos\Delta m t], \\ p_{B_s^0 \to B_s^0}(t|A, \Delta m) &= \frac{1}{2\tau} e^{-t/\tau}[1 + A\cos\Delta m t], \end{aligned} \tag{6}$$

where A is the amplitude of the oscillations and Δm characterizes its frequency. According to the SM, $A = 1$.

There were two important complications with this search. Firstly, the measured time of decay t of a B meson—inferred from the measured displacement of the B meson decay point from the proton antiproton collision point—was measured with an uncertainty that varied from meson to meson. The uncertainty on t was modeled with a normal distribution with a *heteroscedastic* variance, that is, one that varied from one measurement to the next. Secondly, the oscillation signal was contaminated with background arising from other processes. The probability model for t was therefore taken to be a convolution of a normal density with a mixture model comprising an oscillatory signal plus a background. The likelihood for these data is then just a product of M terms, one for each measured time t_i,

$$p(t|A, \Delta m) \sim \prod_{i=1}^{M} \mathrm{N}(t_i|t', \sigma_i^2) \otimes [\alpha\, p(t'|A, \Delta m) + (1 - \alpha)\, b(t')], \tag{7}$$

where α is the signal fraction and $b(t)$ the background density. Since the oscillation frequency Δm is predicted to be high (about 18 cycles per picosecond), it proved more satisfactory to perform a maximum

likelihood fit for the amplitude A, using $p(t|A, \Delta m)$, for different fixed values of Δm. It was found that at $\Delta m = 17.8$ cycles / ps, $A = 1.21 \pm 0.20$, which is consistent with the SM prediction $A = 1$ and inconsistent with $A = 0$. The amplitude A was then set to unity and Δm was measured to be $17.77 \pm 0.10(\text{stat}) \pm 0.07(\text{syst})$ cycles / ps.

The CDF Collaboration quantified the statistical significance of its results using the p-value $p_0 = \Pr[\Lambda < \Lambda_0|H_0]$ based on the statistic $\Lambda = \log[p(t|H_0)/p(t|H_1, \Delta m)]$, where $p(t|H_0) \equiv p(t|A = 0)$ and $p(t|H_1, \Delta m) \equiv p(t|A = 1, \Delta m)$ are the densities for the null and alternative hypotheses, respectively. Small values of Λ provide evidence against the null. At $\Delta m = 17.8$ cycles / ps, the value of the test statistic Λ was observed to be -17.26 [12] from which the p-value of 8×10^{-8} was calculated by a Monte Carlo technique. This p-value, being rather smaller than the traditional threshold for discovery, fully justified the title "Observation of $B_s^0 - \bar{B}_s^0$ Oscillations" of the CDF article announcing this result.

2.4 Search for the Higgs

Since the start of the current millennium—and building on the searches at LEP and earlier machines, high energy physicists have been engaged in a relentlessly intensifying search for the Higgs boson. This particle, or something that mimics it, is a critical ingredient of the SM, being a vestige of the mechanism through which mass is introduced into a theory that would otherwise describe an unrealistic world of massless particles. Its fundamental role in the SM is reason enough to sustain the Higgs search effort that began at LEP and the Tevatron and that continues apace at the LHC.

But, of course, the Higgs boson may not exist. From a certain point of view, it would be a spectacularly exciting outcome were it to be shown convincingly that no such particle exists with a mass less than about 1 TeV. On the other hand, finding it rather than not finding it is pretty exciting too! If a low-mass neutral Higgs boson exists, we would be in a position akin to that during the search for the top quark. During that search, we "knew" everything about the top quark since all of its characteristics, with the exception of its mass, were predicted in detail from the SM. Moreover, the mass of the top quark was inferred from radiative corrections to precision measurements. Likewise for the Higgs searches: if a SM Higgs boson exists, we know a lot about it [13]. Indeed, the searches for the Higgs boson rely extensively on detailed predictions from the SM. When the Tevatron data from CDF and D0 are analyzed in the context of the SM, one obtains the results shown in the left plot of Fig. 8. In this figure is plotted the 95% credible level (C.L.) upper limit R^{up}, given by

$$0.95 = \int_0^{R^{up}} p(R|n, m_H)\, dR, \tag{8}$$

as a function of the Higgs mass hypothesis, where the posterior density is given by

$$p(R|n, m_H) \propto \prod_{i=1}^{N_C} \prod_{j=1}^{N_{bi}} \text{Poisson}(n_{ij}|R\, s_{ij} + \mu_{ij})\, \pi(R, s_{ij}, \mu_{ij}, m_H), \tag{9}$$

and $R \equiv \sigma/\sigma_{\text{SM}}$, with σ_{SM} the predicted SM cross section for the creation of Higgs bosons of a given mass. The index i ranges from 1 to N_C final state channels, while the index j is over the N_{bi} data bins in the i^{th} channel. The quantities s_{ij} are the predicted signals, for a given Higgs boson mass, assuming the validity of the Standard Model. The prior incorporates the uncertainty in these predictions. Systematic uncertainties can be incorporated by representing the prior $\pi(R, s, \mu, m_H)$ as an integration,

$$\pi(R, s, \mu, m_H) = \int \pi(R, s, \mu, m_H|\theta)\, \pi(\theta)\, d\theta, \tag{10}$$

with respect to (hyper) parameters θ that characterize the systematic effects.

In the right plot in Fig. 8 is displayed a summary of the LEP results: the negative log-likelihood as a function of the Higgs mass.

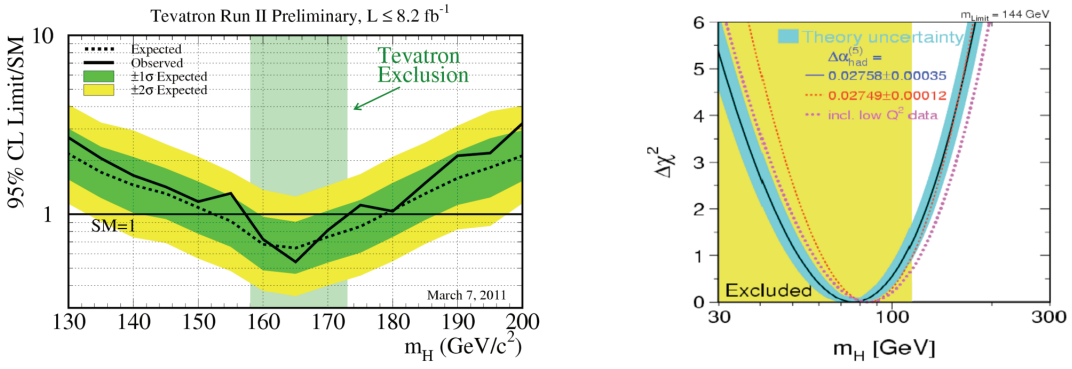


Fig. 8: (left) The plot summarizes the conclusions of the Tevatron New Phenomena and Higgs Working Group (TEVNPHWG) regarding the SM Higgs boson. The dark line is the observed upper limit from a Bayesian calculation, while the bands indicate by how much the limits may be expected to fluctuate in a large number of simulated repetitions of the (combined) CDF and D0 experiments. Note: the bands depend on the ensemble used to calculate them. Therefore, a different ensemble would yield different bands. It is useful to keep this mind and, it is hoped, keep in check the temptation to over-interpret the bands, useful though they are. (right) The plot summarizes the conclusions of the LEP experiments. The negative log-likelihood curve, which favours a low-mass Higgs, is the result of a fit to the LEP precision measurements. The vertical band on the left of the plot ends at 114 GeV, the lower limit set by the LEP experiments as a result of direct searches for the SM Higgs boson. (Courtesy TEVNPHWG.)

So what, if anything, can we say about the Higgs hypothesis, given the assumption that its description in the SM is correct? An answer to this question would require a coherent integration of the information presented in Fig. 8. In principle, the LEP curve provides the probability density $p(m_H|H_1)$ (assuming a flat prior in the Higgs mass—though a reference prior would be better), while the Tevatron results provide $p(s|m_H, H_1)$, where now s denotes the expected Higgs signal and H_1 denotes the Higgs hypothesis and the numerous assumptions on which the LEP and Tevatron results depend[1].

Given the densities, $p(m_H|H_1)$ and $p(s|m_H, H_1)$, it would be natural from a Bayesian viewpoint to compute the expected signal density,

$$p(s|H_1) = \int_{100}^{200} p(s|m_H, H_1)\, p(m_H|H_1)\, dm_H, \tag{11}$$

by marginalizing over the Higgs boson mass, m_H. Then, physicists at the LHC, if so inclined, could use $p(s|H_1)$ as an evidence-based prior $\pi(s)$ in their LHC Higgs searches. How might it be used? It could be used, for example, in conjunction with LHC likelihoods to test the Higgs hypothesis using a Bayes factor (see Berger, these Proceedings),

$$B_{10} = \int_0^\infty p(\text{LHC-data}|s+\mu)\, \pi(\mu)\, \pi(s)\, d\mu\, ds \Big/ \int_0^\infty p(\text{LHC-data}|\mu)\, \pi(\mu)\, d\mu, \tag{12}$$

which can be mapped to a scale akin to n-sigma using the transformation $Z = \sqrt{2 \log B_{10}}$.

3 Conclusions

Numerous discoveries have been made at the Tevatron in spite of our eclectic (and sometimes baroque) approach to interpreting results in a statistical manner. If there is one theme throughout, it is that we

[1] In practice, because these curves are summaries, they do not provide enough information to actually carry out the integration of this information. Here is a compelling case for making the full probability model, plus the observations, available using, for example, `RooStats` workspaces.

remain ferociously fond of exact frequentist coverage. Hence the Herculean efforts to achieve “exact" 5 standard deviation results. Moreover, p-values remain the principle measure of “surprise": if a p-value is small enough, we judge that something surprising and presumably exciting has happened. Rarely has the notion of *power* been explicitly addressed in Tevatron analyses, though simple measures of experimental sensitivity have become routine, such as the notion of expected limits. The Poisson model remains ubiquitous as does the use of the normal distribution as a model for systematic uncertainties. However, there is a growing realization that we can, and should, do a better job of designing probability models using more appropriate functions, such as gamma or log-normal densities, for modeling systematic uncertainties. The `RooFit/RooStats` system now makes this possible (see Schott, these proceedings).

Bayesian methods have made significant inroads, witness for example the discovery of single top by DØ, which was Bayesian through and through, until the very end when a p-value was used to quantify the significance of the observations.

Physicists are still prone to statistical invention, even when perfectly satisfactory alternatives exist. But the good news is that we can be taught!

Acknowledgements

I wish to thank Luc Demortier and the organizers of this very interesting conference, in particular, Louis Lyons and Albert de Roeck.

References

[1] The Tevatron, Fermilab, `http://www.fnal.gov/pub/science/accelerator`.

[2] The Large Hadron Collider, CERN, `http://public.web.cern.ch/public/en/lhc/lhc-en.html`.

[3] CDF Collaboration, F. Abe *et al.*, Phys. Rev. Lett. **74**, 2626 (1995); D0 Collaboration, S. Abachi *et al.*, Phys. Rev. Lett. **74**, 2632 (1995).

[4] See for example, J. Lefrancois, “The standard model: 30 years of glory" in Proceedings: Techniques And Concepts In High-Energy Physics, Eds. H.B. Prosper and M. Danilov, (Dordrecht, The Netherlands, Kluwer, 2001, 411p., Nato Science Series C: Mathematical and Physical Sciences, Vol. 566).

[5] See the contributions on unfolding in these Proceedings.

[6] D0 Collaboration, V.M. Abazov, *et al.*, Phys. Rev. Lett. **101**, 062001 (2008).

[7] D0 Collaboration, V.M. Abazov, *et al.*, Phys. Rev. Lett. B**693**, 539 (2010); aXiv:1006.3469 [hep-ex].

[8] Luc Demortier, these Proceedings.

[9] D0 Collaboration, V.M. Abazov, *et al.*, Phys. Rev. Lett. **98**, 18102 (2007); Phys. Rev. D**68**, 012005 (2008).

[10] CDF Collaboration, T. Aaltonen *et al.*, Phys. Rev. Lett. **103**, 092002 (2009).

[11] D0 Collaboration, V.M. Abazov *et al.*, Phys. Rev. Lett. **103**, 092001 (2009).

[12] CDF Collaboration, A. Abulencia *et al,*, Phys. Rev. Lett. **97**, 242003 (2006).

[13] TEVNPHWG, `http://tevnphwg.fnal.gov`.

Statistical methods used on searches at LHCb, with special emphasis in the search for the very rare decay $B_s \to \mu^+\mu^-$

Jose A. Hernando, on behalf of the LHCb collaboration
Universidade de Santiago de Compostela, Spain

Abstract

The LHCb experiment searches for new physics in CP violation and rare decay processes of B and D mesons. We describe here the strategy followed to search for the rare decay $B_s^0 \to \mu^+\mu^-$. This is one of the key analyses of LHCb and serves as a model for other LHCb searches. Emphasis will be put on the statistical methods used.

1 Introduction

The LHCb experiment is one of the four experiments located at the Large Hadron Collider (LHC) at CERN. The experiment is designed to search for new physics (NP) beyond the Standard Model (SM) in charge-parity (CP) violation and rare decay processes of beauty (B) and charm (D) mesons. During 2010, the experiment collected 37 pb^{-1} of integrated luminosity. One of the first measurements published by the LHCb collaboration was the value of the $b\bar{b}$ cross-section $\sigma(pp \to b\bar{b}X) = (284 \pm 20 \pm 49)\,\mu\mathrm{b}$ [1]. This cross-section is high enough to produce thousands of B mesons per second at the nominal luminosity $\mathcal{L} = 2 - 5 \times 10^{32}\ \mathrm{cm}^{-2}\mathrm{s}^{-1}$. With these large statistical samples, the physics reach of the LHCb experiment does not suffer significantly from the fact that the LHC runs at $\sqrt{s} = 7$ TeV.

The LHCb detector has performed beautifully during its first year of operation. The detector is a forward spectrometer with a vertex detector, a tracking system (before and after a warm dipole magnet), two RICH detectors, an electromagnetic and hadronic calorimeter and a muon system. A detailed description of the detector can be found in Ref. [2]. In order to identify specific B decays and separate them from the large background, the experiment has a flexible trigger system, good particle identification and excellent momentum and vertex resolution. Except for the vertex resolution, where the impact parameter of the tracks has been measured with an uncertainty $\sim 10\%$ greater than expected, the rest of the detector characteristics are within design specifications.

The LHCb physics program includes the search for very rare decays $B^0_{(s)} \to \mu^+\mu^-$, $D \to \mu^+\mu^-$, lepton flavor violating decays $B \to \mu e$, etc. The $B_s^0 \to \mu^+\mu^-$ is one of the key searches of LHCb. The first results [3] obtained with $37\,\mathrm{pb}^{-1}$ integrated luminosity has been recently submitted to Phys. Lett. B. The $B_s^0 \to \mu^+\mu^-$ analysis has defined a strategy that is followed by other LHCb searches. For this reason, and given its mature state, this Paper is dedicated to the description of this analysis. Special emphasis is given to the statistical methods used. However it is a "classical" search as it uses well-known methods. The only exception is the use of a multi-variate method ($\Delta\chi^2$) that will be described here in detail.

2 The $B_s \to \mu^-\mu^+$ search

Within the SM, exclusive dimuon decays of B^0 and B_s^0 mesons occur only via loop diagrams and are helicity suppressed. The SM prediction is $\mathcal{B}(B_s^0 \to \mu^+\mu^-) = (3.2 \pm 0.2) \times 10^{-9}$ [4]. However, within NP models, especially those with an extended Higgs sector, the $\mathcal{B}$ can differ significantly. This is the case, for example, within the minimal supersymmetric SM (MSSM) [5]. The current limits have been set by the CDF and D0 collaborations [6]. The CDF collaboration has presented a preliminary result [7]

with the most stringent limit so far $\mathcal{B}(B_s^0 \to \mu^+\mu^-) < 4.3 \times 10^{-8}$ at 95 % C.L. with $3.7\,\mathrm{fb}^{-1}$ of data analyzed, but this is still an order of magnitude greater than the SM value.

The LHCb experiment is well suited for the search of this decay due to its excellent invariant mass resolution, vertex resolution, muon identification and trigger acceptance. The forward geometry of LHCb allow us to trigger on muons with low transverse momenta. For example, the first level trigger L0, which is a hardware trigger, accepts an event if there is a single muon with $p_T > 1.4$ GeV/c^2 or if there are two muons with $p_T > 0.48\,(0.56)$ GeV/c^2 for the muon with the lowest (highest) p_T. The signal trigger efficiency has been estimated with data to be $(90 \pm 4)\,\%$. $B_s^0 \to \mu^+\mu^-$ events are selected offline with soft criteria designed to remove the most obvious background events while keeping the signal efficiency as high as possible. The selection is based on the existence of two tracks identified as muons, that form a good secondary vertex separated from the primary one by a distance significance (distance/uncertainty) greater than 15. After the trigger and the selection, we expect 18 ± 2 signal events per fb^{-1} according to the SM prediction.

The search for the signal is done in a two-dimensional space. One coordinate of the space is the invariant mass. The second coordinate, which we refer to it as a Geometrical Likelihood (GL), is the output of a multivariate method that combines different discriminant variables taking into account their correlations. The invariant mass resolution of the signal has been measured in control channels and is $26.7 \pm 0.9\,\mathrm{MeV/c^2}$. We expect the overall mass distribution to be a Gaussian distribution for the signal plus an exponential distribution for the background. The background is dominated by random combinations of real muons coming from semileptonic decays of a $b\bar{b}$ pair ($b\bar{b} \to \mu^+\mu^- X$). The GL combines the following variables: the minimum impact parameter of the muons, the distance of closest approach between them, the impact parameter significance of the B candidate, the B proper time and an isolation variable that quantifies if any of the muons is attached to other secondary vertices besides the one from the B. The GL variable has a range between 0 and 1. For the background it peaks at 0 while there are almost no events left for GL> 0.5. The GL has been constructed in such a way that the signal events distribute uniformely between 0 and 1. The GL is defined using a sample of simulated $B_s \to \mu^+\mu^-$ and $b\bar{b} \to \mu^+\mu^- X$ events, but as we will comment later, its distribution has been validated with data. The region GL > 0.5 and with an invariant mass within $60\,\mathrm{MeV/c^2}$ interval around the $B^0_{(s)}$ mass was blinded until the analysis was completely defined.

To construct the GL we use a multivariate method called $\Delta\chi^2$. This method is described in Ref. [8] and there is a first version in Ref. [9]. The method transforms a set of n initial variables $\{x_i\}$ into a set of n new variables $\{s_i\}$ which are distributed according to a Gaussian with zero mean and sigma unity. The transformed variables are mostly uncorrelated and the p.d.f. can be approximated by an n-dimensional Gaussian. Two transformations are defined to separate signal from background: one for the signal events and a second one for background. Given an event, the original $\{x_i\}$ variables are transformed into the $\{s_i\}$ variables using the transformation of the signal, and into the $\{b_i\}$ variables using the transformation of the background. Then we compute the quantities $\chi^2_s = \sum_{i=1}^n s_i^2$ and $\chi^2_b = \sum_{i=1}^n b_i^2$ that are related to the probability that the event is signal or background, and we use the difference between them as the final discriminanting variable $\Delta\chi^2 = \chi^2_s - \chi^2_b$. For practical reasons, we transform the $\Delta\chi^2$ distribution for the signal events into a uniform distribution between 0 and 1.

The process of "Gaussianization" and de-correlation of the input variables is made in two steps. In the first step the initial variables are transformed into gaussian distributed variables. To do so, first the variables are transformed into a uniform distribution using the accumulative function of the distribution of the input variable; and later, they are transformed into a gaussian distribution using the inverse function of the accumulative function of a gaussian distribution. At this stage, the variables are gaussian but they are still correlated. To reduce the correlation, we compute the moments or the symmetry axis of the new variables. Then, we rotate them to the symmetry axis and we re-gaussianize the variables with the process described above. After this point, the variables are now gaussian distributed with unit variance and zero mean and they are mostly uncorrelated. They follow an n-dimensional Gaussian. We can relate

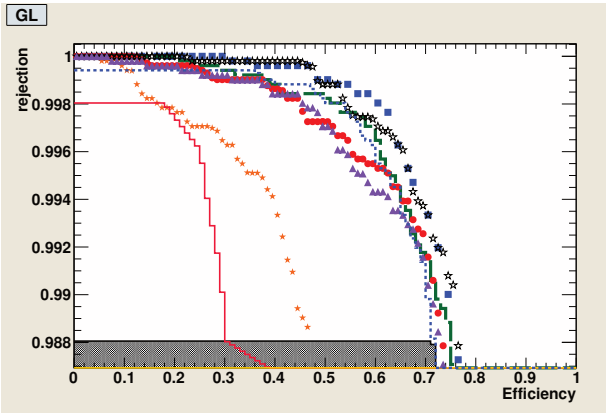


Fig. 1: Performance (rejection of background events *vs* efficiency on signal events) of the $\Delta\chi^2$ method (blue squares) compared with the default configuration of different TMVA methods: BDT (open stars), PDERS (sort dashed), Fisher Discriminant (violet trianges), Best performing NN (red circles), Support Vector Machine (green dashed line), RuleFit (red solid line), FDA (orange stars), kNN (black filled histogram).

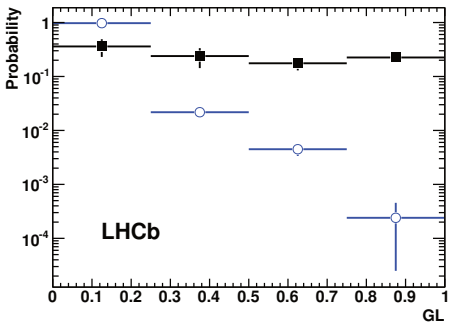


Fig. 2: GL p.d.f for signal (black) and background (blue) calibrated using data.

the probability of an event to belong to that sample (i.e signal) with the $\chi^2_s = \sum_{i=1}^{n} s_i^2$ quantity.

We have compared the performance of this method with some multi-variate methods implemented in the `ROOT TMVA` package [11]. Using as input variables the ones indicated above, the $\Delta\chi^2$ method performs as well as the Boosted Decision Trees (BDT), which of the MVA methods used performs the best. The default configurations of the methods provided by the `TMVA` package were used. The comparisons between all the methods were done under the same conditions. The sample was divided in two identical samples, one used for training the method and the second one to obtain the performance. Figure 1 shows the signal efficiency vs background rejection obtained with the different methods on the signal and background samples. For the background, we used $b\bar{b} \to \mu^+\mu^- X$ simulated events.

One of the strong points of the analysis is the fact that the mass and the GL pdfs have been calibrated with data. The resolution of the B_s mass is obtained from the interpolation between the J/Φ and Υ dimuon resonances. It has also been measured from the mass distribution of $B_s^0 \to K^+K^-$ events. The GL pdf of the signal has been calibrated using $B \to h^+h^{-'}$ events (where $h, h^{'}$ stand for kaon or pion). The selection of these events is identical to the signal ones (except for the muon identification requirement) but they are triggered differently. The effect of the trigger has been corrected using only $B \to h^+h^{-'}$ candidates where the events were triggered not using the candidates themselves. Therefore, they were not biased by the trigger. The GL pdf of the background has been calibrated using the $B_s \to \mu^+\mu^-$ events that are in the sidebands of the invariant mass (they are outside a $60\,\mathrm{MeV/c^2}$ window centered at the B_s mass but inside a larger window of $600\,\mathrm{MeV/c^2}$). The GL calibrated pdfs for

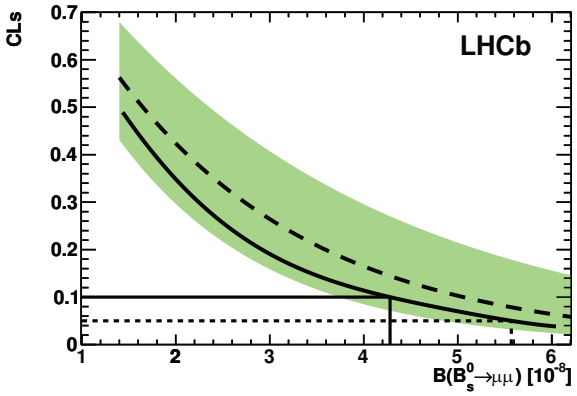


Fig. 3: CL_s vs $\mathcal{B}(B_s^0 \to \mu^+\mu^-)$ expected value if there is no signal events (dashed curve) and observed value (solid curve). The green shaded area contains the $\pm\sigma$ interval of possible results compatible with the expected value when only background is observed. The 90% (95 %) C.L. observed value is shown by the solid (dashed) line.

signal and data are shown in Fig. 2.

The region $\mathrm{GL} > 0.5$ and $60\,\mathrm{MeV/c^2}$ window around the B_s^0 mass has been divided in bins (2 for the GL, and 6 for the mass, equally spaced). In each bin the expected number of background and signal events are computed using the mass and GL calibrated pdfs. We used several control channels $B^+ \to J/\Psi K^+,\ B^0 \to K^+\pi^-,\ B_s \to J/\Psi\phi$ to normalize the signal expected events to a known $\mathcal{B}$ via the relation,

$$\mathcal{B}(B_s^0 \to \mu^+\mu^-) = \mathcal{B}_n \frac{\epsilon_n^{sel}}{\epsilon_s^{sel}} \frac{\epsilon_n^{trg/sel}}{\epsilon_s^{trg/sel}} \frac{f_n}{f_s} \frac{N_s}{N_n}, \tag{1}$$

where $\mathcal{B}_n$ is the branching ratio of any of the normalization channels noted above. The $\frac{\epsilon_n^{sel}}{\epsilon_s^{sel}}$ is the ratio of the selection efficiency (which includes also the acceptance of the detector and the reconstruction efficiency) between the normalization and the signal channel. The $\frac{\epsilon_n^{trig/sel}}{\epsilon_s^{trg/sel}}$ is the ratio of the trigger efficiency on the selected events between the normalization and the signal channels. f_n/f_s is the ratio of the fragmentation fractions, i.e. the ratio of the probabilities that a b quark produces a $B_s(B_n)$ meson. The ratio is 1 when normalizing to B_s, and $f_d/f_s = 3.71 \pm 0.47$ [10] when normalizing to a B^0. Finally N_s and N_n are the number of selected and triggered events for the signal and the control channel, respectively. The first ratio has been computed using MC simulations. The second one has been estimated using the data. Several cross-checks have been performed with data to verify the first ratio. Care has been taken in defining the selection of the normalization of the control channels so that it matches the signal selection as closely as possible in order to minimize the systematic errors.

To set a limit on the $\mathcal{B}$ we have used the CL_s method [12]. The CL_s method is well known in HEP; in particular, it was used in the Higgs searches performed at LEP. It uses as a test-statistic the ratio of the likelihood of the signal plus background hypothesis and the likelihood of the background-only hypothesis. The distribution of the test-statistic of the signal plus background (sb) and background-only (b) hypotheses are used to compute two p-values ($p_{sb} = \mathrm{CL}_{sb},\ p_b = 1 - \mathrm{CL}_b$). The CL_s quantity is computed as the ratio $\mathrm{CL}_s = \frac{\mathrm{CL}_{sb}}{\mathrm{CL}_b}$. This quantity has the advantage (over the pure p_{sb}-value) of not excluding a region where the experiment has no sensitivity to observe a signal. Figure 3 shows the CL_s vs $\mathcal{B}$ when the observation equals the expected number of background events (dashed curve). The shaded area contains the $\pm\sigma$ interval of possible results compatible with the expected value when only background events are observed. The solid curve corresponds to the LHCb observation with $37\,\mathrm{pb}^{-1}$. The horizontal solid (dashed) line corresponds to the 90% (95%) C.L. The LHCb limits are $\mathcal{B}(B_s \to \mu^+\mu^-) < 5.6 \times 10^{-8}$ at 95 % C.L. and $\mathcal{B}(B^0 \to \mu^+\mu^-) < 1.5 \times 10^{-8}$ at 95 % C.L. The systematics errors of the normalization

factor and of the pdfs of the mass and the GL are propagated into the calculation of the $\mathcal{B}$ C.L. using the technique described in Ref. [13]. Currently, the collaboration has started to discuss the possibility of using different methods other than the CL_s to obtain the $\mathcal{B}$ limit.

Previous MC studies [14] have shown the potential of LHCb to observe the SM $B_s^0 \to \mu^+\mu^-$ decay. For the observation we use the p-value of the background-only hypothesis or 1-CL_b. Measuring the SM $\mathcal{B}(B_s \to \mu^+\mu^-)$ at 3 σ it will require collecting more than $2\,\mathrm{fb}^{-1}$ of data.

3 Conclusions

The LHCb detector has performed beautifully during the data taking period of the year 2010 and has collected a data-set corresponding to 37 pb^{-1} of integrated luminosity. The LHCb physics program includes the searches for rare or forbidden B and D meson decays. One of the most relevant LHCb analyses is the measurement of the $\mathcal{B}(B^0_{(s)} \to \mu^+\mu^-)$. The LHCb collaboration has sent for publication in Physics Letter B the first results of this search.

The $B_s \to \mu^+\mu^-$ analysis serves as a model for other LHCb searches. It is based on the definition of a sensitive region, in this case it is a plane defined by the invariant mass and a second variable, the GL, that combines several discriminant variables into one using a multi-variate method (the $\Delta\chi^2$ method). The sensitive region of the plane was blinded until the analysis was completely defined. The main point of the analysis is the use of control channels to calibrate the signal and background pdfs and to normalize the number of observed events to a known $\mathcal{B}$ ratio using several normalization channels. To set a limit on the $\mathcal{B}$ the CL_s method has been used. Discussions are ongoing to use other methods. The limits set by the LHCb collaboration with $37\,\mathrm{pb}^{-1}$ are $\mathcal{B}(B_s \to \mu^+\mu^-) < 5.6 \times 10^{-8}$ at 95 % C.L. and $\mathcal{B}(B^0 \to \mu^+\mu^-) < 1.5 \times 10^{-8}$ at 95 % C.L.

References

[1] R. Aaij *et al.*, [LHCb Collaboration], "Measurement of $\sigma(pp \to b\bar{b}\,X)$ at $\sqrt{s} = 7\,\mathrm{TeV}$ in the forward region", Phys. Lett. B **694** 209 (2010).

[2] A.A. Alves *et al.* [LHCb Collaboration] "The LHCb detector at LHC", JINST **3** (2008) S08005, and references therein.

[3] R. Aaik *et al.* [The LHCb Collaboration] "Search for the rate decays $B_s^0 \to \mu^+\mu^-$ and $B^0 \to \mu^+\mu^-$", arXiv:1103.2465v1 [hep-ex], CERN-PH-EP-2011-029.

[4] A.J. Buras, G. Isidori and P. Paradisi, "EDMs vs CPV in $B_{s,d}$ mixing in two Higgs doublet models with MFV", arXiv:1007.5291; A. J. Buras, "Relations between $\Delta m_{s,d}$ and $B_{s,d} \to \mu^+\mu^-$ in Models with Minimal Flavour Violation", Phys. Lett. B **566**, 115 (2003)

[5] K.S. Baba, C.F. Kolda, "Higgs-mediated $B_q^0 \to \mu^+\mu^-$", Phys. Rev. Lett. **84**, 539 (2010).

[6] V. Abazov *et al.* [D0 Collaboration], "Search for rare decay $B_s^0 \to \mu^+\mu^-$, Phys. Lett. B **693**, 539 (2010). T. Aaltonen *et al.* [CDF Collaboration], "Search for $B_s^0 \to \mu^+\mu^-$ and $B^0 \to \mu^+\mu^-$ decays with $2\,\mathrm{fb}^{-1}$ pf $p\bar{p}$ Collisions", Phys. Rev. Lett. **100**, 101802 (2008).

[7] T. Aaltonen *et al.* [CDF Collaboration], "Search for $B_s^0 \to \mu^+\mu^-$ and $B^0 \to \mu^+\mu^-$ decays with $3.7\,\mathrm{fb}^{-1}$ pf $p\bar{p}$ collisions with CDF II", CDF Public Note 9892.

[8] D. Martinez Santos, J.A. Hernando and F. Teubert, "LHCb Potential to Measure/Exclude the Branching Ratio of the Decay $B_s^0 \to \mu^+\mu^-$", LHCb Note LHCb-2007-033.

[9] D. Karlen, "Using Projections and Correlations to Approximate Probability Distributions", Comp. Phys., **12** 69 (1998).

[10] D. Asner *et al.* [Heavy Flavour Averaging Group], "Averages of b-hadron, c-hadron, and τ-lepton properties", arXiv:1010.1589 [hep-ex] (online update at: http://www.slac.stanfrod.edu/xorg/hfag/osc/).

[11] P. Speckmayer *et al.* "The toolkit for Multivariate Data Analysis TMVA 4", J. Phys. Conf. Ser. **219** 033057 (2010).

[12] A.L. Read, "Presentation of search results: the CL_s technique", J. Phys. G **28**, 2693 (2002).

[13] T. Junk, "Confidence Level computation for combining searches with small statistics", Nucl. Instrum. Meth A **434** 435 (1999).

[14] D. Martinez Santos, "Study of the very rare decay $B_s^0 \to \mu^+\mu^-$ in LHCb", CERN-THESIS-2010-068.

Statistical methods in CMS searches

Amnon Harel, on behalf of the CMS collaboration
University of Rochester

Abstract
A review of the statistical methods used in the first CMS searches for new physics at 7 TeV, from 2010 to January 2011.

1 Introduction

In 2010, the Large Hadron Collider (LHC) started producing pp collisions at a center of mass energy of 7 TeV. By the year's end, a data sample corresponding to over 40 pb^{-1} was recorded in the CMS detector, a multipurpose high-energy particle detector. The CMS collaboration analyzed this data for evidence of physics beyond the standard model (SM). The results of these first searches are consistent with the SM, and limits were placed on the corresponding new physics scenarios. We will review the statistical methods used to set these first limits and to rule out evidence of new physics.

2 The W' search

In Ref. [1] we report a search for the production and decay of a heavy copy of the W boson. Specifically, we search for a W' boson that has W-boson like couplings to fermions and does not couple to other gauge bosons. A previous search at the Fermilab Tevatron Collider ruled out $M_{\text{W'}} < 1.1$ TeV.

We simulate the signal using the PYTHIA V6.422 event generator [2], and scale the production cross section to match next-to-next-to-leading order (NNLO) calculations. We select events with an isolated electron, and a p_T imbalance ($\not\!\!E_T$). The p_T imbalance is reconstructed using the particle-flow technique [3]. The main background processes are W+jets and multijet production. We derive the distributions of key observables for both processes using data-driven techniques. We then fit a linear sum of these distribution to the distribution observed in collision data, with the additional smaller background contributions accounted for according to simulation, as shown in Fig. 1.

Next, we define for each event its visible mass,

$$(M_T)^2 = 2E_T^e \not\!\!E_T \left(1 - \cos\phi_{e,\not\!\!E_T}\right), \tag{1}$$

and look for an excess of high M_T, as shown in Fig. 2. For each W' mass $M_{\text{W'}}$, we defined a priori a search region consisting of M_T values above some minimal value and set limits on the effective cross section of a hypothetical W' boson. No data are observed in any of the search regions.

We use a Bayesian limit setting procedure following Ref. [4] which addresses this canonical scenario: Poisson statistics in each bin of M_T, no interference between the signal and background contributions, and the systematic uncertainties are easily factorized.

The statistical problem is then that for each M_T bin we have:

$$N_{\text{pred}} = b + \mathcal{L}\epsilon\sigma_{\text{eff}}, \tag{2}$$

where $\mathcal{L}$ is the integrated luminosity, ϵ is the selection efficiency for that bin, and σ_{eff} is the effective cross section, i.e. the production cross section (σ) times the branching fraction into the observed channel (B). Then $N_{\text{pred}}(M_T)$ is given for the null (SM) hypothesis ($\sigma_{\text{eff}} = 0$) and for the alternative (SM+signal) hypothesis. We use a constant prior for σ_{eff}, often described as "flat" in HEP papers:

$$f(\sigma_{\text{eff}}) = \begin{cases} \text{const} & \sigma_{\text{eff}} \in [0, \sigma_{\text{max}}] \\ 0 & \text{otherwise} \end{cases}, \tag{3}$$

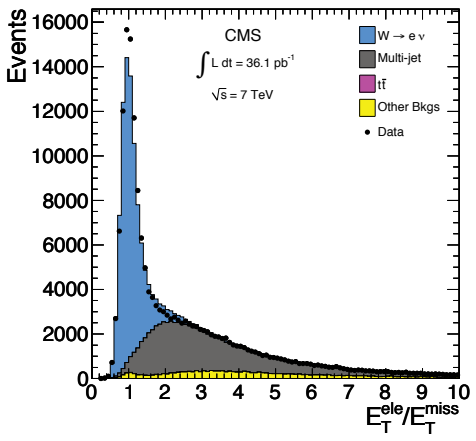


Fig. 1: Sample composition in the W' search and the distributions of the observable used in its fit.

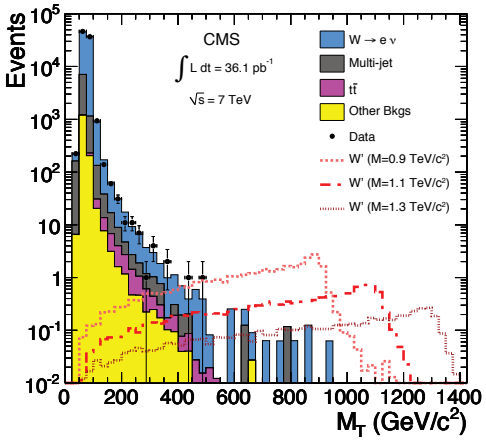


Fig. 2: Data, sample composition, and examples of signal models for the W' search.

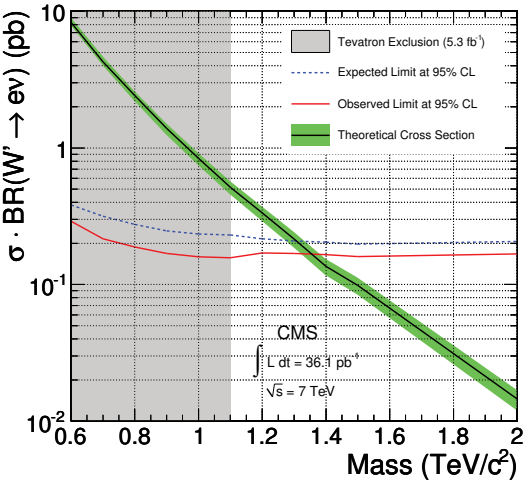


Fig. 3: Limits on W' bosons.

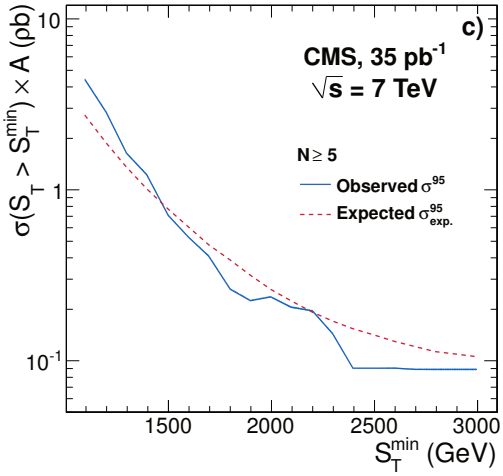


Fig. 4: An example of model-independent limits on microscopic black hole production.

with $\sigma_{\max}$ chosen to be large enough so that the results are not sensitive to its exact value. We use log-normal priors for the nuisance parameters, and integrate them out. In particular, the signal normalization uncertainties from the fit are summarized into one number, which is a typical approximation. The resulting limits are shown in Fig. 3 together with the σ_{eff} predicted for each $M_{\text{W}'}$. From their intersection we rule out at 95% CL the existence of such W' bosons with $M_{\text{W}'} < 1.36$ TeV.

3 Other Bayesian limits

The same statistical treatment was used in the early CMS searches for 1st [5] and 2nd [6] generation leptoquarks and for microscopic black holes [7]. The latter also contains model-independent limits on the effective cross section times acceptance for the different final-state particle multiplicities: ≥ 3, ≥ 4, and ≥ 5 (see example in Fig. 4). Several models are considered with rotating or non-rotating black holes, with or without a stable non-interacting remnant, and with differing values of the Planck scale in the bulk, the number of extra dimensions, and of the minimal black hole mass. In all cases, the model independent limits were only 10% worse than the full model-dependent limits, as in each model there is one particle multiplicity that dominates the limits.

4 Dijet resonance search

CMS published a search for resonant dijet production [8]. Dijet resonances are common in models of new physics beyond the SM. Eight specific models are studied in the paper, which also describes a model-independent study based on three generic signal models: narrow resonances with quark-quark, quark-gluon and gluon-gluon final states.

We consider the two leading (largest p_T) jets as the dijet system. Events are collected using single-jet triggers. Only events where both jets have a pseudo-rapidity $|\eta| < 2.5$ and their unsigned pseudo-rapidity difference is $|\Delta\eta| < 1.3$ are used. For each event we reconstruct the invariant mass of the dijet system, m_{jj}. The observables used in the statistical analysis are the event counts in each of the predefined m_{jj} bins. The width of the m_{jj} bins corresponds to the experimental resolution on m_{jj}. A narrow resonance is one whose width is similar to or smaller than the experimental resolution on m_{jj}.

Non-resonant dijet production is described by a fit to the data of a smooth functional form that does not contain a peak. Three functional forms, used in similar searches in previous colliders, were considered. The best fit to the data ($\chi^2/\text{N.D.O.F.} = 32/31$) was with the form

$$\frac{d\sigma}{dm_{jj}} = p_0 \frac{\left(1 - \frac{m_{jj}}{\sqrt{s}}\right)^{p_1}}{\left(\frac{m_{jj}}{\sqrt{s}}\right)^{p_2 + p_3 \ln\left(\frac{m_{jj}}{\sqrt{s}}\right)}}, \tag{4}$$

where p_i are the fitted parameters, and $\sqrt{s}$ is the collision energy (7 TeV), and is shown in Fig. 5.

To verify the fit's agreement with data and rule out evidence for dijet resonances, we find the biggest excess in the range 0.5 – 2.0TeV, which is for a resonance mass ≈ 0.9 TeV, and quantify its statistical significance. Its local significance, from the log likelihood ratio (LLR), is 1.7σ. We account for the "look elsewhere effect" (LEE) using ensemble tests, and find a similar or locally-more-significant fluctuation in almost half the pseudodatasets (PDSs), so that the overall significance is reduced to 0.02σ.

We set limits on resonance dijet production using an approximate Bayesian procedure. The statistics-only case is treated exactly, using the same method used in the W' search (see Section 2). In this analysis we define $\sigma_{\text{eff}} = \sigma BA$, where A is the acceptance, i.e., the probability that the resonance produces two jets that pass the selection criteria.

The systematic uncertainty is incorporated at each resonance mass by smearing the posterior probability density of σ_{eff} with a Gaussian whose width is set to the systematic uncertainty on the measured σ_{eff}. This is approximate, but here, it is also conservative. In particular, we verified frequentist coverage at 1 TeV for σ_{eff} equal to the limiting value, finding a coverage of $\approx 95\%$ without systematic uncertainties, and $> 98\%$ with systematic uncertainties.

The JES uncertainty is the dominant systematic uncertainty, yielding fractional uncertainties of roughly 20 to 40%, depending on the resonance mass. Other systematic uncertainties, due to the choice of background parametrization, jet energy resolution, and the integrated luminosity, yield fractional uncertainties of $\approx 10\%$ each. The systematic uncertainties increase the cross section limits by 15 to 50%, depending on the resonance mass and its parton content. They decrease the mass limits by $\approx 10\%$.

The limits on resonant dijet production are shown in Fig. 6. We rule out, at the 95% CL, string resonances of mass 0.5–2.5 TeV, excited quarks of mass 0.50–1.58 TeV, axigluons and colorons of mass 0.50–1.17 TeV and 1.47–1.52 TeV, and E_6 diquarks of mass 0.50–0.58 TeV, 0.97–1.08 TeV, and 1.45–1.60 TeV. References to the exact models used are available in the paper [8].

5 The CLs method

The CLs method is to exclude regions of phase space where

$$CL_s = \frac{CL_{s+b}}{CL_b} < 1 - \alpha, \tag{5}$$

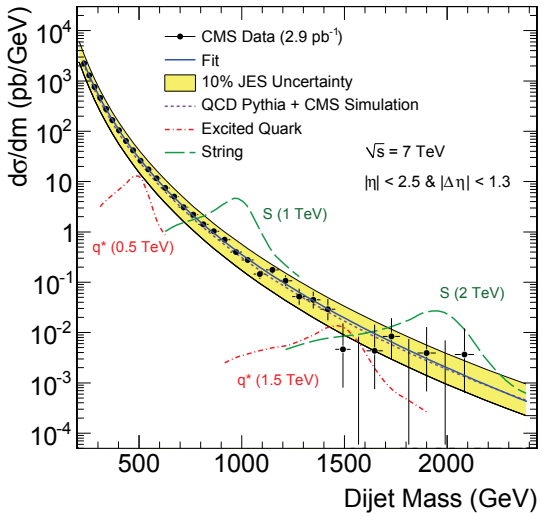


Fig. 5: Data, background fit, and examples of signal models for the dijet resonance search.

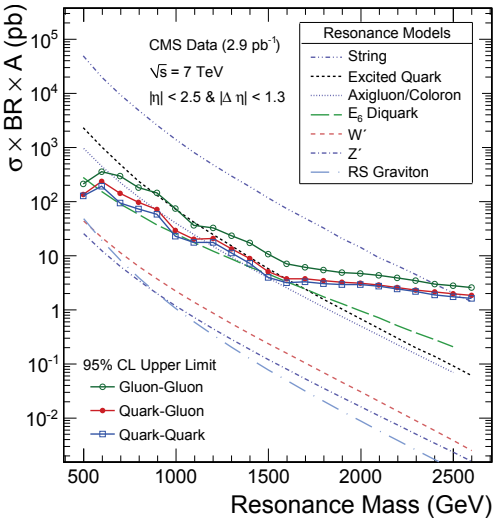


Fig. 6: Limits on dijet resonances.

where α is the desired confidence level, and CL_b and CL_{s+b} are the standard tail probabilities under the null and signal hypotheses (CL_b is the p value). It is recommended [9] to use the LLR as the observable. But often the systematic uncertainties are ignored when calculating the LLR observable, in keeping with the more general prescription [10].

The method's name is very descriptive, but also misleading, as the CLs exclusion region is not a confidence interval. The method is neither purely frequentist nor Bayesian, instead its motivation is practical — it seeks to modify the frequentist CL_{s+b} to avoid false exclusions when the experiment is insensitive to the signal, that it, it is a method of power-constraining frequentist limits. The CLs limit corresponds to the frequentist limits when the experiment is fully sensitive, and the method smoothly degrades the limits as the experiment's power decreases. Despite its shaky foundations in statistical theory, it has been producing sensible results for over a decade.

6 Search for quark compositeness

CMS searched for quark compositeness [11], which is expected to appear at low energies as a contact interaction. Quark contact interactions will enhance low-$|\eta|$ dijet production, in contrast to the SM production, where quantum chromodynamics predicts mostly high-$|\eta|$ jets from t-channel production.

We define the dijet centrality ratio R_η as the number of events where both leading jets are central, with $|\eta| < 0.7$, divided by the number of events where both are less central, with $0.7 < |\eta| < 1.3$. Except for these angular cuts, the event selection follows that of the dijet resonance search above.

The R_η observable is binned in m_{jj} using the same binning as in the dijet resonance search. The background is estimated from next-to-leading-order (NLO) calculations with non-perturbative corrections and with an offset in R_η to match the data in the low m_{jj} region, where no new physics is expected. The fitted offset is $-0.050 \pm 0.021(\text{stat.}) \pm 0.039(\text{syst.})$. Using an ensemble of PDS generated according to the background model, we find the two-sided p value of this offset to be 0.29. The data and background are shown in Fig. 7. At high m_{jj} the data is significantly less signal-like than the SM predictions. But overall, the data and background model are consistent. For example, fitting an offset over the entire m_{jj} range yields $-0.037 \pm 0.007(\text{stat.}) \pm 0.039(\text{syst.})$ with a two-sided p value of 0.34.

In each m_{jj} bin, R_η is distributed as a "Ratio of Poisson means", and we use the standard and extremely useful practice of conditioning this distribution on the total (inner + outer) number of events observed in that bin, simplifying it to a Binomial distribution [12]. We combine data from all m_{jj} bins

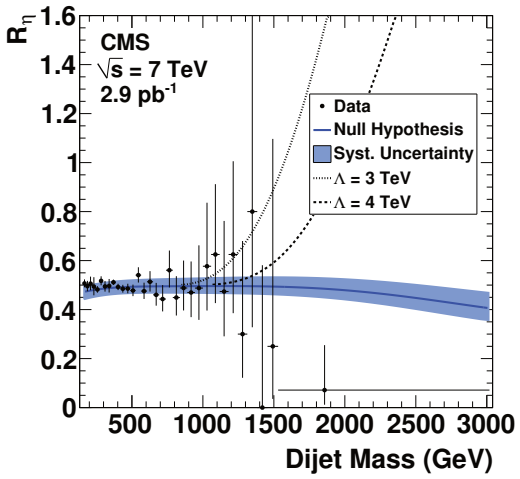


Fig. 7: Data, background model, and examples of signal models for the quark compositeness search.

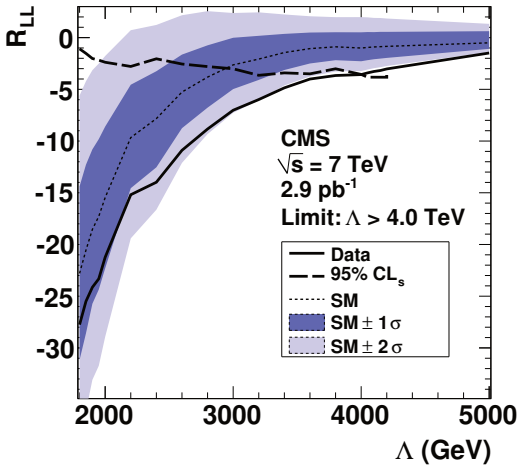


Fig. 8: Test statistics and limits on the scale of quark compositeness.

into one test statistic — the statistics-only LLR for the SM and SM-with-contact-interaction hypotheses. We use the CLs method to set limits on the scale of the contact interactions, Λ. Each Λ value is evaluated separately. The CL_b and CL_{s+b} tails are calculated by ensemble testing, with the nuisance parameters integrated out by varying them for each PDS. The low R_η values at high m_{jj} lead to low CL_b values which require evaluation of the extreme tails of CL_{s+b}, which proved difficult using this integration technique. Large ensembles were needed, some with $> 200\,000$ PDSs. To avoid bias from the choice of ensemble size we formalized stopping rules for the production of additional PDSs: either the Λ value is included/excluded by CLs at the 2σ level, or the statistical error on the CL_s value is $< 0.5\%$.

Fig. 8 illustrates the limit setting procedure. The intersection of the CLs and SM curves indicates the expected limit of $\Lambda > 2.9$ TeV, while the intersection of the CLs and data curves indicates the much higher observed limit of $\Lambda > 4$ TeV.

7 Stopped gluino search

CMS published a search for heavy, quasi-stable particles [13], in particular, for gluinos predicted by split-SUSY which give rise to charged R-hadrons. Such R-hadrons would stop within the CMS detector, and decay at a later time, in contrast to SM decays, whose timing is strongly correlated with LHC collisions.

The search used two observables. The first is the number of events within a time window. The window starts at 50 ns and ends at $1.256\tau_{\text{gluino}}$ (where τ_{gluino} is the gluino lifetime) after each LHC bunch crossing, and excludes 100 ns windows around subsequent bunch crossings. No signal excess was observed and the CLs method was used to derive limits on σ_{eff} for each τ_{gluino} hypothesis (see "Counting" in Fig. 9) and for different stopping scenarios.

The second observable was the time of the selected events, within those same time windows. The signal time-dependence is driven by the timing of the bunch crossing and by τ_{gluino}. The background is mostly from instrumental noises, and is time-independent. We calculate a likelihood as a function of the background amount (per LHC filling scheme) and σ_{eff}, and derive Bayesian limits from the posterior probability using uniform priors in both variables (see "Timing" in Fig. 9).

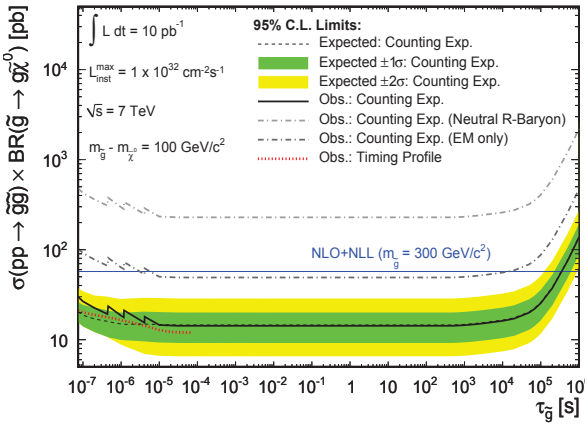


Fig. 9: Limits on stopped gluino production.

8 Summary

The statistical techniques used in the first CMS searches for new physics in pp collisions at $\sqrt{s} = 7$ TeV were reviewed. Consistency with the SM was typically evaluated using p values from ensemble tests. Limits were set using either Bayesian methods or the CLs method.

References

[1] CMS Collaboration, *Search for a heavy gauge boson W' in the final state with an electron and large missing transverse energy in pp collisions at* $\sqrt{s} = 7$ *TeV*, Phys. Lett. B **698** (2011) 21.

[2] T. Sjöstrand, S. Mrenna, and P. Skands, *PYTHIA 6.4 physics and manual*, `http://dx.doi.org/doi:10.1088/1126-6708/2006/05/026` JHEP **2006** (2006) 026.

[3] CMS Collaboration, *Particle-flow event reconstruction in CMS and performance for jets, taus, and missing ET*, CMS PAS PFT-09-001 (2009).

[4] I. Bertram *et al.*, *A recipe for the construction of confidence limits*, technical report, 2000. FERMILAB-TM-2104.

[5] CMS Collaboration, *Search for pair production of first-generation scalar leptoquarks in pp collisions at* $\sqrt{s} = 7$ *TeV*, Phys. Rev. Lett. **106** (2011) 201802.

[6] CMS Collaboration, *Search for pair production of second-generation scalar leptoquarks in pp Collisions at* $\sqrt{s} = 7$ *TeV*, Phys. Rev. Lett. **106** (2011) 201803.

[7] CMS Collaboration, *Search for microscopic black hole signatures at the Large Hadron Collider*, Phys. Lett. B **697** 5 (2011) 434.

[8] CMS Collaboration, *Search for dijet resonances in 7 TeV pp collisions at CMS*, Phys. Rev. Lett. **105** (2010) 211801.

[9] A. L. Read, *Presentation of search results: the CLs technique*, J. Phys. G: Nucl. Part. Phys. **28** (2002) 2693.

[10] A. L. Read, *Modified frequentist analysis of search results (the CLs method)*, in Proceedings of the First Workshop on Confidence Limits, CERN, Geneva, Switzerland (2000) 81.

[11] CMS Collaboration, *Search for quark compositeness with the dijet centrality ratio in pp collisions at* $\sqrt{s} = 7$ *TeV*, Phys. Rev. Lett. **105** (2010) 262001.

[12] J. Przyborowski and H. Wilenski, *Homogeneity of results in testing samples from Poisson series with an application to testing clover seed for dodder*, Biometrika **31** (1940) 313.

[13] CMS Collaboration, *Search for stopped gluinos in pp collisions at* $\sqrt{s} = 7$ *TeV*, Phys. Rev. Lett. **106** (2011) 011801.

Statistical methods used in ATLAS for exclusion and discovery

Diego Casadei for the ATLAS Collaboration
Department of Physics, New York University, New York

Abstract
The statistical methods used by the ATLAS Collaboration for setting upper limits or establishing a discovery are reviewed, as they are fundamental ingredients in the search for new phenomena. The analyses published so far adopted different approaches, choosing a frequentist or a Bayesian or a hybrid frequentist-Bayesian method to perform a search for new physics and set upper limits. In this note, after the introduction of the necessary basic concepts of statistical hypothesis testing, a few recommendations are made about the preferred approaches to be followed in future analyses.

1 Introduction

This note summarizes the statistical methods used so far by the ATLAS Collaboration for setting upper limits or establishing a discovery, and includes the recommended approaches for future analyses, as recently agreed in the context of the ATLAS Statistics Forum. The recommendations aim at achieving a better uniformity across different physics analyses and their ultimate goal is to improve the sensitivity to new phenomena, while keeping robustness as a fundamental request. The best way to be safe against false discoveries is to compare the results obtained using at least two different methods, at least when one is very near the "five sigma" threshold which is required in high-energy physics (HEP) to claim a discovery. One recommended method is explained in this paper (section 4).

We focus here on the searches for some type of "signal" in a sample of events dominated by other ("background") physical sources. The events are the output of a particle detector, filtered by reconstruction algorithms which compute high-level features like an "electron" or a "jet". Large use of simulated samples is required to tune calibrations, characterize the event reconstruction, and compare the outcome of an experiment with the theoretical models.

A typical simulation consists of few different steps. First, one needs to simulate the result of the primary particle interaction with the help of an "event generator". Usually, only one specific process of interest is considered (e.g. Higgs production with a specific channel) and saved to disk, allowing the physicist to study a well defined type of "signal" at the time. Different Monte Carlo (MC) productions are then organized to obtain a variegate set of processes which, depending on the analysis, can be considered either signal or background. The next step is to simulate the effects of the passage of the produced particles (and their decay products) through the detector. This requires the knowledge of the ways energy is deposited in each material and defines the "tracking" of the simulated particles up to the point (if any) in which they stop. Finally, the detector response is simulated: for each energy deposition into an active material, another MC process produces the electronic signal. The latter is processed in a way which closely follows the design of the front-end electronics, obtaining the simulated detector output in the same format as the data coming from the real detector.

Statistical uncertainties arise from fluctuations in the energy deposition in the active materials and from the electronic noise. Systematics due to the limited knowledge of the real detector performance and to the details of the offline reconstruction also contribute to the final uncertainty and need to be addressed case by case. Finally, theoretical uncertainties in the physical models need also to be accounted for. In general, the differences among the event generators cannot be treated as standard deviations, because one usually have just two or three available generators. Hence they should not be summed in quadrature but treated separately.

Section 2 summarizes the statistical aspects relevant to our problems and defines some notation. The methods applied in past ATLAS publications are reviewed in section 3, while section 4 focuses on the methods which can be used in future analyses.

2 Notation

In HEP we deal with hypothesis testing when making inferences about the "true physical model": one has to take a decision (e.g. exclusion, discovery) given the experimental data. In the classical approach proposed by Fisher, one may decide to reject the hypothesis if the *p-value*, which is the probability of observing a result at least as extreme as the test statistic[1] in the assumption that the null hypothesis H_0 is true, is lower than some threshold. In the search for new phenomena, the p-value is interpreted as the probability to observe at least as many events as the outcome of our experiment in the hypothesis of no new physics. Alternatively, one may convert the p-value into the *significance* Z, which is the number of Gaussian standard deviations which correspond to the same right-tail probability[2]: $Z = \Phi^{-1}(1-p)$. The function $\Phi^{-1}(x) = \sqrt{2}\,\mathrm{erf}^{-1}(2x-1)$ is the quantile of the normal distribution, expressed in terms of the inverse error function.

A p-value threshold of 0.05 corresponds to $Z = 1.64$ and is commonly used in HEP for setting upper limits (or one-sided confidence limits) with 95% confidence level or posterior probability. On the other hand, it is customary to require at least a "five sigma" level $Z \geq 5$ (i.e. $p \leq 2.87 \times 10^{-7}$) in order to claim for a discovery of a new phenomenon (if $3 \leq Z \leq 5$ one usually says only that the data suggest the evidence for something new). It is also common to quantify the sensitivity of an experiment by reporting the expected significance under the assumption of different hypotheses.

Another possible approach, suggested by Neyman and Pearson, is to compare two alternative hypotheses (when the null hypothesis is the main focus of the analysis, the alternative H_1 can be defined as the negation of H_0). In this case, two figures of merit are to be taken into account:

- the *size*[3] α of the test, which is the probability of incorrectly rejecting H_0 in favour of H_1 when H_0 is true. α is also the false positive (or "type I error") rate;
- the *power* of the test $(1-\beta)$, which is the probability of correctly rejecting H_0 in favour of H_1 when H_0 is false. β is the probability of failing to reject a false hypothesis, i.e. the false negative (or "type II error") rate.

In the Bayesian approach, one always compares two (or more) different hypotheses. In order to take the decision among the alternatives, one can look at the posterior odds or at the ratio of the marginal likelihoods (or "Bayes' factor"). The former are always well defined and take into account the information accumulated with the performed experiment in the light of the existing prior information, whereas the latter is often very difficult to compute and may be even ill-defined in some problem (for example when comparing two models in which one of the priors is improper), although it does not depend on the prior knowledge about the hypothesis under consideration. The decision is taken in favour of the hypothesis which maximizes the chosen ratio, though the particular value of the latter can can suggest a weak, mild or strong preference for that hypothesis.

In this note, we address two problems, exclusion and discovery, for which the notation is different and sometimes misleading, as illustrated below. For this reason, in the rest of the paper we will speak about the "signal plus background" (sig+bkg or H_{s+b}) hypothesis and about the "background only" (bkg or H_b) hypothesis, without saying what is the null hypothesis.

[1] A test statistic is a function of the sample which is considered as a numerical summary of the data that can be used to perform a hypothesis test.

[2] Here we consider a one-sided test, in which we look for an excess over the expected number of events due to the background processes.

[3] α is also known as "significance level" of the test. We do not use this terminology to avoid confusion with the significance Z defined above.

In the discovery problem, the null hypothesis H_0 describes the background only, while the alternative H_1 describes signal plus background. In the classical approach, one first requires that the p-value of H_0 is found below the given threshold (in HEP one requires $p \leq 2.87 \times 10^{-7}$). If this condition is satisfied, one looks for an alternative hypothesis which can explain well the data[4].

In the exclusion problem, the situation is reversed: H_0 describes signal plus background while the alternative hypothesis H_1 describes the background only. In the classical approach, one just makes use of the null hypothesis to set the upper limit, though this will exclude with probability $\approx \alpha$ parameters values for which one has little sensitivity, obtaining "lucky" results. Historically, this problem has been first addressed in the HEP community by the CL_s method [1,2], whose approach is to reject the sig+bkg hypothesis if $CLs = p_{s+b}/(1 - p_b) \leq \alpha$. CLs is a ratio of p-values which is commonly use in HEP, and one can find a probabilistic interpretation if certain asymptotic conditions are met [3]. Another possibility being discussed by ATLAS physicists is to construct a Power Constrained upper Limit (PCL) by requiring that two conditions hold at the same time: (1) the p-value is lower than the chosen threshold, and (2) the power of the test is larger than a minimum value chosen in advance.

3 Methods used in past ATLAS publications

So far, different ATLAS analyses used different approaches. Converging takes time and is not always possible nor necessarily good, the main reason probably being that different uncertainties are addressed in different ways. Whenever possible, the background is estimated from data. Still, one has to extrapolate to the signal region and this requires the knowledge of the shape, hence depends on the simulation. In addition, in many cases signal and control regions should be treated at the same time: systematics affect both signal and background and often it is impossible to find a signal free region. Finally, in most cases the background is composed by several contributions which are independently simulated but are not really independent: systematic effects act on all of them, making things more and more complicate.

Accounting simultaneously for systematic effects on different components is now possible thanks to HistFactory, a ROOT [4] tool for a coherent treatment of systematics based on RooFit/RooStats [5], initially developed by K. Cranmer and A. Shibata. First used in the top group [6], HistFactory is now being adopted also by other ATLAS groups.

Searches for new physics (for example, Higgs searches) often start by looking for a "bump" in a distribution which is dominated by the background. When the location of the bump is not know, the search is typically repeated in different windows, decreasing the sensitivity [7, 8]. In the ATLAS dijet resonance search [9], a tool for systematic scans with different methods has been applied: BumpHunter, developed G. Choudalakis [10]. The program makes a brute force scan for all possible bump locations and widths, achieving a very good sensitivity, and is appropriate when the bump position and/or width are not known.

A hybrid Bayesian-frequentist approach has been used by the LEP and Tevatron Higgs working groups and is also used in ATLAS Higgs searches. All or some nuisance parameters (modeling systematic effects) are treated in the Bayesian way: a prior is defined for each parameter which is integrated over. On the other hand, for the parameters of interest the frequentist approach is followed, computing p-values and constructing confidence intervals. HistFactory can be used also with this approach, supporting normal, Gamma and log-normal posteriors for nuisance parameters.

In the Higgs combination chapter in the ATLAS "CSC book" [11], the statistical combination of SM Higgs searches in 4 different channels (using MC data) was performed with RooFit/RooStats in the frequentist approach: systematics have been incorporated by profile likelihood. Each search was performed with a fixed mass and repeated for different values, and the limits have been interpolated.

[4]It might happen that more than a single hypothesis can explain the data. In this case there is no conventionally agreed behaviour. A reasonable approach would be to pick up the one with the best agreement with the data, perhaps using a Bayesian approach to assess how strong the preference is.

Many lessons have been learned and the statistical treatment has been refined since then, culminating in the recommended frequentist method explained in section 4.1 below.

4 Present and future analyses

If possible, one may consider using more than a single approach in searches for new phenomena: if they agree, one gains confidence in the result; if they disagree, one must understand why (possibly finding flaws in the analysis). This becomes expecially important when the obtained sensitivity is close to the minimum limit for discovery. Section 4.1 below summarizes the recently proposed frequentist approach which is being recommended for all ATLAS analyses. A possibility is to test the result of the frequentist approach with a Bayesian method. The current dicussion about the Bayesian approach is summarized in section 4.2, but at present there is no official ATLAS recommendation about it.

4.1 Recommended frequentist approach

The problem is formulated by stating that the expected number of events in bin i is the sum $E(n_i) = \mu s_i + b_i$ of two separate contributions, a background expectation of b_i events and a signal contribution given by the product of an intensity parameter μ with the expected number of signal events s_i. For discovery, we test the background-only hypothesis $\mu = 0$. If there is no significant evidence against such hypothesis, we set an upper limit on the magnitude of the intensity parameter.

The Reader will find a full treatment of the recommended method in Ref. [12]. Very shortly, the profile likelihood is used to construct different statistics for testing the alternative bkg and sig+bkg hypotheses. In the asymptotic regime, confidence intervals can be found analytically using such statistics, and the resulting expressions can be used to define approximate intervals for finite samples. Asymptotically, the maximum likelihood estimate $\hat{\mu}$ is Gaussian distributed about the true value with standard deviation σ which can be found numerically by means of the "Asimov dataset", defined as the MC sample which, when used to estimate all parameters, gives their true values. In case of exclusion, the approximate upper limit is $\mu_{\text{up/low}} = \hat{\mu} \pm \sigma_{\text{A}} \Phi^{-1}(1 - \alpha/2)$. In case of discovery, in which one assumes $\mu = 1$, the median significance is

$$\text{med}[Z_0|1] = \sqrt{2[(s+b)\ln(1+s/b) - s]}\,, \tag{1}$$

which is the recommended formula by the ATLAS Statistics Forum when estimating the sensitivity for discovery [12].

4.2 Current discussions about the Bayesian approach

In the Bayesian approach, the full solution to an inference problem about the "true physical model", which is responsible for the outcome of an experiment, is provided by the posterior probability distribution of the parameter of interest. Typicallly, there are several nuisance parameters which model systematic effects or uninteresting degrees of freedom. In order to obtain the marginal posterior probability distribution as a function only of the parameter of interest, one has to integrate over all nuisance parameters. This marginalization procedure contrasts with the frequentist approach based on the profile likelihood, in which the nuisance parameters are fixed at their "best" values.

Prior probabilities need to be specified for all parameters and should model our knowledge about the effects which they refer to. Quite often, one does not want to encode a precise model into the prior or does not assume any relevant prior information. In this case, uniform densities are commonly preferred for computational reasons, but they are often misinterpreted as "non-informative" priors, which is not the case. For example, a uniform density is no more flat, when considered as a function of the logarithm of the given parameter. When attempting to make an "objective" inference, least-informative priors should be used instead. They can be defined, as in the case of the reference priors, as as the ones which maximize the amount of missing information [13]. Reference priors are invariant under reparametrization, are

known (and often identical to Jeffreys' priors) for most common one-dimensional problems in HEP, and can also be used to test the dependence of the result from the choice of the prior [14, 15].

When dealing with discovery or exclusion in the Bayesian approach, one has to make a choice between two alternative hypotheses: background only (H_b) and signal plus background (H_{s+b}). Comparing the posterior probabilities is the best way to account for the whole amount of information provided by the experiment in the light of the previous knowledge. Although values of $O(1000)$ for the posterior odds are interpreted as a strong preference, no widespread agreement exists in the HEP community about a minimum threshold for claiming a discovery.[5] In order to check the impact of the assumptions made before performing the experiment on the final decision, it is also useful to compare the posterior odds against the prior odds (defined as the ratio of prior probabilities for H_b and H_{s+b}, whenever this is well defined).

5 Summary

This note summarizes the statistical approaches used in the past ATLAS analyses and the current ongoing efforts to provide uniformity of statistical treatment across all analyses. Guidelines for estimating the sensitivity with a frequentist method based on profile likelihood ratio have been recently formalized [12]. In this approach, which is recommended for all ATLAS analyses, all nuisance parameters are fixed at their best values and a single MC sample (the Asimov dataset) can be used to find the numerical values of the interesting statistics.

The Bayesian approach can also be considered in the analysis, although no official ATLAS recommendation has been made yet about the best method. In general, the prior densities should be chosen in the way which best models our prior knowledge of the model. Whenever one wants to minimize the impact of the choice of the prior on the result, one should be aware that flat priors are to be considered informative. On the other hand, least-informative priors can be defined for all commong HEP problems and have very appealing properties. In the Bayesian approach, the treatment of systematics is different from the recommended frequentist method, because the whole range of each nuisance parameter is considered in the marginalization. Hence, the comparison between the two approaches may be helpful, expecially near the sensitivity threshold for discovery.

References

[1] T. Junk. *Nucl. Instrum. Methods Phys. Res.*, A 434 (1999) 435.

[2] A.L. Read. *J. Phys.*, G 28 (2002) 2693.

[3] E. Gross, O. Vitells. "Statistics Challenges in High Energy Physics. Search Experiments". *ACAT2010 Conf. Proc.*, 2010.

[4] I. Antcheva et al. *Comp. Phys. Comm.*, Vol. 180, Issue 12 (2009) 2499-2512.

[5] L. Moneta et al. "The RooStats Project". *ACAT2010 Conf. Proc.*, 2010. arXiv: 1009.1003.

[6] The ATLAS Collaboration. "Measurement of the top quark-pair production cross section with ATLAS in pp collisions at $\sqrt{s}$ = 7 TeV". Accepted by *Eur. Phys. J.* C, 2011. arXiv: 1012.1792.

[7] O. Vitells. "Look Elsewhere Effect". These proceedings. CERN, 2011.

[8] G. Ranucci. "An alternative view of the Look Elsewhere Effect". These proceedings. CERN, 2011.

[9] The ATLAS Collaboration. *Phys. Rev. Lett.*, 105 (2010) 161801. arXiv: 1008.2461.

[10] G. Choudalakis. "On hypothesis testing, trials factor, hypertests and the BumpHunter". arXiv: 1101.0390, 2011.

[11] The ATLAS Collaboration. "Expected Performance of the ATLAS Experiment. Detector, Trigger and Physics". Technical Report CERN-OPEN-2008-020, ISBN 978-92-9083-321-5, CERN, 2008.

[5] A possible approach could be to simulate many pseudo-experiments, compute the p-value and follow the "five sigma" rule mentioned above.

[12] G. Cowan et al. *Eur. Phys. J.*, C 71 (2011) 1554. arXiv: 1007.1727.
G. Cowan, "Likelihood ratios for search experiments", these proceedings.

[13] J.O. Berger, J.M. Bernardo and D. Sun, Annals of Statistics 37 (2009) 905.

[14] J.M. Bernardo. *Bayesian Statistics 9*, chapter "Integrated objective Bayesian estimation and hypothesis testing", pages 1–68. Oxford University Press, Oxford, 2011.

[15] J.M. Bernardo. *Philosophy of Statistics* (P. Bandyopadhyay and M. Forster, eds.), chapter "Modern Bayesian Inference: Foundations and Objective Methods". Elsevier, Amsterdam, 2009.

Combined searches for the Higgs boson with ATLAS and CMS

Kyle Cranmer, on behalf of the LHC Higgs Combination Group
New York University

Abstract
This document outlines the conceptual challenges involved in forming the combined statistical model of several ATLAS and CMS searches for the Higgs boson, as well as the technology developed to form such a complicated model and the statistical tests the group is considering for the initial result.

1 Introduction

Perhaps the most pressing open question in fundamental particle physics is why the W and Z bosons, the particles associated with the weak interaction, are massive. The fact that these bosons have mass is associated to a phenomena called electroweak symmetry breaking; however, the mechanism for the electroweak symmetry breaking has not been established. Within the standard model (SM) of particle physics, a specific manifestation of the so-called "Higgs mechanism" gives mass to the W and Z bosons and also predicts a new particle, H, called the Higgs boson, which has not been observed. The search for the Higgs boson is one of the primary goals of the Large Hadron Collider (LHC).

As will be detailed below, the Higgs boson can be produced and decay in many different ways. Dedicated searches are designed for the most promising of these possibilities, or *channels*, and each can be related to the same underlying physics theory. Thus, it is common for a collaboration to bring together the different searches and form a more powerful combined analysis. These combined analyses are not based on the results of the individual searches, as there is no satisfactory way to combine p-values; instead a joint statistical model is formed and tested. Forming this joint statistical model requires coordination as many systematic effects are common to the individual searches. Continuing on this logical path, different experiments at the same accelerator complex have formed combined analyses for the Higgs boson. This was first done by the four LEP experiments at CERN [1] and has also been done at the two Tevatron experiments [2]. The ATLAS and CMS collaborations have now formed an LHC Higgs Combination Group (LHC-HCG) with the goal of providing a first combined result in the summer of 2011.

This document outlines the conceptual issues involved in forming the combined statistical model across experiments, the technology developed to form such a complicated model, and the statistical tests the group is considering for the initial result. The focus will be on lessons learned from a toy combination completed in the summer of 2010 [3] and open questions the LHC-HCG is currently deliberating.

2 The probability model of a typical search channel

The SM is an impressively predictive theory formulated in a more general formalism called quantum field theory (QFT). Within QFT the fundamental mathematical object is called the Lagrangian, which encodes the particle content and the interactions in the theory with some free parameters that must be determined experimentally. All of the parameters of the SM Lagrangian have been measured, except for one: M_H, the mass of the Higgs boson itself. Once one specifies M_H – either by hypothesizing a particular value or measuring it – the SM is a fully specified theory that makes numerous predictions. In particular, the SM predicts the production rates via each of the possible Higgs production modes, branching ratios for different decay modes, and distributions of kinematic properties of the final-state particles.

The LHC is a proton-proton collider, currently with a center-of-mass energy of 7 TeV, giving it the capability to produce Higgs bosons. There are a few different types of interactions $pp \to H + X$ that can produce a Higgs boson (perhaps in association with other particles, generically referred to as X) in an individual collision. Each of these production modes, indexed by I, has an associated production

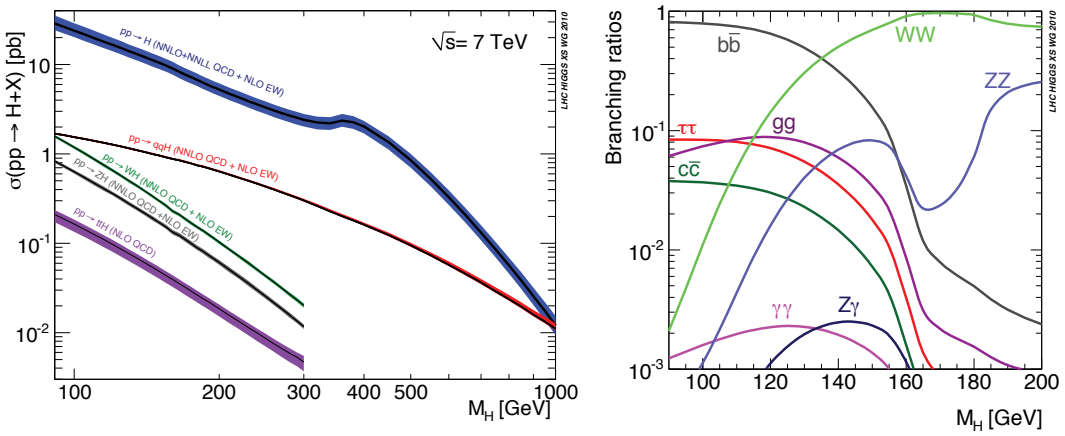


Fig. 1: Production cross-sections and branching ratios of the standard model Higgs boson taken from Ref. [4]

cross-section σ_I (in pb = picobarns = 10^{-36}cm^{-2} in the SM. Figure 1 shows the predicted cross-section in the SM for various production modes as a function of the unknown Higgs mass parameter M_H [4]. The number of collisions predicted to undergo $pp \to H + X$ is given by the product of this production cross-section and the time-integrated luminosity of the proton beams, L, (measured in 1/cm^2) which is controlled by the collider itself.

The Higgs boson is also predicted to be unstable and decay very quickly, well before coming in contact with any component of our particle detectors. The Higgs may decay in several different ways, and the relative proportions for these decays, indexed by f, are called branching ratios, denoted B_f. Figure 1 shows the dependence of the branching ratios for the various decay modes on the unknown Higgs mass parameter M_H [4]. In some cases, such as the decay $H \to ZZ$, there is a further cascade of decays before reaching the final-state particles that interact with the detector. For example, Fig. 2 shows a real collision observed by the CMS detector, that is compatible with the hypothesis of a $H \to ZZ \to \mu^+\mu^-\mu^+\mu^-$ decay. In addition to the overall rate of events like this, the SM also predicts the joint distribution of particles in *phase space*: the space that describes the energies and directions of the final state particles. The interaction of these final state particles with the detector is modeled using detailed computer simulations and reconstruction algorithms are developed to estimate the angles and energies of the particles based on the signatures left in the detector components.

Other types of interactions that may lead to the same final state particles or mimic them in the detector are referred to as backgrounds. Because the production of Higgs bosons is quite rare in comparison to backgrounds, event selection criteria are developed to reject the bulk of these background events, thus defining a *signal region* that is relatively rich in the Higgs signal events. The fraction of signal events from production mode I and decay mode f satisfying the criteria for search channel c is called the *efficiency* of the event selection, ϵ^c_{If}. Thus, the total number of Higgs boson events expected to satisfy the event selection criteria for a given channel is given by

$$s_c = \sum_{I \in \text{production}} \sum_{f \in \text{decay}} \epsilon^c_{If} L \sigma_{I,\text{sm}} B_{f,\text{sm}}. \quad (1)$$

Note that s_c is implicitly a function of the Higgs mass parameter M_H as the production cross-section, branching ratio, and efficiency all depend on M_H.

While the cross-section and branching ratio are predicted to have a specific value in the SM once M_H is specified, it is common to generalize the situation by considering events of a similar efficiency but with a modified rate. This is accomplished by introducing a new parameter $\mu = \sigma_I B_f / \sigma_{I,\text{sm}} B_{f,\text{sm}}$,

where $\mu = 1$ corresponds to the presence of a SM Higgs boson, $\mu = 0$ corresponds to the background-only hypothesis, and other values would indicate a non-standard Higgs boson or possibly some other form of new physics beyond the SM. It is worth noting that in these non-SM situations there is no particular reason to assume a common μ for each production mode I and decay mode f. This common structure across production and decay modes can also be broken by theoretical uncertainties, which will be addressed below.

Of course, we also need an estimate of the background in this selection region. Some backgrounds are estimated using the same first-principles procedure based on theoretical predictions and detector simulation as is used for the signal. However, for some backgrounds this simulation-based approach is highly sensitive to the details of the interaction in the detector or rely heavily on aspects of the theory that require approximate solutions. In these cases, experimentalists prefer to use *data-driven* techniques, where one utilizes known or assumed relationships between the observations in *control regions* and the signal region. Let us refer to the estimates on the number of background events using the simulation-based and data-driven approaches as $b_{\text{sim.}}$ and $b_{\text{d.d.}}$, respectively, and assume that the contributions from those background processes are disjoint, exhaustive, and sum to a total background rate b.

At this point we can write the probability model, neglecting uncertainties, for observing N_c events in the event-selection for the c^{th}-channel based on the background estimates, the total signal expectation for a given M_H, and the parameter of interest μ as $\text{Pois}(N_c|\mu s_c + b)$.

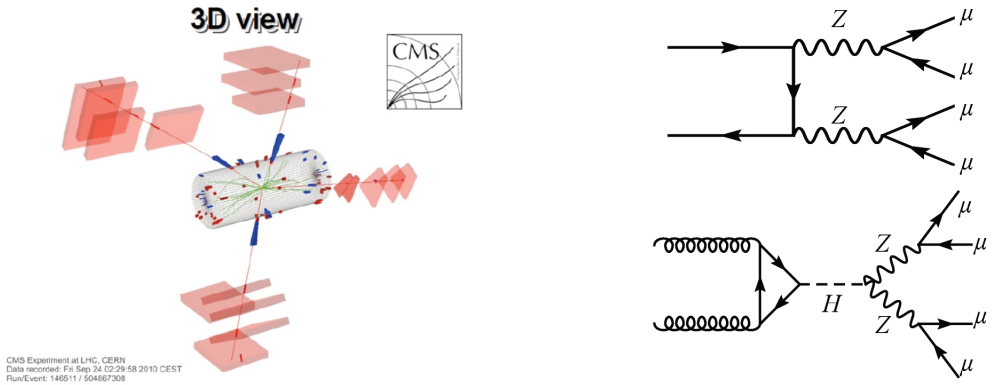


Fig. 2: An event display of a $H \to ZZ \to \mu^+\mu^-\mu^+\mu^-$ candidate event along with Feynman diagrams of compatible background and signal processes

The distribution of final state particles in phase-space can be quite complex and often have pronounced structures. For instance, a certain combination of the angles and energies of the four muons in the $H \to ZZ \to \mu^+\mu^-\mu^+\mu^-$ decay called the invariant mass is a direct estimator of M_H. In some cases, multivariate algorithms such as neural networks and boosted decision trees are used to form a discriminating variable, which may also include information on particle identification. Let us generically denote these discriminating variables, or *marks*, as x and the probability density function describing the distribution of x for a signal and background processes as $f_s(x)$ and $f_b(x)$, respectively. We can extend the simple Poisson model to include this shape information by building what statisticians refer to as a *marked Poisson* process with $\boldsymbol{x}_c = \{x_1, ..., x_{N_c}\}$:

$$P(\boldsymbol{x}_c|\mu) = \text{Pois}(N_c|\mu s_c + b) \prod_j^{N_c} \frac{\mu s_c f_s(x_j) + b f_b(x_j)}{\mu s_c + b} \,. \tag{2}$$

This is the same type of expression physicists use when performing an unbinned extended maximum likelihood fit, where the rate and shape information are both related to the parameter of interest μ.

3 Incorporating uncertainties into the model

Arguably, the most involved and delicate aspect to Higgs searches is controlling and understanding the many sources of uncertainty that modify the expected rate and distributions of signal and background processes. As a result, the model in Eq. 2 is extended to a family of models parametrized with several nuisance parameters $\boldsymbol{\nu}$: $P(\boldsymbol{x}_c|\mu) \rightarrow P(\boldsymbol{x}_c|\mu, \boldsymbol{\nu})$. Modeling of these effects must be coordinated among the different searches as a single source of uncertainty may be common to many searches and the correlations must be taken into account. When combining searches from different experiments, the most acute correlated effects arise from the use of common theoretical tools and the luminosity of the beam.

There are systematic uncertainties associated with detector simulation and performance; statistical uncertainties associated with the auxiliary measurements used in data-driven background estimates; the residual statistical and systematic uncertainty of theoretical predictions associated with the measurement of the strong coupling constant and the parton density functions; and theoretical uncertainties that are not statistical in nature, but result from neglecting higher-order terms in the perturbative expansion of the theory. These uncertainties are quite different in nature, which is reflected in their statistical modeling.

It is helpful to think of the modeling in two steps. The first step is to parametrize the effect of the uncertainty on the primary measurement $\boldsymbol{x}_c$: $P(\boldsymbol{x}_c|\mu) \rightarrow P(\boldsymbol{x}_c|\mu, \boldsymbol{\nu})$. The second is to incorporate additional *constraint terms* $P(\boldsymbol{y}_i|\boldsymbol{\nu})$ that describe how auxiliary measurements $\boldsymbol{y}_i = \{y_1, ..., y_{M_i}\}$ depend on the nuisance parameters. Thus, the probability model expands $P(\boldsymbol{x}_c|\mu, \boldsymbol{\nu}) \rightarrow P(\boldsymbol{x}_c, \boldsymbol{y}|\mu, \boldsymbol{\nu})$.

Parametrization of the effect of uncertainty on the primary measurement of the marks, $P(\boldsymbol{x}_c|\mu, \boldsymbol{\nu})$, is typically dealt with in one of two ways. In the first approach, the model is *explicitly parametrized* in terms of elemental sources of uncertainty, such as electron identification efficiency, jet energy scale uncertainty, parton density functions, etc. A single nuisance parameter ν_i parametrizes the effect of changing a single source of uncertainty, which can be estimated by running the simulation with modified settings or by correcting the simulation in some way. This approach typically requires some form of interpolation as the variation in the selection efficiency $\epsilon(\nu_i)$ and distributions $f(m; \nu_i)$ can only be estimated for discrete values of ν_i. The advantage of this approach is that it is straightforward to understand the correlated effect on the individual searches, by simply identifying the ν_i that correspond to the same source of uncertainty. This approach was used by the CMS inputs to the toy combination [5] and recent ATLAS analyses [6,7]. The second common approach is to use some parametric function that is believed to be flexible enough to capture potential variations due to the underlying sources of uncertainty as an effective model for $P(\boldsymbol{x}_c|\mu, \boldsymbol{\nu})$. These effective models may be well-motivated by understanding of the physical processes, such as an exponentially falling distribution [8], or ad hoc choices, such as polynomials. In this approach, the effect of individual sources of uncertainty are *implicitly parametrized* by the effective model. The advantage of this approach is that the interpolation in $\boldsymbol{\nu}$ is built into the effective model; however, the disadvantage of this approach is that it is difficult to introduce the correlated effect of a specific source of uncertainty across individual searches – an issue the LHC-HCG currently faces.

Ideally, the constraint terms describe other auxiliary measurements $\boldsymbol{y}_i$ in such a way that the uncertainty in $\boldsymbol{\nu}$ can be dealt with in a clear statistical sense. A simple example is the well-studied "on/off" problem [9] in which the unknown background rate in the signal region is related to an auxiliary counting experiment via a known constant τ: $P(\boldsymbol{x}_c, \boldsymbol{y}_i|\mu, b) = \text{Pois}(N_c|\mu s_0 + b) \cdot \text{Pois}(M_i|\tau b)$, where N_c and M_i are the number of events in the main and auxiliary counting experiment and the distributions of x and y are not taken into account. An extension of the "on/off" problem in the marked Poisson model comes from using a control region in the data that is devoid of signal and has a similar distribution for the discriminating variable: $f_b(x; \boldsymbol{\nu}) = f_b(y; \boldsymbol{\nu})$. In most realistic situations, the extrapolation coefficient τ is also uncertain and the distributions in x and y are not identical. In particular, τ is often estimated from simulations and is subject to both experimental and theoretical uncertainties.

It is common that the uncertainty on a nuisance parameter can be estimated from an auxiliary measurement or from experience, but an explicit probability model relating $\boldsymbol{y}_i$ and ν_i is not available

for practical reasons. In these cases, it is common to idealize the situation and choose an ad hoc constraint term that summarizes the auxiliary measurement or captures intuition about the uncertainty in the nuisance parameter. For example, Gaussian constraint terms are very common idealizations of auxiliary measurements. Here Bayesian reasoning is deceptively natural as one often refers to the prior $\pi(\nu_i)$ in informal conversation without recognizing a Bayesian probability inversion. Similarly, one often refers to a "Gamma" prior on a nuisance parameter, which can be interpreted as the posterior resulting from an idealized auxiliary counting experiment with a uniform prior via Bayes theorem: $\pi(\nu_i) \propto \text{Pois}(M|\tau\nu_i) \cdot \text{Uniform}(\nu_i)$. In order use a consistent probability model in both frequentist and Bayesian statistical formalisms, it is important to incorporate the Poisson term into the probability model and separate the original uniform prior $\eta(\nu)$ for Bayesian techniques. Another popular form for an ad hoc constraint term is the log-normal distribution, particularly for non-negative nuisance parameters with large relative uncertainty (>20%). In this case, one must be more careful about what is assumed to be log-normally distributed. If one assumes the observable y in the auxiliary measurement is log-normally distributed (as is implied when invoking multiplicative measurement errors) and uses a uniform prior on ν_i (as in the more familiar Gaussian and Gamma case), then the posterior $\pi(\nu_i)$ does not have a log-normal form. On the other hand, if one means that the posterior is log-normally distributed, then the likelihood function and prior must be specified to provide a consistent frequentist treatment of the problem. While a $\eta(\nu_i) \propto 1/\nu_i$ prior allows both the PDF and the posterior to have a log-normal form, the likelihood function and the posterior are no longer proportional (as they were in the Gaussian and Gamma case). The lesson here is that one cannot simply appeal to idealized measurements and hope for an unambiguous interpretation when there are large uncertainties involved.

3.1 Forming the Combined Model and Technical Implementation with RooFit/RooStats

Once one has prepared the model for each individual search channel together with the associated auxiliary measurements and constraint terms $P(\boldsymbol{x}_c, \boldsymbol{y}|\mu, \boldsymbol{\nu})$, the combined model can be formed by multiplying the individual terms, identifying the common parameters, and perhaps introducing additional terms that would impose non-trivial correlations or functional relationships among the parameters. This requires that the selection regions are disjoint and that the common parameters not only parametrize the same effect, but also have the same conventions. We refer to the combined model by $P(\boldsymbol{x}, \boldsymbol{y}|\mu, \boldsymbol{\nu})$, dropping the c and i subscripts for the individual terms: $\boldsymbol{x} = \{\boldsymbol{x}_c\}$ and $\boldsymbol{y} = \{\boldsymbol{y}_i\}$.

$$P(\boldsymbol{x}, \boldsymbol{y}|\mu, \boldsymbol{\nu}) = \prod_{c \in \text{channel}} P(\boldsymbol{x}_c|\mu, \boldsymbol{\nu}) \prod_{i \in \text{aux.meas.}} P(\boldsymbol{y}_i|\boldsymbol{\nu}) \tag{3}$$

Since the PhyStat conference in 2007, there has been a dedicated effort to develop technologies capable of creating and communicating, and testing complex probability models within the context of the related ROOT projects RooFit and RooStats [10–12]. The `RooWorkspace` class utilizes the ROOT I/O technology to save these complicated models into a persistent file, which can be shared easily. Much of the effort has been to cleanly separate and organize the information needed for the various statistical tests: in particular the `ModelConfig` class keeps track of the PDF $P(\boldsymbol{x}, \boldsymbol{y}|\mu, \boldsymbol{\nu})$, the priors $\eta(\mu)$ and $\eta(\boldsymbol{\nu})$, the parameter of interest μ, the nuisance parameters $\boldsymbol{\nu}$, the observables $\boldsymbol{x}$, and the auxiliary observables $\boldsymbol{y}$.

In the summer of 2010 the ATLAS and CMS collaborations embarked on a toy combination exercise with unofficial, though realistic mock-ups of the searches for $H \to W^+W^- \to l^+l^-\nu\bar{\nu} + 0$ jets with $L = 1$ fb^{-1}, where $l = e, \mu$ [3]. The ATLAS model was based on counting events in the signal region and three control-regions for each of the three $ee, e\mu$, and $\mu\mu$ final states [13]. This setup is very similar to the "on/off" problem where the background rate is a nuisance parameter and auxiliary counting measurements are made explicit; however, here there were three τ coefficients for the different control regions as well as coefficients due to cross-contamination of the different background processes in the individual control regions. The extrapolation coefficients also had large uncertainty that was represented by a Gaussian constraint term (truncated for $\tau < 0$). Unfortunately, the variance of this Gaussian was

given by the sum in quadrature of the variation in τ due to the individual sources of uncertainty, making it impossible to identify the effect of theoretical of uncertainties common to both ATLAS and CMS. In contrast, the CMS model did not incorporate actual auxiliary measurements into the model, but used log-normal constraint terms for 37 nuisance parameters that explicitly parametrized the effect of 37 sources of uncertainty [5]. A visualization of the combined model in terms of a directed acyclic graph is shown in Fig. 3. The models that the LHC-HCG group are considering now are drastically more complex.

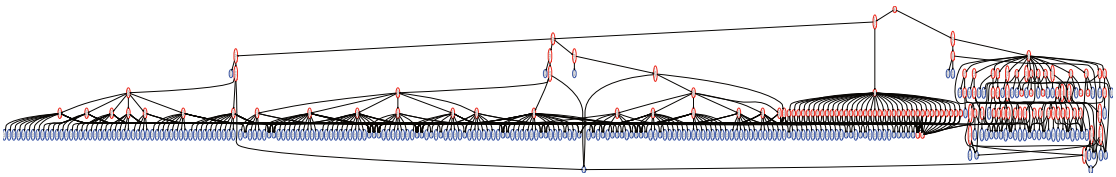


Fig. 3: Visualization of the combined model for the toy ATLAS and CMS Higgs combination performed in the summer of 2010. The top node represents the the likelihood, the left portion of the graph represents the CMS model, the right portion represents the ATLAS model, and the lowest node in the middle of the graph represents μ.

3.2 Statistical Tests

The emphasis on cleanly separating the objective PDF from the Bayesian prior is largely motivated by the desire to retain flexibility in the type of statistical tests that can be used. In particular, the strategy has been to put effort into a single probability model and then consider different statistical procedures. The RooStats framework has implementations of most of the commonly used statistical procedures, including Bayesian methods based on Markov-Chain Monte Carlo, fully frequentist methods based on the Neyman-Construction and hybrid-resampling (also referred to as the 'profile construction') [10, 14–16], likelihood-based methods that utilize the asymptotic results of Wilks and Wald together with numerical procedures for estimating the non-centrality parameter [17–19], as well as mixed Bayesian-Frequentist procedures [20, 21] and CL_s [22, 23].

In the Bayesian realm, the hope is that the auxiliary measurements or idealized constraint terms represented by $P(\boldsymbol{y}|\boldsymbol{\nu})$ are sufficiently informative to dominate the priors on the nuisance parameter $\eta(\boldsymbol{\nu})$, though this has not been studied in much detail. Instead, physicists are more keenly aware of the sensitivity to the prior on the parameter of interest $\eta(\mu)$. While uniform priors on μ reign supreme, there is interest in the use of Jeffreys prior and reference priors [24–26]. Recently, progress has been made in estimating these priors directly [27] and the efficient numerical techniques for calculating the Fisher information matrix, which is a necessary ingredient [19].

On the frequentist side, the emphasis has been on the choice of the test statistic and the details of the ensemble used to compute p-values. At LEP, systematic uncertainties were small and the test statistic was the simple likelihood ratio $Q_{LEP}(\boldsymbol{x}) = P(\boldsymbol{x}|\mu=1)/P(\boldsymbol{x}|\mu=0)$. At the Tevatron uncertainties are larger and a profiled generalization of the LEP test statistic has been used

$$Q_{\mathrm{Tev}}(\boldsymbol{x},\boldsymbol{y}) = \frac{P(\boldsymbol{x},\boldsymbol{y}\,|\,\mu=0,\hat{\hat{\boldsymbol{\nu}}}(\mu=0;\boldsymbol{x},\boldsymbol{y}))}{P(\boldsymbol{x},\boldsymbol{y}\,|\,\mu=1,\hat{\hat{\boldsymbol{\nu}}}(\mu=1;\boldsymbol{x},\boldsymbol{y}))} = \frac{\lambda(\mu=0;\boldsymbol{x},\boldsymbol{y})}{\lambda(\mu=1;\boldsymbol{x},\boldsymbol{y})}\,, \tag{4}$$

where $\lambda(\mu;\boldsymbol{x},\boldsymbol{y})$ is the profile-likelihood ratio

$$\lambda(\mu;\boldsymbol{x},\boldsymbol{y}) = \frac{P(\boldsymbol{x},\boldsymbol{y}\,|\,\mu,\hat{\hat{\boldsymbol{\nu}}}(\mu;\boldsymbol{x},\boldsymbol{y}))}{P(\boldsymbol{x},\boldsymbol{y}\,|\,\hat{\mu}(\boldsymbol{x},\boldsymbol{y}),\hat{\boldsymbol{\nu}}(\boldsymbol{x},\boldsymbol{y}))}\,. \tag{5}$$

Our field has known for many years that the distribution $f(-2\log\lambda(\mu)|\mu,\boldsymbol{\nu})$ is asymptotically related to a chi-square distribution and independent of $\boldsymbol{\nu}$. However, we have only recently appreciated that

$f(-2\log\lambda(\mu)|\mu',\boldsymbol{\nu})$ is asymptotically a non-central chi-square distribution with a non-centrality parameter that depends on μ' and $\boldsymbol{\nu}$. As a result, $f(Q_{\rm TeV}|\mu=1$ or $\mu=0,\boldsymbol{\nu})$ necessarily depend on the true values of the nuisance parameters [19]. For this reason, there is growing support in the LHC-HCG to move to $\lambda(\mu)$ for the test statistic, perhaps restricted to a one-sided alternative. There has also been some progress in understanding how the the look-elsewhere effect modifies the distribution of the test statistic as M_H is a free parameter that cannot be identified in the $\mu=0$ hypothesis [28].

Another area of development that can be compared to the Tevatron procedure is the precise way in which the ensembles are generated. The Tevatron Higgs combination group primarily uses a mixed Bayesian-Frequentist procedure in which the nuisance parameters are marginalized with respect to $\pi(\boldsymbol{\nu})$ in the process of generating pseduo-experiments [20, 21]. The fully frequentist procedure emerging at the LHC is based on the hybrid-resampling procedure [10, 14–16], in which the distribution $f(-2\log\lambda(\mu)|\mu',\hat{\hat{\boldsymbol{\nu}}}(\mu';\boldsymbol{x},\boldsymbol{y}))$ is constructed at a particular value of the nuisance parameter expected to be most relevant given the data. This approach has been used by ATLAS in its first Higgs results [6,7,29] and Fig. 4 is a demonstration in the context of discovery with $\mathcal{O}(10^7)$ pseudo-experiments and a model of realistic complexity. As a result of going to this fully frequentist approach, the ensemble includes both variations in $\boldsymbol{x}$ as well as $\boldsymbol{y}$. The presence of variations in $\boldsymbol{y}$ breaks the discreteness in the test statistic and modifies familiar rules of thumb when the observed count N is zero and expected background rate b is small. This is relevant in cases such as $H\to ZZ\to 4l$ [7] and gives results similar to the Lancaster's mid-P [30].

A new element to the discussion is the potential for conditioning. In the simplest situation represented by the "on/off" problem $\text{Pois}(N|\mu s+b)$ $\text{Pois}(M|\tau b)$, the hypothesis test of $\mu=0$ can be reformulated in terms of the ratio of Poisson means $\beta=(\mu s+b)/\tau b$ where it is clear that the total count $N+M$ has no information on the ratio [31–34]. However, in the context of confidence intervals on μ this conditioning is not appropriate as the total count does carry information on the magnitude of μ. Thus, it is not yet clear to the LHC-HCG if there is an appropriate conditioning procedure in this context.

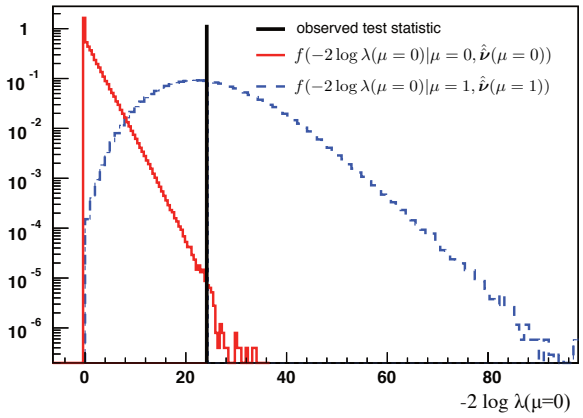


Fig. 4: An example distribution of $-2\log\lambda(\mu=0)$ evaluated $\sim 10^7$ pseudo-experiments for background-only and signal plus background hypotheses. Evaluating $-2\log\lambda(\mu=0)$ requires two fits to the full model, which typically has $\mathcal{O}(50)$ nuisance parameters. This requires batch or PROOF enabled computing clusters.

4 Conclusions

The LHC is performing exceptionally well breaking records in both energy and intensity at a hadron collider. Sensitivity studies suggest that within the next year or two ATLAS and CMS will be in the position to make very strong statements about the existence or non-existence of a SM Higgs boson. The LHC-HCG is aiming to show the first combined results from ATLAS and CMS Higgs searches in the

summer of 2011 – roughly two months from the writing of this document.

Acknowledgements

The author would like to thank his colleagues in the LHC-HCG and the statistics forums and Higgs groups of ATLAS and CMS. This work was supported by NSF grants PHY-0854724 and PHY-0955626.

References

[1] R. Barate et al. *Phys.Lett.*, B565:61–75, 2003.

[2] The CDF and D0 Collaborations. 2011. arxiv:1103.3233.

[3] The ATLAS and CMS Collaborations. Jul. 2010. `http://indico.cern.ch/conferenceDisplay.py?confId=100458`.

[4] LHC Higgs Cross Section Working Group, S. Dittmaier, C. Mariotti, G. Passarino, and R. Tanaka (Eds.). *CERN-2011-002*, 2011. [arxiv:1101.0593].

[5] W. Quayle. Jul. 2010. `http://indico.cern.ch/conferenceDisplay.py?confId=100458`.

[6] The ATLAS Collaboration. Mar 2011. ATLAS-CONF-2011-026.

[7] The ATLAS Collaboration. Mar 2011. ATLAS-CONF-2011-048.

[8] (ATLAS-CONF-2011-004), Feb 2011.

[9] R. D. Cousins, J. T. Linnemann, and J. Tucker. *Nucl.Instrum.Meth.*, A595:480–501, 2008.

[10] K. Cranmer. *PHYSTAT-LHC Workshop on Statistical Issues for LHC Physics*, 2008. oai:cds.cern.ch:1021125. `http://cdsweb.cern.ch/record/1099969`.

[11] W. Verkerke. *PHYSTAT-LHC Workshop on Statistical Issues for LHC Physics*, 2008. oai:cds.cern.ch:1021125. `http://cdsweb.cern.ch/record/1099988`.

[12] L. Moneta, K. Belasco, K. Cranmer, A. Lazzaro, D. Piparo, et al. *PoS*, ACAT2010:057, 2010. [arxiv:1009.1003].

[13] The ATLAS Collaboration. Jul 2010. ATL-PHYS-PUB-2010-009.

[14] G. Feldman. 2000. Talk at the FermiLab Workshop on Confidence Limits.

[15] C. Chuang and T. L. Lai. *Statist. Sinica*, 10:1–50, 2000. `http://www3.stat.sinica.edu.tw/statistica/oldpdf/A10n11.pdf`.

[16] M. Walker B. Sen and M. Woodroofe. *Statist. Sinica*, 19:301–314., 2009. `http://www3.stat.sinica.edu.tw/statistica/oldpdf/A19n116.pdf`.

[17] S.S. Wilks. *Ann. Math. Statist.*, 9:60–2, 1938.

[18] A. Wald. *Transactions of the American Mathematical Society*, 54, No. 3, 1943.

[19] G. Cowan, K. Cranmer, E. Gross, and O. Vitells. *Eur.Phys.J.*, C71:1554, 2011.

[20] R. D. Cousins and V. L. Highland. *Nucl.Instrum.Meth.*, A320:331–335, 1992. Revised version.

[21] T. Junk. *Nucl.Instrum.Meth.*, A434:435–443, 1999.

[22] A. L. Read. *J. Phys. G: Nucl. Part. Phys.*, 28, 2002.

[23] A. L. Read. *in Proceedings of the First Workshop on Confidence Limits, CERN*, 2000.

[24] J. M. Bernardo. *J. R. Statist. Soc. B*, 41:113, 1979.

[25] J. O. Berger and J. M. Bernardo. *J. Amer. Statist. Assoc.*, 84:200, 1989.

[26] J. O. Berger and J. M. Bernardo. *Biometrika*, 79:25, 1992.

[27] L Demortier, S. Jain, and H. B. Prosper. *Phys.Rev.*, D82:034002, 2010.

[28] E. Gross and O. Vitells. *Eur. Phys. Jour. C*, 70:525–530.

[29] The ATLAS Collaboration. Feb 2011. ATLAS-CONF-2011-005.

[30] H.O. Lancaster. *Biometrika*, 39:419–422, 1949.

[31] J. Przyborowski and H. Wilenski. *Biometrika*, 31:313, 1940.
[32] F. James and M. Roos. *Nuclear Physics B*, 172:475, 1980.
[33] N. Reid. *Stat. Sci.*, 10:138, 1995.
[34] R. D. Cousins, K. E. Hymes, and J. Tucker. *Nucl.Instrum.Meth.*, A612:388–398, 2010.

Use of the profile likelihood function in searches for new physics

Glen Cowan
Physics Department, Royal Holloway, University of London, Egham TW20 0EX, UK

Abstract
We describe likelihood-based statistical tests for use in high energy physics for the discovery of new phenomena and for construction of confidence intervals. Explicit formulae for the asymptotic distributions of test statistics based on the profile likelihood ratio are derived using results of Wilks and Wald. We motivate and justify the use of a representative data set, called the "Asimov data set", which provides a simple method to obtain the median experimental sensitivity of a search or measurement as well as fluctuations about this expectation.

1 Introduction

This paper summarizes results recently published in Ref. [1]. These allow one to carry out statistical tests in searches for processes that have been predicted but not yet seen, such as production of a Higgs boson. The statistical significance of an observed signal can be quantified by means of a p-value or its equivalent Gaussian significance (discussed below). It is useful to characterize the sensitivity of an experiment by reporting the expected (e.g., mean or median) significance that one would obtain for a variety of signal hypotheses.

Finding both the significance for a specific data set and the expected significance can involve Monte Carlo calculations that are computationally expensive. The approximate methods reported here are based on results due to Wilks [2] and Wald [3], which allow one to obtain both the significance for given data as well as the full sampling distribution of the significance under the hypothesis of different signal models, all without recourse to Monte Carlo.

In Sec. 2 the formalism of a search as a statistical test is outlined. Several test statistics based on the profile likelihood ratio are defined in Sec. 3 that can be used for establishing a discovery or setting upper limits. Example applications are shown in Sec. 4, and conclusions are given in Sec. 5.

2 Formalism of a search as a statistical test

In this section we outline the general procedure used to search for a new phenomenon in the context of a frequentist statistical test. For purposes of discovering a new signal process, one defines the null hypothesis, H_0, as describing only known processes, here designated as background. This is to be tested against the alternative H_1, which includes both background as well as the sought after signal. When setting limits, the model with signal plus background plays the role of H_0, which is tested against the background-only hypothesis, H_1.

To summarize the outcome of such a search one quantifies the level of agreement of the observed data with a given hypothesis H by computing a p-value, i.e., a probability, under assumption of H, of finding data of equal or greater incompatibility with the predictions of H. One can regard the hypothesis as excluded if its p-value is observed below a specified threshold. In particle physics one usually converts the p-value into an equivalent significance, Z, defined such that a Gaussian distributed variable found Z standard deviations above its mean has an upper-tail probability equal to p. That is, $Z = \Phi^{-1}(1-p)$, where Φ^{-1} is the quantile (inverse of the cumulative distribution) of the standard Gaussian.

It is often useful to quantify the sensitivity of an experiment by reporting the expected (or more precisely, the median) significance one would obtain with a given measurement under the assumption

of various hypotheses. For example, the sensitivity to discovery of a given signal process H_1 could be characterized by the median value, under the assumption of H_1, of the value of Z obtained from a test of H_0.

Consider an experiment where for each selected event one measures the values of certain kinematic variables, and thus the resulting data can be represented as one or more histograms. Using the method in an unbinned analysis is a straightforward extension. Suppose for each event in the signal sample one measures a variable x and uses these values to construct a histogram $\vec{n} = (n_1, \ldots, n_N)$. The expectation value of n_i can be written $E[n_i] = \mu s_i + b_i$, where the mean number of entries in the ith bin from signal and background are

$$s_i = s_{\rm tot} \int_{{\rm bin}\,i} f_s(x; \vec{\theta}_s)\, dx \;, \qquad b_i = b_{\rm tot} \int_{{\rm bin}\,i} f_b(x; \vec{\theta}_b)\, dx \,. \tag{1}$$

Here the parameter μ determines the strength of the signal process, with $\mu = 0$ corresponding to the background-only hypothesis and $\mu = 1$ being the nominal signal hypothesis. The functions $f_s(x; \vec{\theta}_s)$ and $f_b(x; \vec{\theta}_b)$ are the probability density functions (pdfs) of the variable x for signal and background events, and $\vec{\theta}_s$ and $\vec{\theta}_b$ represent parameters that characterize the shapes of pdfs. The quantities $s_{\rm tot}$ and $b_{\rm tot}$ are the total mean numbers of signal and background events. Below we will use $\vec{\theta} = (\vec{\theta}_s, \vec{\theta}_b, b_{\rm tot})$ to denote all of the nuisance parameters. The signal normalization $s_{\rm tot}$ is not, however, an adjustable parameter but rather is fixed to the value predicted by the nominal signal model.

In addition to the measured histogram $\vec{n}$ one often makes subsidiary measurements that help constrain the nuisance parameters. For example, one may select a control sample where one expects mainly background events and from them construct a histogram of some chosen kinematic variable. This then gives a set of values $\vec{m} = (m_1, \ldots, m_M)$ for the number of entries in each of the M bins. The expectation value of m_i can be written $E[m_i] = u_i(\vec{\theta})$, where the u_i are calculable quantities depending on the parameters $\vec{\theta}$. One often constructs this measurement so as to provide information on the background normalization parameter $b_{\rm tot}$ and also possibly on the signal and background shape parameters. The likelihood function is the product of Poisson probabilities for all bins:

$$L(\mu, \vec{\theta}) = \prod_{j=1}^{N} \frac{(\mu s_j + b_j)^{n_j}}{n_j!} e^{-(\mu s_j + b_j)} \;\; \prod_{k=1}^{M} \frac{u_k^{m_k}}{m_k!}\, e^{-u_k} \;. \tag{2}$$

To test a hypothesized value of μ we consider the profile likelihood ratio (see, e.g., [4]),

$$\lambda(\mu) = \frac{L(\mu, \hat{\hat{\vec{\theta}}})}{L(\hat{\mu}, \hat{\vec{\theta}})} \;. \tag{3}$$

Here $\hat{\hat{\vec{\theta}}}$ in the numerator denotes the value of $\vec{\theta}$ that maximizes L for the specified μ, i.e., it is the conditional maximum-likelihood (ML) estimator of $\vec{\theta}$ (and thus is a function of μ). The denominator is the maximized (unconditional) likelihood function, i.e., $\hat{\mu}$ and $\hat{\vec{\theta}}$ are the ML estimators.

In many analyses, the contribution of the signal process to the mean number of events is assumed to be nonnegative, which is to say that any physical estimator for μ must be nonnegative. Even if we regard this to be the case, however, it is convenient to define an effective estimator $\hat{\mu}$ as the value of μ that maximizes the likelihood, even this gives $\hat{\mu} < 0$ (but providing that the Poisson mean values, $\mu s_i + b_i$, remain nonnegative). This will allow us in Sec. 3 to model $\hat{\mu}$ as a Gaussian distributed variable, and in this way we can determine the distributions of the test statistics that we consider. Therefore in the following we will always regard $\hat{\mu}$ as an effective estimator which is allowed to take on negative values.

3 Test statistics for discovery and upper limits

In this section we present test statistics based on the profile likelihood ratio. To compute p-values and sensitivities one requires the sampling distributions of these statistics. These are given below in an approximate form valid in the large-sample limit. More details can be found in Ref. [1].

3.1 Test statistic $t_\mu = -2\ln\lambda(\mu)$

From the definition of $\lambda(\mu)$ in Eq. (3), one can see that $0 \le \lambda \le 1$, with λ near 1 implying better agreement between the data and the hypothesized value of μ. Equivalently it is convenient to use the statistic $t_\mu = -2\ln\lambda(\mu)$ as the basis of a statistical test. Higher values of t_μ thus correspond to increasing incompatibility between the data and μ. To quantify the level of disagreement we compute the p-value,

$$p_\mu = \int_{t_{\mu,\mathrm{obs}}}^{\infty} f(t_\mu|\mu)\, dt_\mu \ , \tag{4}$$

where $t_{\mu,\mathrm{obs}}$ is the value of the statistic t_μ observed from the data and $f(t_\mu|\mu)$ denotes the pdf of t_μ under the assumption of the signal strength μ. The p-values for all of the statistics considered here are obtained in an analogous fashion.

To find the distribution of t_μ as well as that of other related statistics, we can use a relation due to Wald [3], who showed that for the case of a single parameter of interest,

$$-2\ln\lambda(\mu) = \frac{(\mu-\hat{\mu})^2}{\sigma^2} + \mathcal{O}(1/\sqrt{N}) \ . \tag{5}$$

Here $\hat{\mu}$ follows a Gaussian distribution with a mean μ' and standard deviation σ, and N represents the data sample size. The approximations presented here are valid to the extent that the $\mathcal{O}(1/\sqrt{N})$ term can be neglected.

If $\hat{\mu}$ is Gaussian distributed and we neglect the $\mathcal{O}(1/\sqrt{N})$ term in Eq. (5), then one can show that the statistic $t_\mu = -2\ln\lambda(\mu)$ follows a *noncentral chi-square* distribution for one degree of freedom,

$$f(t_\mu;\Lambda) = \frac{1}{2\sqrt{t_\mu}}\frac{1}{\sqrt{2\pi}}\left[\exp\left(-\frac{1}{2}\left(\sqrt{t_\mu}+\sqrt{\Lambda}\right)^2\right) + \exp\left(-\frac{1}{2}\left(\sqrt{t_\mu}-\sqrt{\Lambda}\right)^2\right)\right] \ , \tag{6}$$

where the noncentrality parameter is $\Lambda = (\mu-\mu')^2/\sigma^2$. For the special case $\mu' = \mu$ one has $\Lambda = 0$ and the pdf of $-2\ln\lambda(\mu)$ approaches a chi-square distribution for one degree of freedom, a result shown earlier by Wilks [2].

3.2 The statistic q_0 for discovery

Often one wishes to test $\mu = 0$ in a class of models where we assume $\mu \ge 0$. Rejecting $\mu = 0$ amounts to discovering a new (positive) signal. In such a case one can define the test such that the data are only regarded as discrepant with the hypothesis of $\mu = 0$ if one observes an excess of events, i.e., one finds $\hat{\mu} > 0$. That is, we define the statistic

$$q_0 = \begin{cases} -2\ln\lambda(0) & \hat{\mu} \ge 0 \ , \\ 0 & \hat{\mu} < 0 \ , \end{cases} \tag{7}$$

where $\lambda(0)$ is the profile likelihood ratio for $\mu = 0$ as defined in Eq. (3).

Assuming the validity of the Wald approximation, one can show that the pdf of q_0 has the form

$$f(q_0|\mu') = \left(1 - \Phi\left(\frac{\mu'}{\sigma}\right)\right)\delta(q_0) + \frac{1}{2}\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{q_0}}\exp\left[-\frac{1}{2}\left(\sqrt{q_0} - \frac{\mu'}{\sigma}\right)^2\right] . \tag{8}$$

From Eq. (8) the corresponding cumulative distribution is found to be

$$F(q_0|\mu') = \Phi\left(\sqrt{q_0} - \frac{\mu'}{\sigma}\right) . \tag{9}$$

The p-value of the hypothesis $\mu = 0$, p_0, is obtained from these distributions by using $\mu' = 0$. For the significance one finds the simple formula

$$Z_0 = \Phi^{-1}(1 - p_0) = \sqrt{q_0} . \tag{10}$$

3.3 The statistic q_μ for upper limits

For purposes of establishing an upper limit on the strength parameter μ, one can define

$$q_\mu = \begin{cases} -2\ln\lambda(\mu) & \hat{\mu} \le \mu , \\ 0 & \hat{\mu} > \mu , \end{cases} \tag{11}$$

where $\lambda(\mu)$ is the profile likelihood ratio from Eq. (3). The reason for setting $q_\mu = 0$ for $\hat{\mu} > \mu$ is that when setting an upper limit, one would not regard data with $\hat{\mu} > \mu$ as representing less compatibility with μ than the data obtained, and therefore this is not taken as part of the rejection region of the test. From the definition of the test statistic one sees that higher values of q_μ represent greater incompatibility between the data and the hypothesized value of μ. A closely related statistic, which we call $\tilde{q}_\mu$, is discussed in Ref. [1]. In the large-sample limit they are equivalent.

Assuming the validity of the Wald approximation, the pdf $f(q_\mu|\mu')$ is found to be

$$f(q_\mu|\mu') = \Phi\left(\frac{\mu' - \mu}{\sigma}\right)\delta(q_\mu) + \frac{1}{2}\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{q_\mu}}\exp\left[-\frac{1}{2}\left(\sqrt{q_\mu} - \frac{\mu - \mu'}{\sigma}\right)^2\right] , \tag{12}$$

and the cumulative distribution is

$$F(q_\mu|\mu') = \Phi\left(\sqrt{q_\mu} - \frac{\mu - \mu'}{\sigma}\right) . \tag{13}$$

Using these ingredients with $\mu' = 0$, one can obtain the p-value of a hypothesized value of μ, p_μ, and the corresponding significance, Z_μ, which is found to be

$$Z_\mu = \Phi^{-1}(1 - p_\mu) = \sqrt{q_\mu} . \tag{14}$$

3.4 Asimov data set, variance of $\hat{\mu}$, median significance

Some of the formulae above require the standard deviation σ of $\hat{\mu}$. A useful way of estimating σ involves a special, artificial data set that we call the "Asimov data set". This is defined such that when it is used to evaluate the estimators for all parameters, one obtains the true parameter values. One can show that under conditions generally satisfied in practice, this amounts to setting the Poisson data values equal to their expectation values, which can be estimated using a very large Monte Carlo data sample. That is, the Asimov values for the measured histograms $\vec{n}$ and $\vec{m}$ are $n_{i,\mathrm{A}} = E[n_i] = \mu' s_i(\vec{\theta}) + b_i(\vec{\theta})$ and $m_{i,\mathrm{A}} = E[m_i] = u_i(\vec{\theta})$.

We can use the Asimov data set to evaluate the "Asimov likelihood" $L_{\rm A}$ and the corresponding profile likelihood ratio $\lambda_{\rm A}$. Because the Asimov data set corresponding to a strength μ' gives $\hat{\mu} = \mu'$, from Eq. (5) one finds

$$-2\ln\lambda_{\rm A}(\mu) \approx \frac{(\mu-\mu')^2}{\sigma^2} = \Lambda \; . \tag{15}$$

That is, the Asimov data set provides an estimate of the noncentrality parameter Λ that characterizes the distribution of $-2\ln\lambda(\mu)$. Equivalently, one can use Eq. (15) to obtain the variance of $\hat{\mu}$, σ^2.

When the statistics q_0 and q_μ are evaluated with an Asimov data set (denoted $q_{0,{\rm A}}$ and $q_{\mu,{\rm A}}$) one obtains good estimates for their median values, and these lead to simple expressions for the corresponding median significance. From Eqs. (10) and (14) one sees that the significance Z is a monotonic function of q, and therefore the median Z is simply given by the corresponding function of the median of q. For discovery using q_0 one wants the median discovery significance assuming a strength parameter μ' and for upper limits one is particularly interested in the median exclusion significance assuming $\mu' = 0$, med$[Z_\mu|0]$. Using the corresponding Asimov data set for each case, one finds

$$\text{med}[Z_0|\mu'] \;=\; \sqrt{q_{0,{\rm A}}} \; , \tag{16}$$

$$\text{med}[Z_\mu|0] \;=\; \sqrt{q_{\mu,{\rm A}}} \; . \tag{17}$$

4 Tests of asymptotic formulae

Several tests of the validity of the asymptotic formulae given above are described in Ref. [1]. Here as an example we consider a measurement consisting of a number of events n assumed to be Poisson distributed with a mean $\mu s+b$, and a control measurement m modeled as following a Poisson distribution with mean τb. Here s and τ are treated as known with $\tau = 1$, b is a nuisance parameter and μ is the parameter of interest. Figure 1(a) shows the distributions from the asymptotic formula as well as histograms from Monte Carlo using different values of b. One can see that even for b as low as 2, the asymptotic curve agrees out to $q_0 \approx 10$, corresponding to a discovery significance of $Z_0 \approx \sqrt{10}$. To establish a 5σ effect one needs to model the distribution beyond $q_0 = 25$, which is achieved reasonably well here for $b = 20$.

As a second example, Fig. 1(b) shows the median discovery significance with which one would reject $\mu = 0$ assuming data distributed according to $\mu = 1$ in an experiment where n is Poisson distributed with mean $\mu s + b$, but here b is known exactly and there is no control measurement. The exact values shown as points are determined from Monte Carlo, and the jumps are a consequence of the discreteness of the data. Using the Asimov data value $s + b$ to approximate the median significance, one finds

$$\text{med}[Z_0|1] = \sqrt{q_{0,{\rm A}}} = \sqrt{2\left((s+b)\ln(1+s/b) - s\right)} \; . \tag{18}$$

Expanding the logarithm to second order in s/b one finds med$[Z_0|1] = (s/\sqrt{b})\,(1 + \mathcal{O}(s/b))$. Although $Z_0 \approx s/\sqrt{b}$ has been widely used for cases where $s + b$ is large, this final approximation is strictly valid only for $s \ll b$, as can be seen in Fig. 1(b).

5 Conclusions

Statistical tests are described for use in planning and carrying out a search for new phenomena; further details can be found in Ref. [1]. Approximate formulae are given for the distributions of test statistics used to characterize the level of agreement between the data and the hypothesis being tested, as well as the related expressions for p-values and significances. The formulae are implemented in the `RooStats` software package [5].

(a) (b)

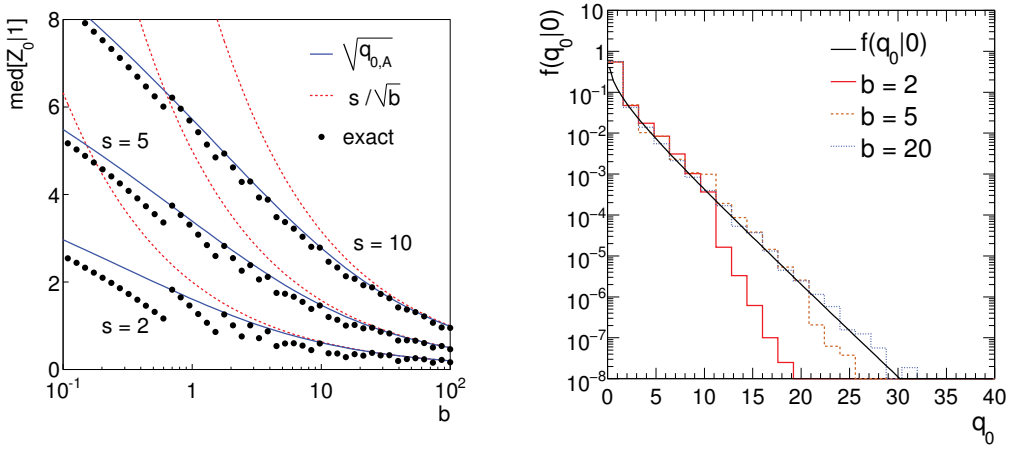


Fig. 1: (a) The pdf $f(q_0|0)$ for the counting experiment. The solid curve shows $f(q_0|0)$ from the asymptotic formula and the histograms are from Monte Carlo using different values of b (see text). (b) The median, assuming $\mu = 1$, of the discovery significance Z_0 for different values of s and b (the plot shown here corrects a minor numerical error in Fig. 7 of Ref. [1]).

The asymptotic formulae free one from the need to carry out lengthy Monte Carlo calculations, which in the case of a discovery at 5σ significance could require simulation of around 10^8 measurements. The approximations used are valid in the limit of a large data sample. Tests with Monte Carlo indicate, however, that the formulae are in fact reasonably accurate even for fairly small samples, and thus can have a wide range of practical applicability. For very small samples and in cases where high accuracy is crucial, one is always free to validate the approximations with Monte Carlo.

Acknowledgements

I thank my coauthors in Ref. [1], Eilam Gross, Kyle Cranmer and Ofer Vitells; this report is presented here on all of our behalf. I also thank the organizers of PHYSTAT 2011, especially Louis Lyons and Harrison Prosper, for an enjoyable and stimulating meeting.

References

[1] G. Cowan, K. Cranmer, E. Gross and O. Vitells, Eur. Phys. J. C (2011) 71:1554; arXiv:1007.1727 [physics.data-an].

[2] S.S. Wilks, *The large-sample distribution of the likelihood ratio for testing composite hypotheses*, Ann. Math. Statist. **9** (1938) 60-2.

[3] A. Wald, *Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations is Large*, Transactions of the American Mathematical Society, Vol. **54**, No. 3 (Nov., 1943), pp. 426-482.

[4] A. Stuart, J.K. Ord, and S. Arnold, *Kendall's Advanced Theory of Statistics*, Vol. 2A: *Classical Inference and the Linear Model* 6th Ed., Oxford Univ. Press (1999), and earlier editions by Kendall and Stuart.

[5] L. Moneta, K. Belasco, K. Cranmer *et al.*, *The RooStats Project*, proceedings of ACAT, 2010, Jaipur, India; arXiv:1009.1003 [physics.data-an].

Incorporating Nuisance Parameters in Likelihoods for Multisource Spectra

J. S. Conway
University of California, Davis, USA

Abstract

We describe here the general mathematical approach to constructing likelihoods for fitting observed spectra in one or more dimensions with multiple sources, including the effects of systematic uncertainties represented as nuisance parameters, when the likelihood is to be maximized with respect to these parameters. We consider three types of nuisance parameters: simple multiplicative factors, source spectra "morphing" parameters, and parameters representing statistical uncertainties in the predicted source spectra.

1 Overview

In particle physics one often encounters the general problem of estimating physical parameters such as particle masses or cross sections from the spectra of observables calculated in each event. In the case of a known, well-established signal process, the dominant technique by far is to use a binned likelihood assuming a Poisson distribution in each bin [1], and find the parameters which maximize the likelihood.

In the case of a search for a new particle or effect resulting in either a discovery or null result, binned likelihoods have also been employed successfully to quote statistical significance or exclusion bounds, respectively. From a certain point of view there is a desirable consistency in utilizing the same basic statistical method for searches, discoveries, and measurements.

A key requirement here, however, is that the likelihood somehow incorporate the effects of all systematic uncertainties present in the analysis. In frequentist inspired methods, the effect of systemtic uncertainties is very often incorporated by the non-frequentist procedure of generating distributions of many pseudoexperiments, where from one pseudoexperiment to the next the values of all parameters are varied within their assumed distributions. In a formal Bayesian treatment, the nuisance parameters are removed by marginalization: integrating them out, assuming some prior pdf. Both of these approaches are computationally very expensive.

In measuring parameters using binned Poisson likelihoods, as mentioned above, one simply maximizes the likelihood (in practice one minimizes the negative log of the likelihood) with respect to all m free parameters, and then constructs the standard error ellipsoid in m-dimensional space. The fit values of the nuisance parameters are typically of no interest, leaving one to interpret the intervals for just the parameters of interest in a straightforward way [2].

We define in this paper three main types of nuisance parameters representing systematic uncertainties on the source distributions, and describe how to incorporate them into a binned Poisson likelihood.

We further argue in this paper that this maximum likelihood method, also called the profile likelihood, can be applied to searches and discoveries as well, either by a pseudo-Bayesian interpretation of the profile likelihood as leading to a posterior density in the parameter(s) of interest (after suitable inclusion of a prior), or by likelihood ratio methods. The profile likelihood requires significantly less computer time, often as much as two orders of magnitude less, than frequentist or frequentist-inspired methods such as $\mathrm{CL_s}$ [3]. That in turn allows much more detailed study of the properties of the fit results.

2 Core of the Poisson Likelihood

Suppose we observe in a set of N events an observable or in general a set of observables $\bar{x}$. If we define a set of n_{bin} bins (which can be of literally any shape we choose) in the space of the observables, then

the number of events n_i in each bin i, is assumed to be Poisson-distributed according to

$$\mathcal{P}(n_i|\mu_i) = \frac{\mu_i^{n_i} e^{-\mu_i}}{n_i!} \tag{1}$$

where μ_i is the number of expected events in the bin. Typically we can write

$$\mu_i = \sum_{j=1}^{n_{source}} L\sigma_j \epsilon_{ji} \tag{2}$$

for integrated luminosity L, cross section σ_j for source j, and efficiency ϵ_{ji} for source j in bin i, often obtained from MC simulation of the process. The sources here include the signal process of interest and all background processes. Again, since we are dealing with a possibly multidimensional space of observables, the index i can actually label the bins in multiple dimensions.

The Poisson likelihood for the full observed spectrum is simply the product of the Poisson probabilities:

$$\mathcal{L} = \prod_{i=1}^{N} \mathcal{P}(n_i|\mu_i) \quad . \tag{3}$$

In the absence of any systematic uncertainties one can simply minimize $-\ln \mathcal{L}$ with respect to all unknown parameters in the problem and interpret the resulting standard error ellipsoid in the normal way to obtain estimates of the unknown parameters and associated confidence intervals.

3 Normalization Uncertainties

Normalization uncertainties provide the simplest example of systematic uncertainties that can be represented by nuisance parameters in profile likelihoods. As an example, let us assume that the integrated luminosity is measured in some auxiliary study, and results in a 2% uncertainty. We would rewrite the likelihood as

$$\mathcal{L} = \prod_{i=1}^{N} \mathcal{P}(n_i|\mu_i)\mathcal{G}(L|\tilde{L}, \sigma_L) \tag{4}$$

for the measured value $\tilde{L} \pm \sigma_L$. The function $\mathcal{G}$ is a normalized Gaussian of mean $\tilde{L}$ and width σ_L, which serves to constrain the value of the new nuisance parameter L to its measured value. Note that it is L and not $\tilde{L}$ that is used to calculate the μ_i. The negative log likelihood is thus

$$-ln\mathcal{L} = \sum_i [-n_i \ln \mu_i + \mu_i] + \frac{(L-\tilde{L})^2}{2\sigma_L^2} \tag{5}$$

and thus the remnant of the Gaussian term can be regarded as a penalty on the negative log likelihood. It is in principle possible to use functions other than Gaussians to constrain the values of the nuisance parameters. In Bayesian terms the constraint functions are simply the prior probability densities of the nuisance parameters.

Any normalization uncertainty can be represented in the likelihood this way, including uncertainties on cross sections, overall efficiencies, and the like, simply by introducing multiplicative nuisance parameters into Eq. 2 as needed, for any or all sources.

In many cases, however, the allowed physical bound on a multiplicative nuisance parameter is that it remain positive. If we are representing the constraint by a Gaussian, then when the uncertainty in the nuisance parameter is large the Gaussian is truncated and an appropriate normalization factor should be included. It must also be realized that any such truncation shifts the mean of the distribution and tends to introduce a bias away from the most probable value of the parameter. In such cases one might also consider constraining the parameter with a log normal or other probability density which does not allow the parameter to become negative.

4 Shape Uncertainties and Morphing

Many systematic uncertainties result in an overall distortion in the shape of the observed spectrum. A good example is an energy scale uncertainty which affects all jet energies in an event in the same direction. If there are energy thresholds in the event selection, changes in not only the shape but the overall normalization of the efficiency (represented here by the ϵ_{ji} for source j in bin i) as a function of the observables can result.

Such spectral distortions can be modeled by altering parameters (like the energy scale) in the MC simulation and recalculating the "shifted" set of efficiencies. If we were, for example, to raise and lower the energy scale by one standard deviation, recalculating the efficiencies, we would then have three measures of the shape (and normalization) of the bin efficiency distribution, which we can call ϵ^-_{ji}, ϵ^0_{ji}, and ϵ^+_{ji}. Clearly one can obtain more measures from other alterations of the energy scale, though this can often be computationally very expensive.

We then face the question of how to turn our three measures of the spectral shape into a continuous estimate in each bin as a function of the energy scale factor. To do this we introduce a "morphing" parameter which we will call f, and which is nominally zero (in the case of no scale shift), and which has some uncertainty (usually Gaussian) σ_f=1.

In this general technique, usually called "vertical morphing", we interpolate quadratically between the three efficiencies in a bin for $|f| < 1$ and extrapolate linearly beyond that range. This does result in the exact measured behavior of the spectrum at $f = \pm 1$ but avoids large deviations from linear behavior outside the range. The value of the efficiency at any $|f| < 1$ can be determined by Lagrange interpolation:

$$\epsilon_{ji} = \frac{f(f-1)}{2}\epsilon^-_{ji} - (f-1)(f+1)\epsilon^0_{ji} + \frac{f(f+1)}{2}\epsilon^+_{ji} \tag{6}$$

Calculation of the linear extrapolation beyond this range is a straightforward exercise for the reader.

Clearly if a more accurate representation of the morphing behavior is required, one can, at the expense of computation and bookkeeping time, obtain additional shifted efficiency spectra and interpolate using a higher order polynomial. A good measure of whether this is a worthwhile exercise is to examine the behavior of one's morphing parameters as a function of the parameter of interest; if they tend to go far from the sampled region (corresponding to one standard deviation in the uncertainty) then it may be desirable to obtain more measurements there, and parametrize the measured region with a higher order polynomial.

We also note that there are somewhat more sophisticated methods such as Alex Read's "horizontal morphing" [4] method. These are more computationally intensive, but could be advantageous. However they are not straightforwardly defined in more than one dimension.

The morphing method presented here can be extended to several morphing parameters for different independent systematic effects simply by adding linearly the deviations from the nominal efficiency due to each effect.

5 Statistical Uncertainties in Efficiencies

Typically one estimates the efficiency of each source in each bin using a Monte Carlo simulation, and hence the statistical accuracy of the estimate of the efficiency in each bin depends on the number of MC events falling there. Likewise, in other, possibly data-driven methods for estimating the expected number of events from some source in some bin, there may be some known statistical uncertainty in each bin.

Barlow and Beeston [5] proposed a method for representing such systematic uncertainties wherein one introduces a separate nuisance parameter multiplying the expected number of events from each source in each bin. Nominally the value of these parameters is 1, and one can then constrain the parameters, which we call β_{ji}, according to the prior pdf assumed for the number of MC events in the efficiency calculation. Barlow and Beeston assumed a Poisson distribution (though one might argue a binomial is

the most correct form to assume); other choices such as log normal avoid the parameters possibly tending to negative values.

Though this method introduces a very large number of new free parameters in the likelihood, the problem can be seen to be tractable in the profile likelihood case since the values of the β_{ji} which maximize the likelihood within a bin can be found independently of those in all the other bins.

Assuming a Gaussian constraint on the β_{ji}, we can write the contribution to the negative log likelihood in a particular bin as

$$-ln\mathcal{L}_i = -n_i \ln(\sum_j \beta_{ji}\mu_{ji}) + \sum_j \beta_{ji}\mu_{ji} + \sum_j \frac{(\beta_{ji}-1)^2}{2\sigma_{ji}^2} \quad . \tag{7}$$

This contribution can be minimized with respect to the β_{ji} by setting the derivative with respect to each to zero. Dropping the bin index i for clarity we write

$$\frac{\partial(-\ln\mathcal{L})}{\partial\beta_j} = \mu_j \left[1 - \frac{n}{\sum_k \beta_k\mu_k}\right] + \frac{\beta_j - 1}{\sigma_j^2} = 0 \quad . \tag{8}$$

We thus arrive at a set of nonlinear equations for the β_j in a bin. These can be approximately solved by iterative Newton-type methods, or by more sophisticated methods.

In the context of performing the profile likelihood using MINUIT minimization, one can implement this Barlow-Beeston type method by solving for the β_{ji} within the "objective" function which provides to MINUIT the value of $-\ln L$ given the values of all the parameters in the fit, and includes the contribution of the deviations of the β_{ji} from unity to $-\ln L$.

However, a problem arises in this approach. Any minimization algorithm can only approximate the values of the parameters and, hence, the true minimum of a function. There is always some last step which meets the convergence criterion, and somewhere in the space of the input μ_{ji} to the minimization for the β_{ji}, one will find the place where that last step is not taken. Near such points the values of the resulting β_{ji} and their associated contribution to $-\ln L$ undergo a small discontinuous jump. Such jumps can (and do) dramatically confuse MINUIT's MIGRAD minimizer, which attempts to measure the Hessian matrix by finite differences. These jumps cause the resulting parameter covariance matrix to become non-positive-definite. When MINUIT detects such a situation it attempts to circumvent it by adding to the offending diagonal element of the matrix an amount necessary to restore positive-definiteness. Sometimes this works but in many cases all is lost: MINUIT is now dealing with a false measure of the Hessian matrix and it tends to send the free parameters in the fit to wild values. We have found no solution to this behavior short of rewriting MINUIT.

The full-blown Barlow-Beeston method for dealing with bin statistical uncertainties is not absolutely required to represent them properly in the likelihood. What matters, in a bin, is the *overall* statistical uncertainty of the predicted number of events from all sources. The statistical uncertainties for each source in each bin are independent, and can be readily combined, particularly if they are Gaussian or Poisson in nature. Thus, a single Barlow-Beeston type parameter is sufficient to represent the statistical uncertainty.

If we make the approximation that the overall uncertainty in the bin can be approximated by a Gaussian of some width, then the value of this parameter, and its contribution to $-\ln L$, can be calculated exactly by solving a quadratic equation. Using a simplified notation for a single bin, we write

$$-\ln\mathcal{L} = -n\ln\beta\mu + \beta\mu + \frac{(\beta-1)^2}{2\sigma_\beta^2} \tag{9}$$

where here μ is the total number of expected events in the bin, given the values of all the other parameters, and σ_β is the relative (statistical) uncertainty in the prediction. Setting the derivative to zero we find the

quadratic equation

$$\beta^2 + (\mu\sigma_\beta^2 - 1)\beta - n\sigma_\beta^2 = 0 \tag{10}$$

which can be solved readily and the correct root taken. The extension to other constraint functions is straightforward though it may result in transcendental equations to solve.

6 Practical Considerations

Care must be taken in using the approach described in this paper to avoid a number of potential pitfalls which we discuss here.

Sparsely Populated Bins

In multi-bin spectra (particularly multi-dimensional spectra) one can encounter situations where the number of events per bin varies by orders of magnitude. This can sometimes lead to situations where

- there can be regions of zero-content bins, surrounded by bins populated by single MC events;
- such single MC-event-bins can migrate under the influence of the morphing systematic effects, spoiling the vertical morphing method;
- single data events can appear in bins where there is no predicted rate.

All of these situations must be avoided. The most straightforward is to generate sufficient Monte Carlo in all bins, but this may not be practical or even possible. The best alternative is to combine bins according to some algorithm (which does not use the observed data distribution!) which ensures some minimum statistical threshold in every bin in the fit.

Bins Entering/Leaving the Likelihood

It is also necessary to ensure that no bin enters or leaves the likelihood as the parameters change. It is not impossible for MINUIT to drive parameters to regions where the contribution from a source, or even all sources, vanishes in a bin. For example, when studying the profile likelihood as a function of some new particle signal, in general one wants to evaluate the likelihood for the case of zero signal. But if there are bins populated by signal only, this can cause the contribution to go to zero, the logarithm of which is of course $-\infty$.

Simply excluding bins from the likelihood when there are no expected events is not a sufficient solution to this problem, as a moment's reflection will make clear. To avoid bins entering/leaving the fit, therefore, the bins to be used or not used must be established *a priori* by finding all bins where some contribution is expected, and making sure there are no bins with data but no expected contribution. Once determined, this set must remain fixed for the duration of the calculation.

One way to ensure that no bin leaves the calculation is to always have it contribute at least some tiny amount. For example to circumvent the zero-signal issue, we always ensure that the signal cross section is no less than 10^{-10} pb, and that no source in any bin used in the fit ever contributes less than 10^{-10} expected events. Though this is a somewhat inelegant solution to a nevertheless important problem, we note that our final results do not depend on these minimum values in practice.

7 Pseudo-Bayesian Posterior Densities

For measuring physical parameters, the profile likelihood can be directly interpreted using the usual $\Delta(\ln L)$ approach to derive confidence intervals in multi-dimensional parameter space.

To extend this treatment to setting exclusion bounds on parameters such as a hypothetical new particle's cross section σ_X, we can simply derive a posterior density by treating the profile likelihood,

which we shall denote $\mathcal{L}_{prof}$ as one would any likelihood using Bayes' Theorem:

$$\mathcal{P}(\sigma_X) = \frac{\mathcal{L}_{prof}(\sigma_X)\pi(\sigma_X)}{\int_0^\infty \mathcal{L}_{prof}(\sigma_X)\pi(\sigma_X)d\sigma_X} \tag{11}$$

where here $\pi(\sigma_X)$ is the assumed prior pdf in σ_X.[1]

But does the profile method really result in a posterior density that can be interpreted in this way? The most proper Bayesian treatment would not maximize the likelihood with respect to the nuisance parameters, but marginalize instead, resulting in what we might denote as $\bar{\mathcal{L}}(\sigma_X)$ to highlight the fact that the marginalized likelihood is in a sense the core likelihood averaged over the prior-weighted values of the nuisance parameters.

We have performed both calculations, profiling and marginalization, in a variety of complex spectrum fits, and it is our experience that the posterior densities derived both ways are nearly identical, though the marginalized one takes orders of magnitude more compute time. Due to this practical consideration alone we employ the profile method and consider it to be a near-perfect representation of a full and proper Bayesian marginalization treatment.

8 Conclusions

We present in this paper the basic mathematical and numerical approach to fitting multi-source spectra using a profile likelihood in which various types of systematic uncertainties are incorporated by representing them by nuisance parameters. This method, we believe, offers a unified approach to setting exclusion bounds, making discoveries, and ultimately performing measurements on a wide range of particle physics data analyses.

Acknowledgements

I wish to thank my CDF t' colleagues Andrew Ivanov, Robin Erbacher, Alison Lister, David Cox, Will Johnson, and Thomas Schwarz for their insights and ideas in developing these methods. This work was supported by the US Department of Energy Office of Science.

References

[1] G. Cowan, *Statistical Data Analysis*, Oxford University Press, 1998.

[2] C. Amsler et al. (Particle Data Group), Physics Letters B667 (2008) 1; available at http://pdg.lbl.gov/2009/reviews/rpp2009-rev-statistics.pdf.

[3] A. Read, J. Phys. G: Nucl. Part. Phys. 28 (2002) 2693.

[4] A. Read, Nucl. Instrum. Meth. Res. A 425 (1999) 357-369.

[5] R. Barlow and C. Beeston, Comp. Phys. Comm. 77 (1993) 219.

[1] All the usual cautions against improper priors apply at this point. We would like to point out, however, that in nearly every case of which we are aware, where such a posterior is used to quote confidence intervals on the parameter of interest in an actual *measurement* of that parameter, no one typically uses a prior other than a uniform one.

Parton distributions: determining probabilities in a space of functions

The NNPDF Collaboration: *Richard D. Ball*[1]*, Valerio Bertone*[2]*, Francesco Cerutti*[3]*, Luigi Del Debbio*[1]*, Stefano Forte*[4]*, Alberto Guffanti*[5]*, José I. Latorre*[3]*, Juan Rojo*[4*] *and Maria Ubiali*[6]*.*
[1] School of Physics and Astronomy, University of Edinburgh,
JCMB, KB, Mayfield Rd, Edinburgh EH9 3JZ, Scotland
[2] Physikalisches Institut, Albert-Ludwigs-Universität Freiburg,
Hermann-Herder-Straße 3, D-79104 Freiburg i. B., Germany
[3] Departament d'Estructura i Constituents de la Matèria, Universitat de Barcelona,
Diagonal 647, E-08028 Barcelona, Spain
[4] Dipartimento di Fisica, Università di Milano and INFN, Sezione di Milano,
Via Celoria 16, I-20133 Milano, Italy
[5] The Niels Bohr International Academy and Discovery Center,
The Niels Bohr Institute, Blegdamsvej 17, DK-2100 Copenhagen, Denmark
[6] Institut für Theoretische Teilchenphysik und Kosmologie, RWTH Aachen University,
D-52056 Aachen, Germany

Abstract
We discuss the statistical properties of parton distributions within the framework of the NNPDF methodology. We present various tests of statistical consistency, in particular that the distribution of results does not depend on the underlying parametrization and that it behaves according to Bayes' theorem upon the addition of new data. We then study the dependence of results on consistent or inconsistent datasets and present tools to assess the consistency of new data. Finally we estimate the relative size of the PDF uncertainty due to data uncertainties, and that due to the need to infer a functional form from a finite set of data.

1 The NNPDF approach to parton distributions

The determination of parton distributions (PDFs) and their uncertainties [1] poses a difficult problem because one is trying to determine the probability distribution for a set of functions. Given that this is necessarily done from a finite set of data it requires some assumptions: some of these, such as a certain degree of smoothness, may be physically motivated, but it is important to check that they do not bias the result and in particular that they do not destroy its statistical interpretation. The most common way of implementing these assumptions is to assume a functional form for the PDFs, each parametrized by a small number of parameters (typically between two and five) which are determined by fitting a suitable set of data. The NNPDF collaboration has developed an alternative approach [2–9] which tries to avoid the bias associated to this procedure.

The NNPDF approach is based on four main ingredients:

– *Monte Carlo by importance sampling.* NNPDF produces a Monte Carlo sampling of the probability density in the (function) space of PDFs. To adequately sample this space by simple binning would be simply impossible: for example assuming seven PDFs (the three light quarks and antiquarks and the gluon) sampled at ten points, binning the probability distribution in each direction with five bins one would end up with $5^{70} \sim 10^{49}$ bins. The problem is solved by importance sampling: most bins are empty and only those with data are relevant. Hence, one starts by constructing a set of data replicas, which reproduces the statistical features of the original data. It then turns out

*Now at PH Department, TH Unit, CERN, CH-1211 Geneva 23, Switzerland

that a sample of 1000 pseudo-data replicas is large enough to reproduce central values, uncertainty and correlations of the starting data to a few percent accuracy

– *Neural networks as universal unbiased interpolants.* Each of the underlying functions is parametrized with a feed-forward multilayer neural network. The architecture chosen corresponds to 37 free parameters for each of the seven PDFs. It can then be checked that results do not depend on the parametrization by verifying that they are unchanged if the size of the neural network is reduced.
– *Genetic Algorithms for neural network training.* The best fit is determined by using a genetic algorithm, and starting from a random initialization of parameters. This ensures that the presumably wide space of equivalent minima can be adequately explored.
– *Determination of the best fit by cross-validation.* Because the parametrization is very large, the best fit is not the minimum of the χ^2, which would correspond to fitting noise. The best fit is then found by dividing randomly data in two sets (training and validation) for each experiment, minimizing the χ^2 of the training set while monitoring the χ^2 of both sets. The best-fit is obtained when the χ^2 of the validation set starts increasing despite the fact that the χ^2 of the training set still decreases.

2 Statistical consistency

Our starting point is the NNPDF2.1 NLO [9] PDF set: we would like to test that it behaves in a statistically consistent way. For a start, in Table 1 we show the statistical estimators for this PDF fit: $\chi^2_{\rm tot}$ is the result of the comparison to data of the best-fit PDFs (defined as the average over the $N_{\rm rep} = 1000$ replicas of the Monte Carlo sample); $\left\langle \chi^{2(k)} \right\rangle$ is the average of the values obtained by comparing each PDF replica to the data, and $\langle E \rangle$ is the value of the same figure of merit, but obtained obtained comparing each PDF replica to the corresponding data replica. For the latter, the training and validation values are also shown. All figures of merit are computed using the full covariance matrix, with normalization uncertainties included using the so-called t_0 method of Ref. [10]; they are all normalized to the number of data points $N_{\rm dat}$. The fact that $\left\langle \chi^{2(k)} \right\rangle \sim 1$ while $\langle E \rangle \sim 2$, and also that $\chi^2_{\rm tot} < \left\langle \chi^{2(k)} \right\rangle$ are both consistent with the fact that the fit is "learning" an underlying law: the fitted PDFs are closer to the data than the data replicas (despite being fitted to the latter), and the best fit (obtained averaging replicas) is yet closer to the data than any of the individual replicas.

	Reference	Central Values	Average Fixed Partitions
$\chi^2_{\rm tot}$	1.16	1.14	1.15
$\langle E \rangle \pm \sigma_E$	2.24 ± 0.09	1.25 ± 0.11	1.24 ± 0.07
$\langle E_{\rm tr} \rangle \pm \sigma_{E_{\rm tr}}$	2.22 ± 0.11	1.25 ± 0.12	1.23 ± 0.07
$\langle E_{\rm val} \rangle \pm \sigma_{E_{\rm val}}$	2.28 ± 0.12	1.27 ± 0.11	1.26 ± 0.08
$\left\langle \chi^{2(k)} \right\rangle \pm \sigma_{\chi^2}$	1.25 ± 0.09	1.25 ± 0.11	1.24 ± 0.07

Table 1: Table of statistical estimators for NNPDF2.1 with $N_{\rm rep} = 1000$ replicas (first columns). The subsequent columns show the corresponding results, to be discussed in Sect. 4, for fits to central data and with fixed partitions, with $N_{\rm rep} = 100$ replicas each. All entries in the last column are obtained repeating the procedure for five random choices of fixed partition and averaging the final results. All values are normalized to the number of data points.

More detailed tests can be performed by looking at the distance between estimators extracted from PDF sets, defined as follows. Given a set of $N^{(i)}_{\rm rep}$ PDF replicas, the estimator for any quantity q computed from the PDFs (including the PDFs themselves) is the mean $\langle q \rangle_{(i)} = \frac{1}{N^{(i)}_{\rm rep}} \sum_{k=1}^{N^{(i)}_{\rm rep}} q_k$. The distance between two determinations of q from sets $q^{(1)}_i$, $q^{(2)}_i$ is then

$$d^2\left(\langle q^{(1)} \rangle, \langle q^{(2)} \rangle\right) = \frac{\left(\langle q^{(1)} \rangle_{(1)} - \langle q^{(2)} \rangle_{(2)}\right)^2}{\sigma^2_{(1)}[\langle q^{(1)} \rangle] + \sigma^2_{(2)}[\langle q^{(2)} \rangle]}, \tag{1}$$

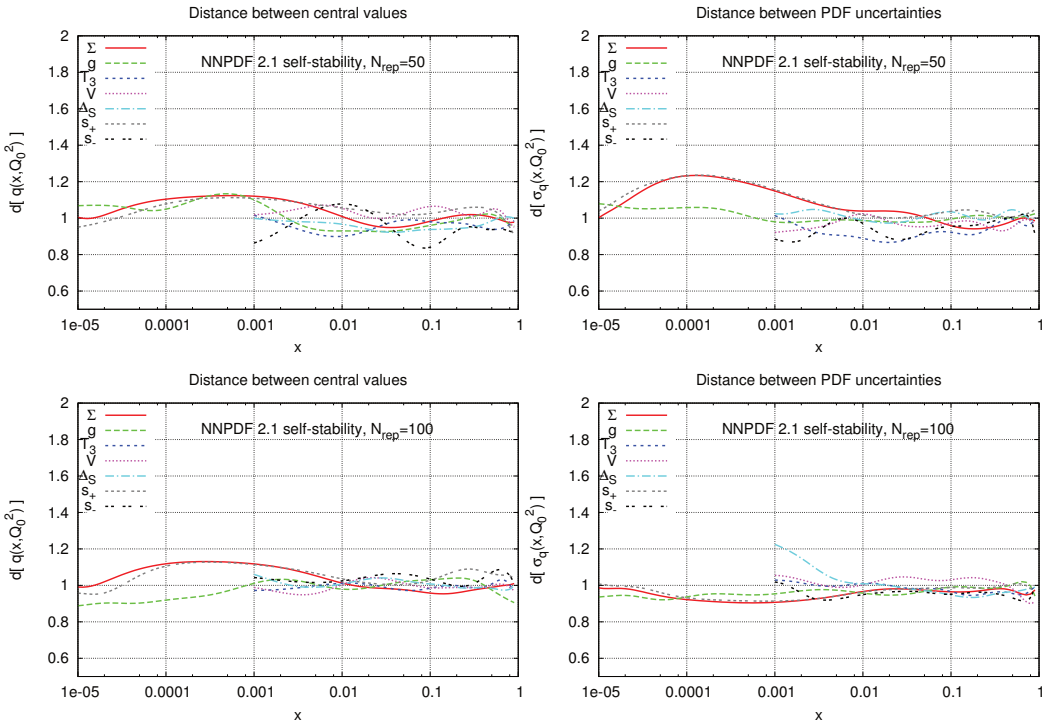


Fig. 1: Distances between central values and uncertainties of PDFs computed from two distinct sets of $N_{\rm rep} = 50$ (top) or $N_{\rm rep} = 100$ replicas (bottom).

with the variance of the mean given by

$$\sigma^2_{(i)}[\langle q^{(i)}\rangle] = \frac{1}{N^{(i)}_{\rm rep}}\sigma^2_{(i)}[q^{(i)}] \tag{2}$$

in terms of the variances $\sigma^2_{(i)}[q^{(i)}]$ of the variables $q^{(i)}$ (which a priori could come from two distinct probability distributions). The distance between uncertainties can be defined in a similar way. By construction, the probability distribution for the distance coincides with the χ^2 distribution with one degree of freedom, and thus it has mean $\langle d\rangle = 1$, and $d \lesssim 2.3$ at 90% confidence level.

An immediate use of the distance is to check that PDF sets computed from different sets of replicas are statistically equivalent (i.e. that $\langle q^{(k)}\rangle_{(i)}$ has the expected distribution). This is shown in Fig. 1 (top row): indeed distances fluctuate about $d \sim 1$. Furthermore, one can check (Fig. 1, bottom row) that the distance does not change as the number of replicas is varied: because of the explicit factor of $\frac{1}{N^{(i)}_{\rm rep}}$ in Eq. (2), this verifies that indeed the uncertainty of the mean decreases as $1/\sqrt{N^{(i)}_{\rm rep}}$ as $N^{(i)}_{\rm rep}$ is increased. Note that this means that the distance between two PDFs that barely overlap within error bands at 68% C.L. with $N_{\rm rep} = 100$ replicas is $\langle d\rangle \sim 7$ (because the distance is computed averaging results from subsets of $N_{\rm rep}/2 = 50$ replicas [7]).

Next, we check the independence of results of the parametrization. This is done by constructing a new set of PDF replicas with a different choice of architecture for neural networks, and checking that results are statistically equivalent. In Fig. 2 we show the distances between PDFs based on the default architecture 2–5–3–1. and PDFs based on the smaller 2–4–3–1 architecture. This corresponds to removing 6 free parameters from the parametrization of each PDF, i.e. removing of 42 free parameters overall. The similarity of Figs. 1 and 2 proves the stability of results. Note that, in order to make sure that the parametrization is indeed redundant, the larger architecture is used as a default.

Finally, we turn to our most detailed test of the statistical consistency of PDFs determined with

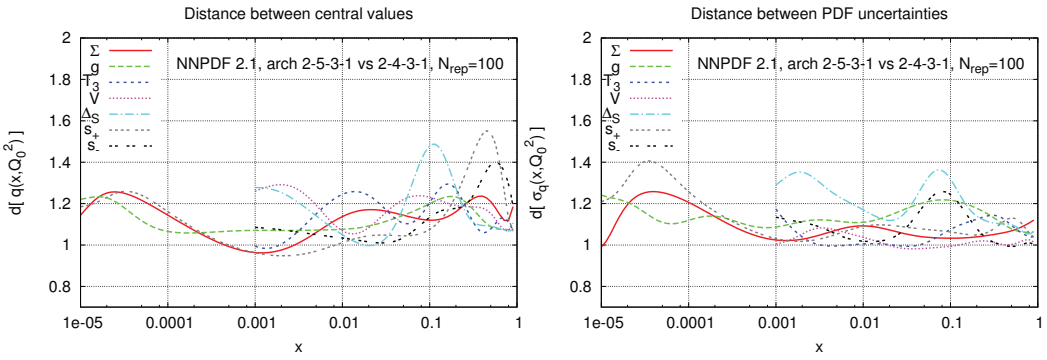


Fig. 2: Distances between PDFs with the default neural network architecture (2–5–3–1) and a reduced architecture (2–4–3–1).

the NNPDF methodology. Namely, we exploit the fact that given the probability distribution $\mathcal{P}_{\rm old}(f)$ for PDFs determined from a certain starting dataset, the effect of the inclusion of the information from new data can be determined using Bayes' theorem. It is then possible to compare the probability distribution $\mathcal{P}_{\rm new}(f)$ obtained in this way, with a determination of $\mathcal{P}_{\rm new}(f)$ found by simply performing a fit to an extended dataset including both the starting dataset and the new data. Statistical equivalence of the two determinations of $\mathcal{P}_{\rm new}(f)$ shows that the NNPDF methodology treats the information contained in the data in a consistent way. In fact, repeating this test for all of the data used for the fit, to the extent that for a large enough dataset results are independent of the prior assumption, would amount to a proof that the set of data and the set of PDFs determined from it contain the same information ("closure test"): indeed, such a Bayesian procedure was suggested in Ref. [11] as a way of arriving at a fully unbiased and self-consistent PDF determination.

We have performed such a test for an individual subset of data included in the NNPDF2.1 NLO PDF determination. The formalism to do so was developed in Ref. [12, 13], correcting a previous proposal of Ref. [14]. The way it works is the following: assume we want to include n new data $y = \{y_1, y_2, \cdots, y_n\}$ which had not been originally included in the determination of the initial probability density distribution. We view this data as a point y in an n-dimensional space, with uncertainties given as a $n \times n$ experimental covariance matrix. We update the probability density $\mathcal{P}_{\rm old}(f)$ using the conditional probability of the new data, which is proportional to the probability density of the χ^2 to the new data conditional on f:

$$\mathcal{P}(\chi^2|f) \propto (\chi^2(y,f))^{\frac{1}{2}(n-1)} e^{-\frac{1}{2}\chi^2(y,f)}, \tag{3}$$

where $y_i[f]$ is the value predicted for the data y_i using the PDF f. By Bayes' theorem then

$$\mathcal{P}_{\rm new}(f) = \mathcal{N}_\chi \mathcal{P}(\chi|f)\, \mathcal{P}_{\rm old}(f), \tag{4}$$

(with $\mathcal{N}_\chi$ an f–independent normalization factor).

Using Eq. (3) in Eq. (4) immediately implies that the inclusion of the new data can be viewed as a reweighting of the prior probability distribution $\mathcal{P}_{\rm old}(f)$. Namely, if the expectation value of some observable $\mathcal{O}$ with the distribution $\mathcal{P}_{\rm old}(f)$ is

$$\langle \mathcal{O} \rangle = \frac{1}{N} \sum_{k=1}^{N} \mathcal{O}[f_k]\,, \tag{5}$$

then, by Eq. (4), its expectation value according to $\mathcal{P}_{\rm old}(f)$ is

$$\langle \mathcal{O} \rangle_{\rm new} \;=\; \frac{1}{N} \sum_{k=1}^{N} \mathcal{N}_\chi \mathcal{P}(\chi|f_k) \mathcal{O}[f_k] = \frac{1}{N} \sum_{k=1}^{N} w_k\, \mathcal{O}[f_k]\,, \tag{6}$$

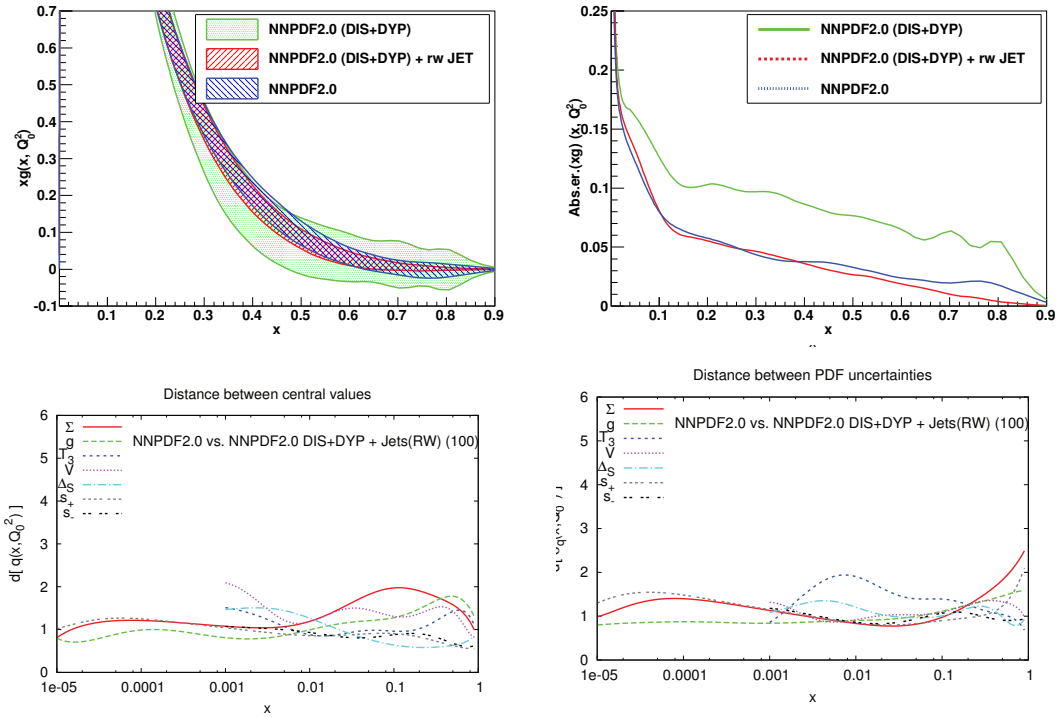


Fig. 3: Top: the gluon distribution (left) and its uncertainty (right) of the NNPDF2.0(DIS+DY) fit before and after reweighting with the inclusive jet data compared to the refitted gluon from NNPDF2.0. Bottom: distances between the refitted and reweighted results for central values (left) and uncertainties (right).

with

$$w_k = \frac{(\chi^2_k)^{\frac{1}{2}(n-1)} e^{-\frac{1}{2}\chi^2_k}}{\frac{1}{N}\sum_{k=1}^{N}(\chi^2_k)^{n/2-1} e^{-\frac{1}{2}\chi^2_k}} \,. \tag{7}$$

The weights w_k, when divided by $N = N_{\rm rep}$, are just the probabilities of the replicas f_k, given the χ^2 to the new data.

The comparison between the "reweighted" result Eq. (6-7) and the refitted one is shown in Fig. 3: it is apparent that the two procedures lead to the same result, except possibly at very large $x \gtrsim 0.7$ where the determination becomes unreliable because of the lack of experimental information. This is a very strong check that PDF uncertainties admit a *bona fide* statistical interpretation, and thus should not be viewed of theoretical uncertainties with unknown distribution. Note that because NNPDF results are delivered as a Monte Carlo sample, any feature of the distribution of results, such as confidence intervals or higher moments, can be determined explicitly.

3 Dataset dependence

One important feature of the NNPDF approach is that the same methodology can be used to determine PDFs from datasets of rather different size and nature: this, in particular, follows from the extreme redundancy of the parametrization, and the ensuing parametrization independence, explicitly checked in the previous section. In fact, NNPDF results are even stable upon the addition of new independently parametrized PDF, as seen in Ref. [5, 6] where light quark and gluon PDFs were found to be stable upon addition of an independent parametrization of strangeness. This is to be contrasted to the approach used by other groups, where a larger dataset requires the introduction of more parametrs. As a consequence, in the NNPDF approach, unlike in other approaches, the addition of new compatible data results in error reduction, as has been checked explicitly in benchmark studies [5, 15].

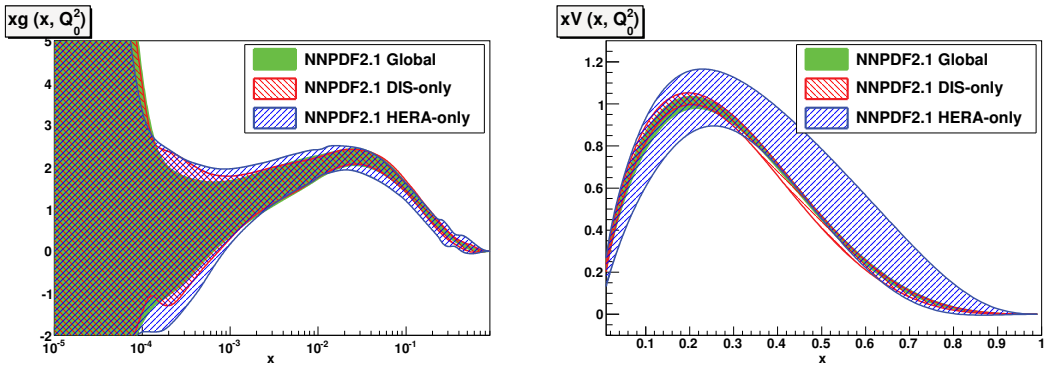


Fig. 4: Comparison of PDFs obtained to fits to different datasets: a global fit, a DIS–only fit and a HERA–only fit. The gluon (left) and total valence (right) PDFs are shown.

By comparing the results of fits to different datasets it is then possible to study the effect of individual data on PDFs and verify their consistency. For example, in Fig. 4, we compare the default NNPDF2.1 PDF set to PDFs obtained using only the DIS data or only the HERA DIS data from the global dataset. On the one hand, it is apparent from this comparison that these fits are mutually consistent; on the other hand it is clear that the HERA data determine well the small x gluon, the DIS data also determine well the total valence (mostly due to neutrino data), while the global dataset further improves the large x gluon. Detailed studies of this kind are performed in Refs. [8, 9] (see Ref. [1] for a general discussion of the expected impact of different data on PDFs).

A more detailed consistency check is performed by comparing fits in which a certain "new" dataset is added to different pre-existing datasets, and verifying that the impact of new data is independent of the choice of the dataset to which they are added, thereby also verifying the mutual consistency of the various data subsets involved. One such comparison (within the framework of the NNPDF2.0 [8] PDF set) is shown in Fig. 5 in which the effect of Drell-Yan data on the total valence and strange valence PDFs are compared when these data are added to a fit to DIS data only, or to a fit to DIS+jet data. More tests of this kind were shown in Ref. [1] and demonstrated equally good consistency.

The consistency of different data can be addressed quantitatively using the Bayesian reweighting technique of Ref. [12] summarized in Sect. 2. Namely, assume that the covariance matrix for a given dataset is rescaled by a common factor α, $\sigma_{ij} \to \alpha\sigma_{ij}$ so that for that experiment $\chi^2 \to \chi^2/\alpha^2$. It is then easy to show [12] that the probability density $\mathcal{P}(\alpha)$ for α given the data is

$$\mathcal{P}(\alpha) \propto \frac{1}{\alpha}\sum_{k=1}^{N} w_k(\alpha), \tag{8}$$

where $w_k(\alpha)$ are the weights Eq. (7) evaluated with the rescaled covariances. If $\mathcal{P}(\alpha)$ peaks close to one the new data are consistent, while if it peaks far above one, then it is likely that the errors in the data have been underestimated. As an example, we show in Fig. 6 $\mathcal{P}(\alpha)$ computed for two of the Tevatron D0 lepton asymmetry datasets analyzed in [12]. For muon data [16] $\mathcal{P}(\alpha)$ is peaked close to one, implying that this dataset is consistent with the other sets in the global fit. For muon data $\mathcal{P}(\alpha)$ is peaked far from one, suggesting that experimental uncertainties have been underestimated by about a factor two.

4 Functional and Data components of the PDF uncertainty

Because PDFs are functions determined from a finite set of data, one may expect that on top of the propagated uncertainty due to the uncertainty in the data there might be a further uncertainty due to existence (for sufficiently general parametrization) of many PDFs which give a fit of the same quality to the data.

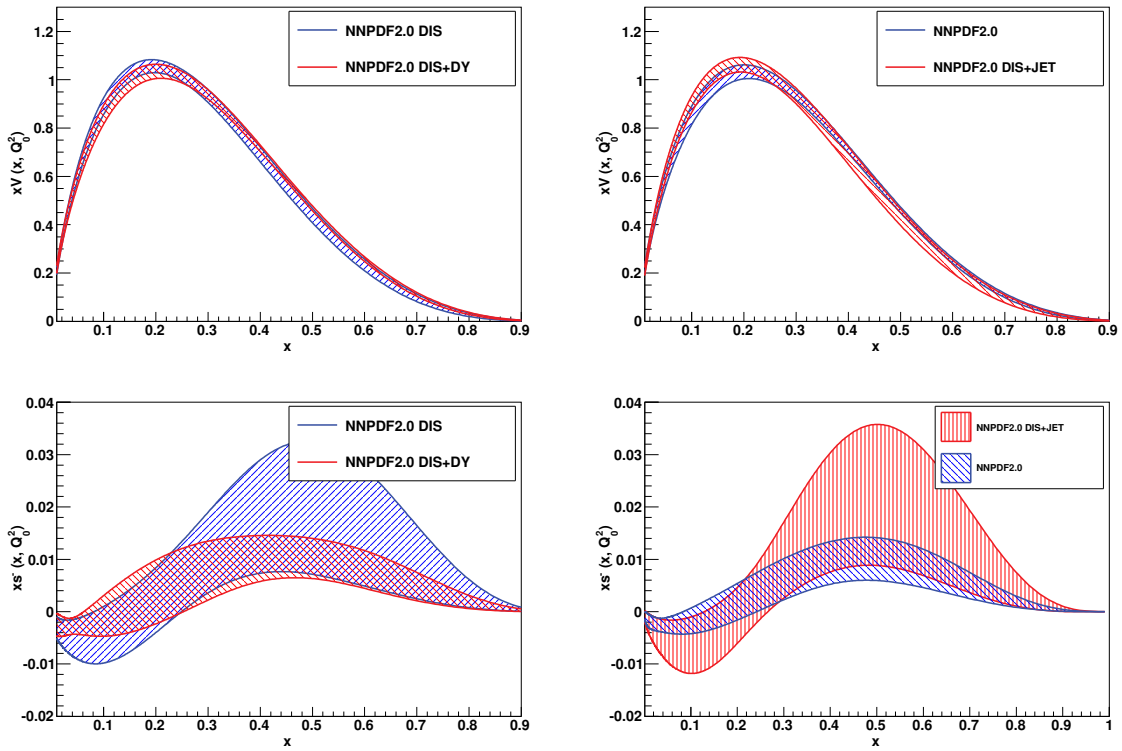


Fig. 5: Impact of the inclusion of Drell–Yan data in a fit with DIS data only (left), and in a fit with DIS and jet data (right). From top to bottom, total valence and $s - \bar{s}$ PDFs.

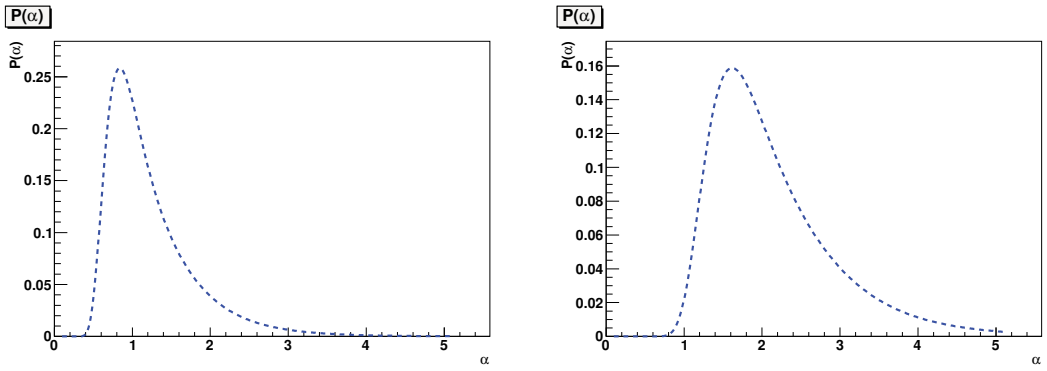


Fig. 6: The probability distribution $\mathcal{P}(\alpha)$ for the D0 lepton asymmetry data uncertainty to be underestimated by a factor α: muon data [16] (left) and electron data [17] (right).

For definiteness, we will call these different sources of uncertainty "data" and "functional" uncertainty respectively. If one were to accept infinitely coarse (e.g. fractal) PDF shapes the functional uncertainty would be infinite, but even if it is kept under control by some smoothness assumption it will generally still be nonzero. In fact, it was recently argued in Ref. [18] that the so called "tolerance" criterion [19] in PDF fits which make use of underlying functional forms with a relatively small number of parameters, and amounts to a rescaling of the $\Delta\chi^2$ range used to determine the one–σ range, mostly accounts for the fact that the choice of a fixed functional form with few parameters substantially underestimates the functional uncertainty.

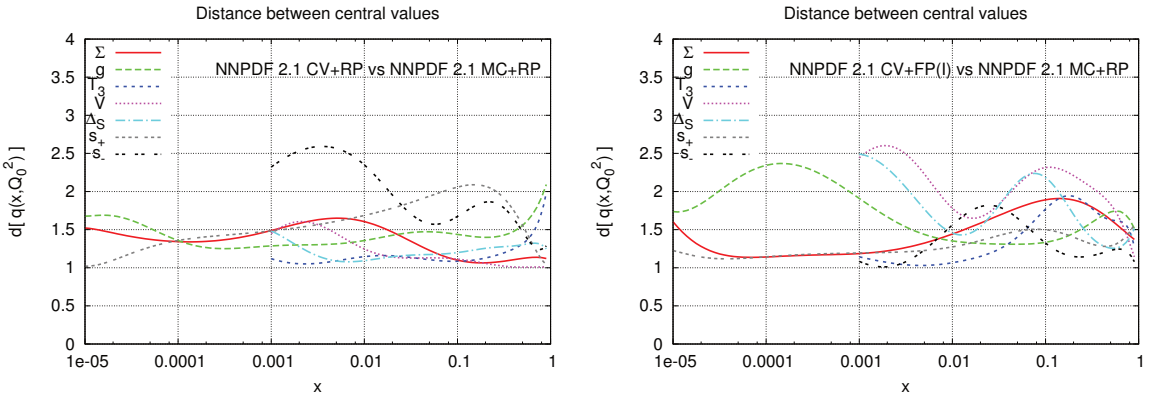


Fig. 7: Distances between central values of the reference PDFs and those fitted to different partitions of central values (left) or to a fixed partition of central values (right).

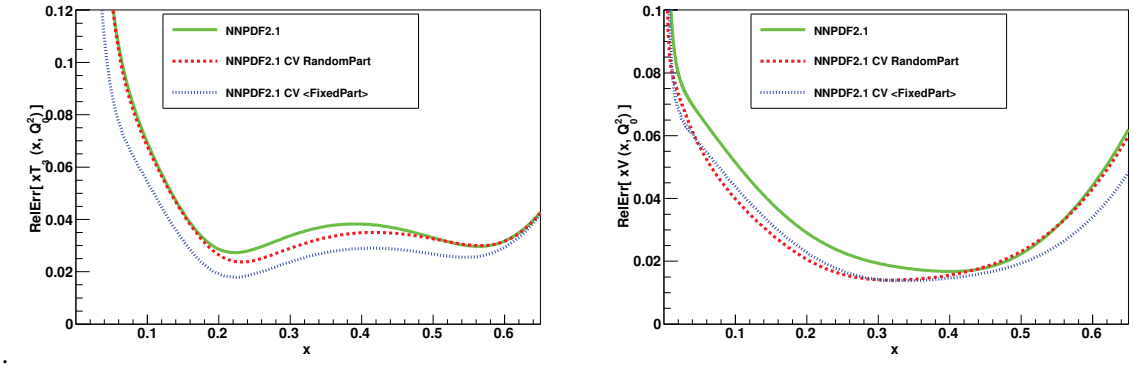


Fig. 8: Comparison between relative PDF uncertainties of the reference PDF set, a fit to varying partitions of central values and a fit to a fixed partition of central values for the isotriplet $T_3(x)$ (left) and the total valence $V(x)$ (right).

In the NNPDF approach, we can actually estimate the relative size of the data and functional uncertainty by constructing PDF replica sets based on a frozen set of data, as we now discuss. First, we switch off the pseudodata generation. Each PDF replica is then fitted to the same central data values (CV fit). However, each replica is still fitted to a different subset of data because for each replica the data are randomly divided in a training and validation set. Next, we also switch off the random partitioning of data for each replica, and we simply fit all PDF replicas to the same partition of central values (FP). In the latter case, the procedure is repeated five times, with different choices of the fixed partition in each case, in order to make sure that there is nothing special about the single partition that has been chosen in the first place, and results are the averaged.

Results for the statistical estimators for these fits are compared to those of the default case in Table 1. Furthermore, in Fig. 7 we display the distances between central values of PDFs obtained in the various cases, while in Fig. 8 we compare the relative percentage uncertainties for a couple representative PDFs. The central values appear to be very stable (distances of order one) and indeed the fit quality as measured by $\chi^2_{\rm tot}$ is essentially the same in all cases. When the pseudodata generation is switched off, $\langle E \rangle$, the average quality of the fit of each replica to the corresponding data replica now by construction coincides with $\left\langle \chi^{2(k)} \right\rangle$ (the same quantity but computed for central data). Interestingly, the value of $\left\langle \chi^{2(k)} \right\rangle$ in the reference and CV fit is identical: this confirms that the fitting methodology is very efficient in removing the extra fluctuation of the pseudodata about their central values induced by the pseudodata

Dataset	σ Data (%)	σ Ref. (%)	σ CV (%)	σ FP (%)
TOTAL	11.3	3.7	3.8	3.1 ± 0.2
NMC-pd	1.9	0.5	0.5	0.4 ± 0.03
NMC	5.0	1.6	1.6	1.4 ± 0.2
SLAC	4.4	1.7	1.7	1.4 ± 0.3
BCDMS	5.7	2.6	2.8	2.3 ± 0.3
HERAI-AV	2.5	1.3	1.3	1.1 ± 0.1
CHORUS	15.1	4.5	5.3	3.4 ± 0.3
FLH108	72.0	4.1	3.9	3.9 ± 0.5
NTVDMN	21.1	14.5	14.1	12.7 ± 1.6
ZEUS-H2	13.4	1.3	1.3	1.1 ± 0.2
ZEUSF2C	23.3	3.1	3.1	2.8 ± 0.2
H1F2C	17.3	2.9	2.9	2.6 ± 0.2
DYE605	22.3	8.1	7.0	6.1 ± 0.3
DYE886	20.1	9.1	8.3	8.2 ± 0.4
CDFWASY	6.0	4.5	3.4	3.1 ± 0.3
CDFZRAP	11.5	3.5	3.6	3.5 ± 0.5
D0ZRAP	10.2	2.8	3.0	2.9 ± 0.5
CDFR2KT	22.8	4.8	4.4	4.4 ± 0.2
D0R2CON	16.8	5.5	5.1	5.1 ± 0.2

Table 2: The average percentage uncertainty for each datasets for the reference, central value, and fixed partition PDF sets.

generation. It also suggests that the pseudodata generation is barely necessary. In fact, one could take this CV fit as a default: the fluctuations in central data are then just reproduced by bootstrap, by the process of choosing different partitions. Indeed, comparison of PDF uncertainties in the reference and CV case shows that they are very close and only moderately larger in the reference case, so that even if the pseudodata generation is viewed as a more conservative way of estimating uncertainties, in practice it is seen to have little effect.

However, the most striking result is given by the PDF uncertainties in the FP case: these uncertainties, though somewhat smaller, are still of the same order of magnitude as those of the the standard fit. This means that different replicas constructed by refitting exactly the same data over and over again still have a non-negligible spread and thus uncertainty. This is only possible because of the random nature of the fitting algorithm, and it shows that indeed there is a nontrivial space of almost equivalent minima. It should be noticed that indeed the fluctuation of $\left\langle \chi^{2(k)} \right\rangle$ for this replica set is significantly smaller than for the reference and CV sets, consistent with the hypothesis that one is now exploring a space of equivalent or almost equivalent minima.

A more quantitative insight on the relative size of various contributions to the uncertainties can be obtained by computing the average uncertainty on the prediction for the fitted observables obtained using each PDF set. These are shown, both for the global and individual dataset, in Table 4, where the starting data uncertainty is also shown for comparison. The uncertainties obtained fitting to central data or to pseudodata replicas are almost identical: as already noticed, one might as well fit to central data. Both are significantly smaller than the original data uncertainty, thereby showing that an underlying law has been learnt. The residual uncertainty in the FP case is still sizable. If one assumes that the uncertainty in the FP case is the functional uncertainty, while in the CV case it is the sum in quadrature of data and functional uncertainty, then one concludes that the functional uncertainty is rather more than half the total uncertainty.

5 Outlook

Having verified that PDFs determined with the NNPDF methodology are consistent with statistical expectations and free of parametrization bias, it is natural to think that some of the statistical tools discussed here, as well as more refined statistical tests, may be used to guide and validate further improvements.

Two aspects of the methodology may be may be amenable to improvement. The first has to do with the underlying functional form. At present, PDFs are parametrized as a neural network, multiplied by a preprocessing function of the form $x^\alpha(1-x)^\beta$. The exponents are then randomly varied in a reasonable range. The preprocessing speeds up the fitting of the neural network, and ensures that outside the data region the behaviour of the PDF does not fluctuate too wildly. This procedure is much more general and unbiased than that used in fits such as MSTW or CTEQ, in which the functional form also incorporates the same small- and large-x behaviour, but the exponents α and β are fitted (instead of being varied in a range around their best fit) and the residual number of parameters is smaller by more than one order of magnitude. But the preprocessing could still be a source of residual bias, so one should check whether results are stable upon completely different choices of preprocessing. The second has to do with the determination of the best fit. While cross-validation is quite efficient on average, it could still lead to some specific dataset being under- or overlearnt; it involves some arbitrariness, for instance in deciding the precise form of the stopping criteria; and it could lead to an excessively wide and thus sub-optimal space of minima. Hence alternative methods to determine the optimal fit should be explored.

Correspondingly, two sets of statistical investigations may be worth pursuing in order to guide and validate these improvements. On the one hand, it may be interesting to study the form of the probability distributions of PDF replicas: for instance, this could allow one to directly address the question of what in a conventional procedure is the $\Delta\chi^2$ range which correponds to a 68% confidence interval. On the other hand, it may be useful to investigate systematically the statistical impact of each dataset, with the aim of arriving at a full "closure test" — a proof that there is no information loss in extracting PDFs from data. These improvements may be useful and even necessary for precision phenomenology at the LHC.

Acknowledgments: We thank G. Cowan, L. Lyons and H. Prosper for discussions and encouragement. M.U. is supported by the Bundesministerium für Bildung and Forschung (BmBF) of the Federal Republic of Germany (project code 05H09PAE). This work was partly supported by the Spanish MEC FIS2007-60350 grant.

References

[1] Stefano Forte. Parton distributions at the dawn of the LHC. *Acta Phys.Polon.*, B41:2859–2920, 2010.

[2] Stefano Forte, Lluis Garrido, Jose I. Latorre, and Andrea Piccione. Neural network parametrization of deep-inelastic structure functions. *JHEP*, 05:062, 2002.

[3] Luigi Del Debbio, Stefano Forte, Jose I. Latorre, Andrea Piccione, and Joan Rojo. Unbiased determination of the proton structure function f2(p) with faithful uncertainty estimation. *JHEP*, 03:080, 2005.

[4] Luigi Del Debbio, Stefano Forte, Jose I. Latorre, Andrea Piccione, and Joan Rojo. Neural network determination of parton distributions: The nonsinglet case. *JHEP*, 03:039, 2007.

[5] Richard D. Ball et al. A determination of parton distributions with faithful uncertainty estimation. *Nucl. Phys.*, B809:1–63, 2009.

[6] Juan Rojo et al. Update on Neural Network Parton Distributions: NNPDF1.1. 2008.

[7] Richard D. Ball et al. Precision determination of electroweak parameters and the strange content of the proton from neutrino deep-inelastic scattering. *Nucl. Phys.*, B823:195–233, 2009.

[8] Richard D. Ball et al. A first unbiased global NLO determination of parton distributions and their uncertainties. *Nucl. Phys.*, B838:136–206, 2010.

[9] Richard D. Ball, Valerio Bertone, Francesco Cerutti, Luigi Del Debbio, Stefano Forte, et al. Impact of Heavy Quark Masses on Parton Distributions and LHC Phenomenology. *Nucl.Phys.*, B849:296–363, 2011.

[10] Richard D. Ball et al. Fitting Parton Distribution Data with Multiplicative Normalization Uncertainties. *JHEP*, 05:075, 2010.

[11] Walter T. Giele, Stephane A. Keller, and David A. Kosower. Parton distribution function uncertainties. 2001.

[12] Richard D. Ball et al. Reweighting NNPDFs: the W lepton asymmetry. *Nucl.Phys.*, B849:112–143, 2011.

[13] Richard D. Ball, Valerio Bertone, Francesco Cerutti, Luigi Del Debbio, Stefano Forte, et al. Reweighting and Unweighting of Parton Distributions and the LHC W lepton asymmetry data. 2011.

[14] Walter T. Giele and Stephane Keller. Implications of hadron collider observables on parton distribution function uncertainties. *Phys. Rev.*, D58:094023, 1998.

[15] M. Dittmar et al. Parton Distributions. 2009.

[16] V. M. Abazov et al. Measurement of the muon charge asymmetry from W boson decays. *Phys. Rev.*, D77:011106, 2008.

[17] V. M. Abazov et al. Measurement of the electron charge asymmetry in $p\bar{p} \to W + X \to e\nu + X$ events at $\sqrt{s}$ = 1.96-TeV. *Phys. Rev. Lett.*, 101:211801, 2008.

[18] Jon Pumplin. Parametrization dependence and Delta Chi-squared in parton distribution fitting. *Phys. Rev.*, D82:114020, 2010.

[19] J. Pumplin et al. New generation of parton distributions with uncertainties from global QCD analysis. *JHEP*, 07:012, 2002.

Nonlinear estimators for the detection of small and rare features

Sylvain Sardy
Section de Mathématiques, University of Geneva, Switzerland

Abstract
We illustrate in three settings (i.e., wavelet smoothing, total variation density estimation and wavelet-based inverse problem) the need of nonlinear estimators to retrieve small or rare features hidden in data. Such nonlinear nonparametric methods could be specifically developed for inverse problems at CERN.

1 Introduction

Consider the regression setting

$$Y_n = \mu_n + \epsilon_n, \quad n = 1, \dots, N, \tag{1}$$

where Y_n are measurements of the signal μ_n with noise ϵ_n. In the following, we write vectors in bold, e.g., $\mathbf{Y} = (Y_1, \dots, Y_N)$. Suppose $\hat{\boldsymbol{\mu}}_\lambda$ is an estimator of $\boldsymbol{\mu} = (\mu_1, \dots, \mu_N)$ indexed by a regularization parameter λ (which use will become clear). To measure the quality of this estimator, the risk of $\hat{\boldsymbol{\mu}}_\lambda$ is defined as $\mathrm{R}(\lambda) = \mathrm{E}[(\hat{\boldsymbol{\mu}}_\lambda - \boldsymbol{\mu})^2]$, where E stands for expectation. In practice the risk is unknown, but can be estimated from the data. Importantly the risk has a bias-variance decomposition

$$\mathrm{R}(\lambda) = \mathrm{bias}^2(\lambda) + \mathrm{Var}(\lambda).$$

In some settings, estimators like the maximum likelihood estimator (MLE) or least squares (LS) have no bias, but have a very high variance; conversely, other estimators have no variance but a high bias. The goal of regularization is to propose appropriate ways to introduce bias and to control it well with a good selection of the regularization parameter λ.

We distinguish two regression problems to illustrate regularization.

1.1 Nonparametric estimation

If μ is a univariate function (or an image) observed at points x_n (in which case $\mu_n = \mu(x_n)$) then one can try to recover μ from the data Y_n without making any strong parametric assumption on μ. Hence the linear smoothing splines estimator [1] assumes μ belongs to a Sobolev space which only imposes a smoothness class. Such an estimator performs well to estimate smooth functions.

Recently Waveshrink [2] provides a nonlinear estimator capable of detecting small and sharp features such as peaks, discontinuities or small bursts. They assume the underlying signal μ expands linearly on N orthonormal wavelets, with corresponding wavelet coefficients $\boldsymbol{\alpha}$, which form a basis of Besov spaces which include Sobolev spaces as particular cases. One can extract an orthonormal regression matrix W of dimension $N \times N$ from this representation such that (1) becomes in vector and matrix notation

$$\boldsymbol{Y} = W\boldsymbol{\alpha} + \boldsymbol{\epsilon}.$$

Assuming Gaussian independent noise $\boldsymbol{\epsilon}$, the maximum likelihood estimate is obtained by applying the discrete wavelet transform (DWT) to the data: $\hat{\boldsymbol{\alpha}}^{\mathrm{MLE}} = W^{\mathrm{T}}\boldsymbol{Y}$. It has no bias but high variance. Importantly, projected on a wavelet basis, most functions μ have a sparse wavelet representation (i.e., most entries of $\boldsymbol{\alpha}$ are zero). So [2] propose to regularize the MLE by applying componentwise a nonlinear function that enforces sparsity, for instance, the so-called soft-thresholding function

$$\hat{\boldsymbol{\alpha}}_\lambda = \left\{1 - \frac{\lambda}{|\hat{\alpha}_n^{\mathrm{MLE}}|}\right\}_+ \hat{\alpha}_n^{\mathrm{MLE}}, \quad n = 1, \dots, N, \tag{2}$$

where $\{x\}_+ = 0$ if x is negative. The smoothing parameter λ controls the bias-variance trade-off: if $|\hat{\alpha}_n^{\mathrm{MLE}}|$ is abnormally large with respect to λ, it will be kept as a significant coefficient; otherwise, it will be seen as noise and set to zero by the thresholding function. Then the estimator of the underlying signal is $\hat{\boldsymbol{\mu}}_\lambda = W\hat{\boldsymbol{\alpha}}_\lambda$. Note that for $\lambda = 0$, no thresholding/regularization is performed and we get back the MLE. Reference [3] derives near minimax results for Waveshrink.

1.2 Parametric estimation

Here we also assume that covariates $(x_1, \ldots, x_P)_n$ are observed along with Y_n for $n = 1, \ldots, N$. Linear parametric regression assumes $\mu_n = \sum_{p=1}^P \alpha_p x_{np}$, which in vector form is

$$\mathbf{Y} = X\boldsymbol{\alpha} + \boldsymbol{\epsilon}.$$

The least squares estimate may have a high variance if the matrix X is badly conditioned, so the linear ridge regression estimator [4] adds a quadratic penalty to control the bias, and estimate the coefficients by solving

$$\min_{\boldsymbol{\alpha}} \|\mathbf{Y} - X\boldsymbol{\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_\eta^\eta \tag{3}$$

for $\eta = 2$, where $\|\cdot\|_\eta$ stands for the ℓ_η norm, e.g., $\|\boldsymbol{\alpha}\|_2 = \sqrt{\sum_{p=1}^P \alpha_p^2}$ and $\|\boldsymbol{\alpha}\|_1 = \sum_{p=1}^P |\alpha_p|$. Recently [5] developed the nonlinear lasso estimator for $\eta = 1$; interestingly, lasso performs model selection in the sense that the solution to (3) is a sparse vector (the larger λ the more sparse the estimated vector). Moreover, when the matrix X is orthonormal (e.g., a wavelet matrix), then (3) has a closed form solution via the soft thresholding function (2). More recently, adaptive lasso [6] is a variation of lasso that is oracle in the sense that it selects the right model with a high probability and is root-N consistent for the non-zero coefficients.

2 Nonlinear estimation to detect sharp and rare features

Based on Sections 1.1 and 1.2, we address estimation and detection of rare and sharp features in more complex settings that may be of interest towards solving inverse problems in particle physics at CERN.

2.1 Wavelet smoothing from several captors

Gravitational wave bursts are rare events expected to be produced by energetic cosmic phenomena such as the collapse of a supernova [7]. The signal-to-noise ratio is believed to be low, so that only the joint information recorded by Q captors at a high frequency of 5MHz may help prove the existence of such wave bursts. The noise is colored and possibly non-Gaussian. A good model for these data is (1) for Q signals and for $n = t$ (for time), namely

$$Y_t^{(q)} = \mu_t^{(q)} + \epsilon_t^{(q)}, \quad t = 1, \ldots, T,\ q = 1 \ldots, Q \tag{4}$$

where the noises $\epsilon^{(q)}$ and $\epsilon^{(q')}$ are independent between captors $q \neq q'$. Importantly, most of the time the underlying signal $\mu^{(q)}(t) = 0$ for all q, but, if $\mu^{(q)}(t) \neq 0$ for a given time t and captor q, then $\mu^{(q')}(t) \neq 0$ for all other captors q' at the same time t. Moreover when a wave burst occurs, we may not have $\mu^{(q)}(t) = \mu^{(q')}(t)$, but only a proportionality constant relates them, because the incoming wave burst may not hit the captor with the same angle, or the captors may not have the same sensitivity.

Assuming a wavelet representation of each $\mu^{(q)}$ for captors $q = 1, \ldots, Q$, one can estimate the wavelet coefficients from the data by

$$\hat{\boldsymbol{\alpha}}^{(q)} = W^{\mathrm{T}}\mathbf{Y}^{(q)},$$

where the DWT also has a decorrelating property [8]. Letting $\hat{\boldsymbol{\alpha}}_n = (\hat{\alpha}_n^{(1)}, \ldots, \hat{\alpha}_n^{(Q)})$ be the block of Q wavelet coefficients corresponding to the nth wavelet used in the linear expansion of the Q underlying

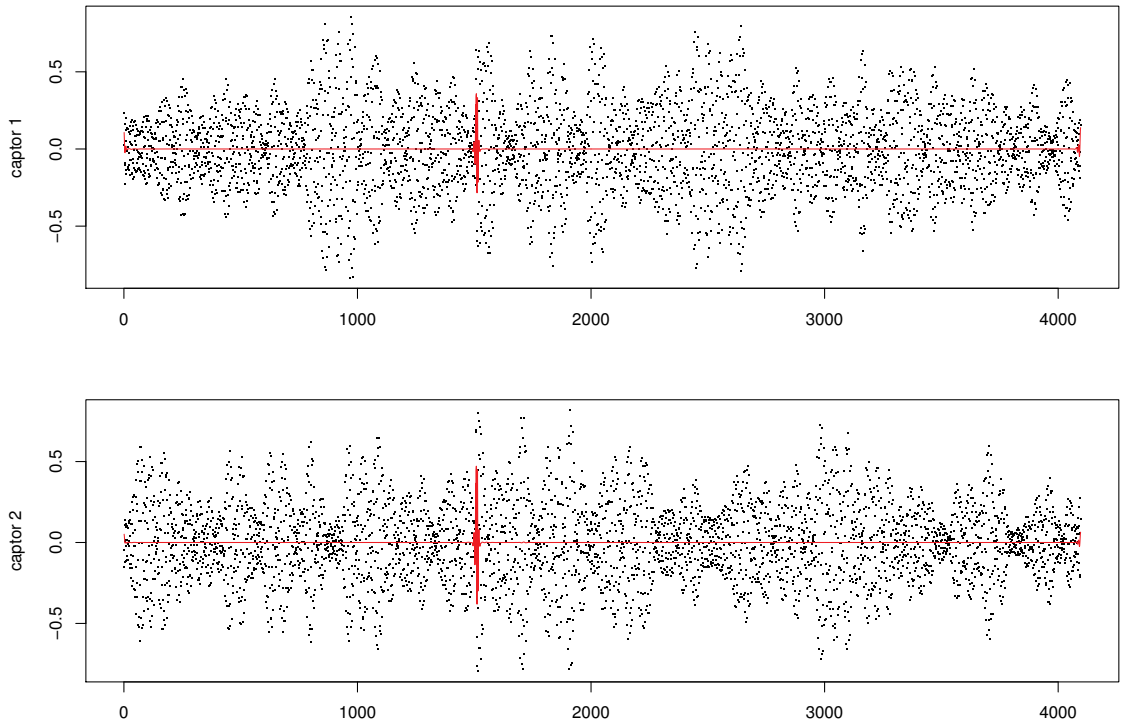


Fig. 1: Gravitational wave burst detection: concomitant and independent noisy signals recorded on $Q = 2$ captors (black dots), and the block thresholded estimator (red line).

signals μ_q, $q = 1 \ldots, Q$, we want to decrease the variance of this block by thresholding it towards zero. If this vector is abnormally large, then we will believe it contains an important feature; otherwise, all of its components will be set to zero concomitantly. To enforce this concomitant sparsity while preserving abnormally large blocks, we generalize (2) by applying the following block soft-thresholding function:

$$\hat{\boldsymbol{\alpha}}_{n,\lambda} = \left\{1 - \frac{\lambda}{\|\hat{\boldsymbol{\alpha}}_n\|_2}\right\}_+ \hat{\boldsymbol{\alpha}}_n, \quad n = 1, \ldots, N, \tag{5}$$

where $\|\boldsymbol{\alpha}\|_2 = \sqrt{\alpha_1^2 + \ldots + \alpha_Q^2}$. We can then estimate the underlying signal on each captor using the inverse DWT. Figure 1 shows typical time series recorded by the two captors (dots) in which an artificial signal resembling a wave burst has been "injected;" the red curve is the estimate based on (5). Figure 2 zooms around the time of the injection. We observe that the artificial wave burst is well detected and that the noise is well removed otherwise, although the signal to noise ratio was small.

2.2 Density estimation

Density estimation is an old problem in statistics [9–11]. Suppose a sample of size N from a density function f has been collected, and let $x_1, \ldots, x_N$ be the corresponding order statistics. The goal is to estimate f from the data x_n, $n = 1, \ldots, N$. The histogram is the commonly used nonparametric estimator, but is unstable to the choice of the binwidth and the left point. Moreover the histogram can show too many modes/bumps, as illustrated on the top graph of Figure 3. Taut string [12] is a more recent nonparametric estimator that controls the numbers of modes and that has some connection with the total variation estimator [13]. That latter estimator regularizes the likelihood with an ℓ_1-based penalty

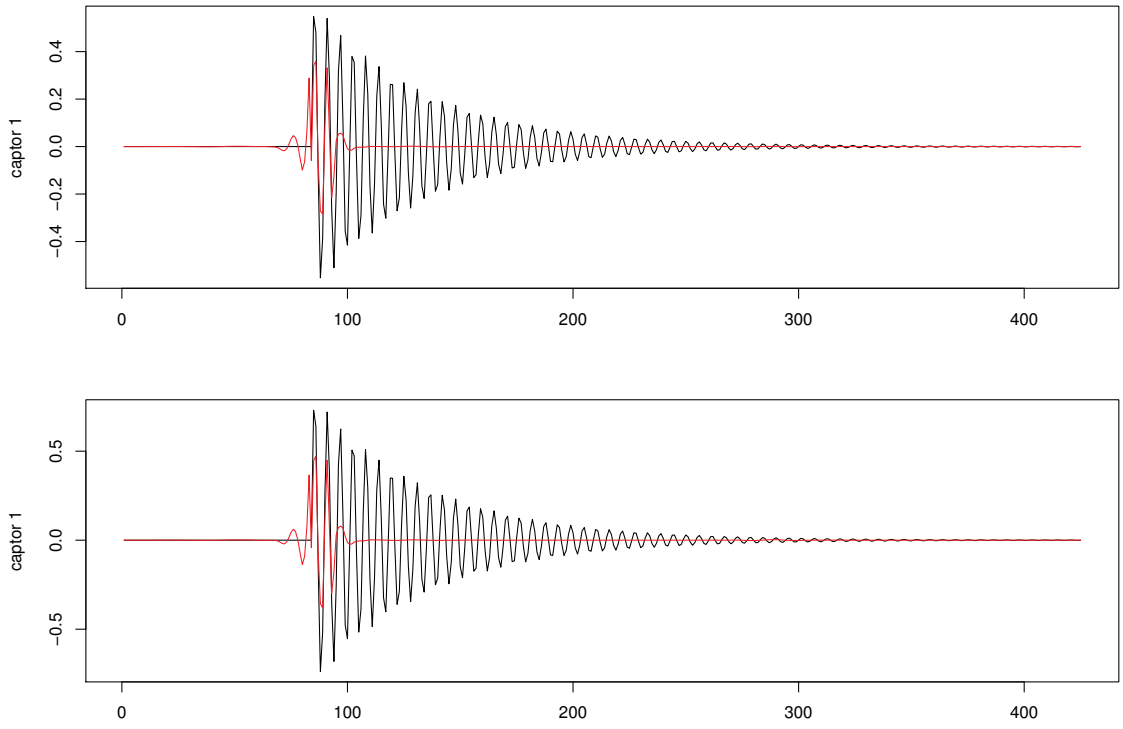


Fig. 2: Gravitational wave burst detection: zoom around the time of an "injection" (black line), and the block thresholded estimator (red line).

by solving

$$\min_{\mathbf{f}\in\mathbb{R}^N} -\sum_{i=1}^{N} \log f_i + \lambda \sum_{i=2}^{N} |f_n - f_{n-1}|, \quad \text{s.t.} \quad \mathbf{a}^T\mathbf{f} = 1, \tag{6}$$

where the equality constraint forces the estimated function to integrate to one (to be a density), and where $a_1 = (x_2 - x_1)/2$, $a_N = (x_N - x_{N-1})/2$ and $a_n = (x_{n+1} - x_{n-1})/2$ for $n = 2, \ldots, N-1$. Here λ controls the smoothness of the estimate. The total variation estimate (middle graph) of Figure 3 illustrates its ability to estimate the underlying density without unnecessary bumps.

2.3 Inverse problem

Likewise in the inverse problem, one can retrieve bumps from data quite well by developing an appropriate nonlinear estimator. Suppose the sample $Y_1, \ldots, Y_N$ measures with noise an unknown function f through a known linear operator $\mathcal{K}$ at known locations $\boldsymbol{t} = (t_1, \ldots, t_N)$ in Ω according to

$$Y_n = \mathcal{K}f(t_n) + \epsilon_n, \quad n = 1, \ldots, N. \tag{7}$$

We propose to expand f linearly on a wavelet basis W and regularize the least squares problem with lasso, i.e., (3) with $\eta = 1$ to enforce a sparse wavelet estimation. Hence we solve

$$\min_{\boldsymbol{\alpha}} \|\mathbf{Y} - KW\boldsymbol{\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1,$$

where the smoothing parameter λ is chosen to minimize an estimate of the risk. Figure 4 illustrates the power of this nonlinear estimator (red curve) to retrieve peaks (the green curve is the underlying function to retrieve) from a blurred and noisy signal (black line). Some peaks that had disappeared with the blurring can be retrieved surprisingly well.

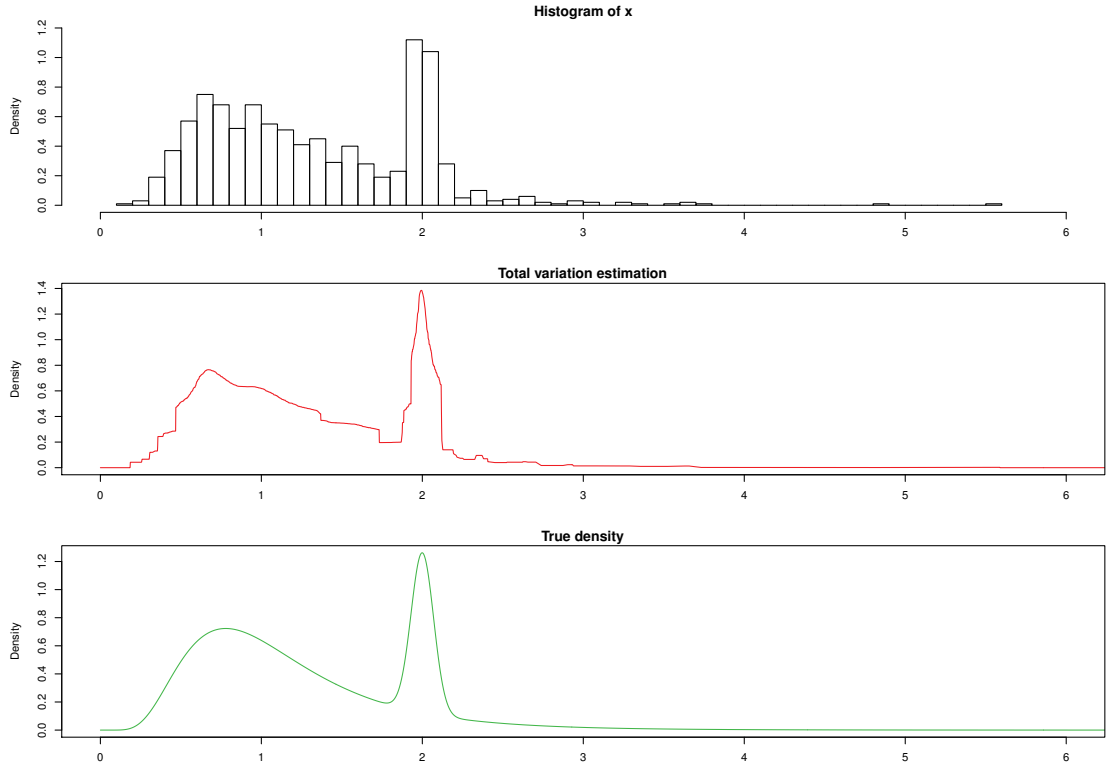


Fig. 3: Looking for bumps in a density. Histogram (top), total variation estimate (middle), true density (bottom).

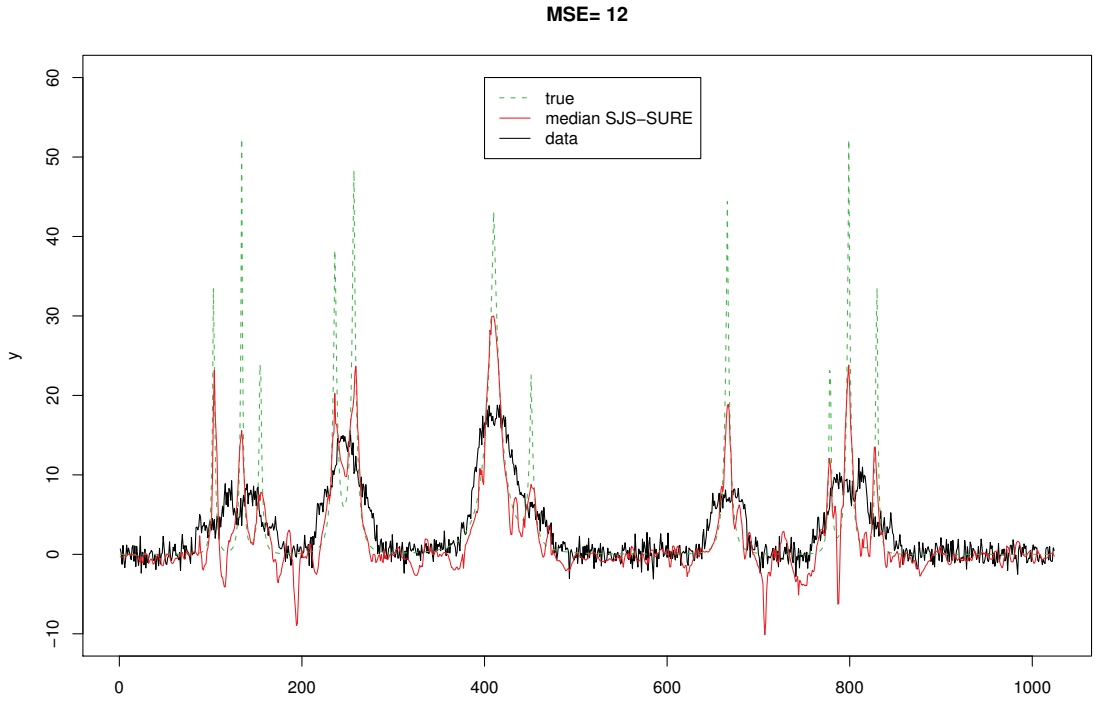


Fig. 4: Looking for bumps in a blurring inverse problem: true underlying function (green), data (black) and nonlinear estimate (red).

3 Conclusion

The three settings considered (i.e., regression, density estimation and inverse problem) illustrate the ability of nonlinear nonparametric estimators to retrieve sharp and rare features from data. Developing such estimators for the specificities of inverse problems encountered at CERN is challenging and will reveal whether these estimators can enhance discoveries on CERN real applications.

4 Acknowledgements

We thank David Hand for his comments on a first draft, and Stefano Foffa, Roberto Terenzi and the ROG group for providing a sample of the astrophysics data in Figure 1.

References

[1] Grace Wahba, *Spline Models for Observational Data*, SIAM, 1990.

[2] David Donoho and Iain Johnstone, Ideal Spatial Adaptation via Wavelet Shrinkage, Biometrika, p. 425–455, 1994.

[3] David Donoho, Iain Johnstone, Gérard Kerkyacharian and Dominique Picard, Wavelet Shrinkage: Asymptopia? (with discussion), Journal of the Royal Statistical Society, Series B: Methodological, p. 301–369, 1995.

[4] Arthur Hoerl and Robert Kennard, Ridge regression: biased estimation for nonorthogonal problems, Technometrics, p. 55–67, 1970.

[5] Robert Tibshirani, Regression Shrinkage and Selection Via the Lasso, Journal of the Royal Statistical Society, Series B: Methodological, p. 267–288, 1996.

[6] Hui Zou, The Adaptive LASSO and Its Oracle Properties, Journal of the American Statistical Association, vol. 101, p. 1418–1429, 2006.

[7] Sergey Klimenko and Guenakh Mitselmakher, A wavelet method for detection of gravitational wave bursts, Classical and Quantum Gravity, p. 1819–1830, 2004.

[8] Iain Johnstone and Bernard Silverman, Wavelet Threshold Estimators for Data with Correlated Noise, Journal of the Royal Statistical Society, Series B: Methodological, p. 319–351, 1997.

[9] David Scott, *Multivariate Density Estimation: Theory, Practice, and Visualization*, Wiley, 1992.

[10] Bernard Silverman, *Density Estimation for Statistics and Data Analysis*, Chapman & Hall, 1986.

[11] Jeffrey Simonoff, *Smoothing Methods in Statistics*, Springer-Verlag, 1996.

[12] Laurie Davies and Arne Kovac, Densities, spectral densities and modality, The Annals of Statistics, p. 1093–1136, 2004.

[13] Sylvain Sardy and Paul Tseng, Density estimation by total variation penalized likelihood driven by the sparsity ℓ_1 information criterion, Scandinavian Journal of Statistics, p. 321–337, 2010.

Signal discovery in sparse spectra: a Bayesian analysis

F. Beaujean[1]*, A. Caldwell*[1*]*, D. Kollár*[2]*, K. Kröninger*[3]*, S. Pashapour*[3]
1 Max-Planck-Institut für Physik, München, Germany
2 CERN, Geneva, Switzerland
3 II Physikalisches Institut, Universität Göttingen, Germany
* Corresponding author

Abstract
A Bayesian analysis of the probability of a signal in the presence of background is described. As an example, the method was used to calculate the sensitivity of the GERDA experiment to neutrinoless double beta decay. In addition, we discuss the use of consensus priors, the look-elsewhere-effect in Bayesian analysis and other topics.

1 Introduction

Scientific knowledge, i.e., justified belief, comes from inductive reasoning. Experimental tests allow us to build our justification for believing in particular models. In the context of the models, frequency distributions can be produced and probabilities of different outcomes calculated. However, it is impossible to make a statement on the truth of the model without considering all possible models which could give similar results, and assigning prior beliefs to the models. Frequentist approaches avoid using priors and therefore in principle do not allow statements on how strongly we should believe in a particular model. Statements of belief in a model become maximally subjective - each interpreter of the data is advised to reach their own conclusions on what to believe [1]. In contrast, in the Bayesian approach the prior beliefs are explicitly stated so that posterior beliefs can be evaluated. While the posterior beliefs are also subjective, the reasoning which led to the conclusion is made clear. Given that the goal is to make a statement on how strongly we believe our models, the Bayesian approach seems to us appropriate.

2 Signal discovery in an event counting setting

Imagine we have a collections of events where we have measured some physical quantity x which can take on a continuous range of values. We assume that we have a background model, with background contribution B, for the distribution of the values of x, possibly with nuisance parameters involved, and we can predict the distribution of x values for some new physics, which could depend on parameters of interest (e.g., for a Gaussian distribution for signal events, we have some position μ, width parameter σ, and amplitude S). To proceed, we need our prior belief that the background model accounts completely for the observations, $P_0(H_1)$, and the prior belief that there could be new physics contributing to the observations, $P_0(H_2) = 1 - P_0(H_1)$. For the models, we also need the prior beliefs in the possible values of the parameters: e.g., for the 'new physics' model $P_0(\mu, S, \sigma|H_2)$. We then group the observations $\{x\}$ in intervals Δx_i and compare the predictions with the observations. Using D to represent the data, we have for the posterior belief in H_2:

$$P(H_2|D) = \frac{P(D|H_2)P_0(H_2)}{P(D|H_2)P_0(H_2) + P(D|H_1)P_0(H_1)} \tag{1}$$

where

$$\begin{aligned} P(D|H_2) &= \int P(D|\mu, S, \sigma, B)P_0(\mu, S, \sigma|H_2)P_0(B)d\mu dS d\sigma dB \\ P(D|H_1) &= \int P(D|B)P_0(B)dB \end{aligned}$$

and

$$P(D|\mu, S, \sigma, B) = \prod_i \frac{e^{-\nu_i}\nu_i^{n_i}}{n_i!} \quad (2)$$

$$P(D|B) = \prod_i \frac{e^{-\lambda_i}\lambda_i^{n_i}}{n_i!} \quad (3)$$

and

$$\lambda_i(B) = \int_{\Delta x_i} f_B(x|B)dx \quad (4)$$

$$\nu_i(\mu, A, \sigma, B) = \lambda_i(B) + \int_{\Delta x_i} f_S(x|\mu, A, \sigma)dx \quad (5)$$

with n_i the observed number of events in bin i, λ_i the expectation for bin i for the background model, and ν_i the expectation including the new physics signal given the parameter values.

3 Sensitivity analysis for GERDA

This analysis method was used to estimate the sensitivity of the GERDA experiment to neutrinoless double beta decay [2]. In the GERDA case, the location and shape of the signal are known (i.e., μ and σ above are fixed), so that the only physics parameter is the expectation for the number of signal events.

Given the lack of theoretical consensus on the Majorana nature of neutrinos and the cloudy experimental picture, the prior probabilities for H_1 and H_2 were chosen to be equal, i.e.

$$P_0(H_1) = 0.5, \quad (6)$$

$$P_0(H_2) = 0.5\,. \quad (7)$$

The prior probability for the number of expected signal events, assuming H_2, was taken flat up to a maximum value, S_{max}, consistent with existing limits[1]. It should be noted that the prior probability for H_1 depends on the maximum allowed signal rate. $S_{\max}$ was chosen so that the probability for the hypothesis H_1 is 50 %, which is a reasonable assumption. The effect of choosing a different prior for the number of signal events was studied in Ref. [2].

The overall background contribution B was chosen to be Gaussian with mean value $\mu_{\rm B} = B_0$ and width $\sigma_{\rm B} = B_0/2$. The prior probabilities for the expected signal and background contributions were taken as

$$P_0(\mu, S, \sigma|H_2) = P_0(S|H_2) = \frac{1}{S_{\max}},\ 0 \leq S \leq S_{\max},\ p_0(S) = 0 \text{ otherwise}, \quad (8)$$

$$P_0(B) = \frac{e^{-\frac{(B-\mu_{\rm B})^2}{2\sigma_{\rm B}^2}}}{\int_0^\infty e^{-\frac{(B-\mu_{\rm B})^2}{2\sigma_{\rm B}^2}} dB},\ B \geq 0,\ P_0(B) = 0 \text{ otherwise}\,. \quad (9)$$

Ensemble tests were then used to evaluate the sensitivity of the experiment to both signal discovery and probability limits on the half-life $T_{1/2}$ for neutrinoless double beta decay. An example data set as well as the resulting discovery sensitivity are given in Fig. 1.

[1] $S_{\max}$ was calculated assuming a half-life of $T_{1/2} = 0.5 \cdot 10^{25}$ years.

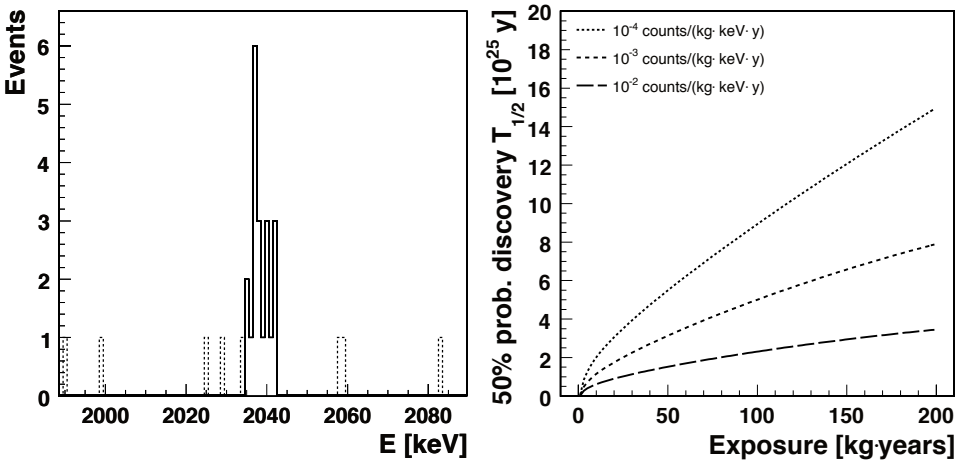


Fig. 1: Left: an example data set generated with $T_{1/2} = 2 \cdot 10^{25}$ yr, a background index of $1 \cdot 10^{-3}$/(keV·kg·yr) and an exposure of 100 kg·yr. Right: the curves indicate the half-life where an experiment would have a 50 % chance of claiming a discovery as a function of exposure, and for different background indices. Discovery was defined in [2] as $P(H_1|D) < 0.0001$.

4 Error bars

No error bars are shown in Fig. 1, since error bars on distributions of observed numbers of events are at best misleading. There is certainly no uncertainty on the number of observed events. The only uncertainty comes when the observed number of events is used to estimate the mean of the underlying Poisson distribution. There are different ways in which this mean can be extracted, and placing the estimate for the mean at the number of observed events is in any case not always the best choice. The second problem arises with the size of the error bar. This is routinely plotted as the square root of the number of events, taking the Poisson result that the variance is equal to the mean. However, this definition does not lead to an error bar which contains 68 % probability. The probability range covered varies dramatically for small numbers of events and is asymmetric around the point. This leads to great confusion when non-experts analyze data/model agreement 'by eye'. We would strongly favor ending the practice of putting error bars on the number of observed events. It is better to give no extra information than to give misleading information.

5 The Look-Elsewhere Effect (LEE) in Bayesian Analysis

There is no look-elsewhere effect in the GERDA example since the location of the signal is known. In general, the LEE is suppressed in Bayesian analysis, since a penalty is built into the prior for allowing a signal to appear in different places during a search. This is demonstrated here for a simple example of searching for a signal in a 1-D distribution. Assume that the resolution (width of the peak, σ) for the potential signal is known as well as the amplitude, but we allow a search with the location of the signal free. Define H_1 as the null hypothesis - only known backgrounds are present. H_2 is the hypothesis that in addition to the known backgrounds, there is also a signal. In this case, using μ as the location of the new physics signal, we have

$$P(H_2|D) = \frac{\int P(D|H_2,\mu)P_0(H_2,\mu)d\mu}{\int P(D|H_2,\mu)P_0(H_2,\mu)d\mu + P(D|H_1)P_0(H_1)} \quad . \tag{10}$$

where D represents the data and we assume that the null hypothesis has no free parameters.

Taking a simple example,

$$P_0(H_2,\mu) \quad = \quad P_0(H_2)P_0(\mu|H_2)$$

$$P_0(H_2) \quad = \quad P_0(H_1) = 1/2$$

our equation (10) becomes

$$P(H_2|D) = \frac{\int P(D|H_2,\mu)P_0(\mu)d\mu}{\int P(D|H_2,\mu)P_0(\mu)d\mu + P(D|H_1)} \quad . \tag{11}$$

Now assume we can use a flat prior for μ, given by

$$P_0(\mu) \quad = \quad \frac{1}{L_\mu}$$

where L_μ is the range over which the parameter can vary. Our equation further simplifies to

$$P(H_2|D) = \frac{\int P(D|H_2,\mu)d\mu}{\int P(D|H_2,\mu)d\mu + L_\mu P(D|H_1)} \tag{12}$$

The integral can be written as

$$\int P(D|H_2,\mu)d\mu = P(D|H_2,\mu^*)\delta_\mu$$

where μ^* is the parameter value which maximizes the probability of the data, and δ_μ is an effective width of the distribution $P(D|H_2,\mu)$. We expect $\delta_\mu \approx \sqrt{2\pi}\sigma$. Using these results, we find

$$P(H_2|D) = \frac{P(D|H_2,\mu^*)\delta\mu}{P(D|H_2,\mu^*)\delta\mu + L_\mu P(D|H_1)} \quad .$$

The probability $P(D|H_2,\mu^*)$ tends to grow relative to $P(D|H_1)$ as we allow searches over bigger ranges (new data sets). However, there is a penalty δ_μ/L_μ for allowing the signal to appear anywhere in the spectrum, and this will shrink as L_μ is expanded, compensating for the larger $P(D|H_2,\mu^*)$. Since the search range L_μ is presumably much greater than the resolution σ, the penalty factor can be quite small. Every additional parameter (dimension in which we search) will bring such a reduction factor.

6 *p*-values and incomplete sets of models

A full Bayesian analysis is only possible if we have a complete set of models. In the GERDA example, we performed a kind of either/or (background model or background+specific signal). We are often in a situation where we are not sure if we have found a complete (enough) set of models. What do we do if we want to include also other possibilities (other types of signals could be present, the background estimate could be faulty) ? We may not even know whether we should include other possibilities. A hierarchical structure can be set up as was done for the BAT solution to the BANFF challenge [3]. The logic for searching for new physics in this case is shown diagrammatically in Fig. 2. The logic is based on using p-values, and is implicitly a Bayesian argument (see [4]). It is assumed that incorrect models have p-value distributions sharply peaked at 0, so that a small p-value gives reason to believe that we have found an incorrect model. Without specifying prior beliefs, the argumentation remains vague.

7 Consensus priors

Our degree-of-belief that we have found new physics depends on both the data and the prior belief. The discussion in the physics community on how many sigmas are needed to define a discovery clearly reflects the need for the definition of consensus priors. Different signals will clearly have different priors. E.g., it would come as no great surprise to find the Higgs particle with a mass around 120 GeV. A search for the Higgs in this mass range would start with a sizeable prior belief. On the other hand, signals for

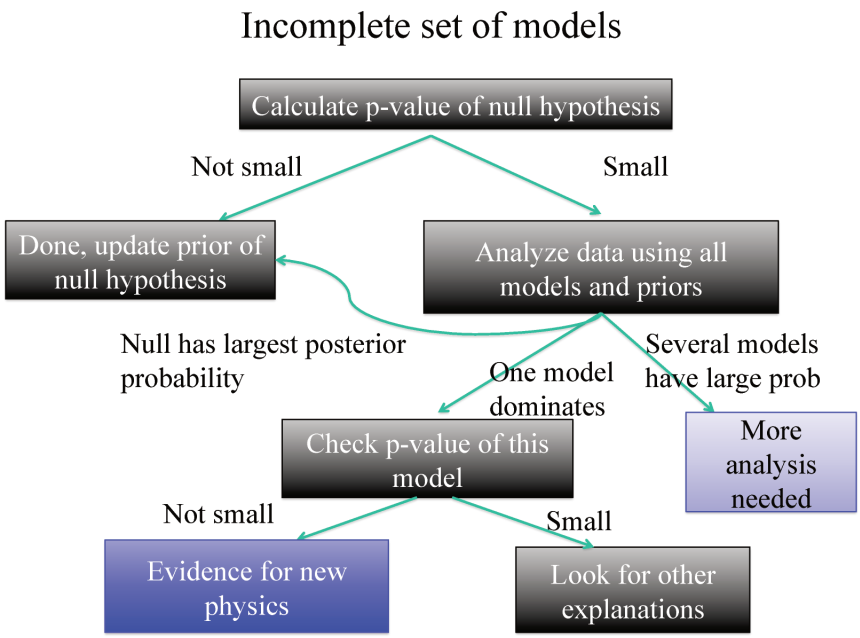


Fig. 2: The logical flow for claiming evidence for new physics (or not) in situations where we are reluctant to define a set of models summing to prior probability of one.

large extra dimensions are a priori considered much more unlikely in this mass range, and would come with a much smaller prior belief. The PHYSTAT community could be a good place to start a discussion towards consensus priors for new physics, at the LHC and also for other experimental searches (direct dark matter detection, neutrinoless double beta decay, ...). For each type of new physics searched for, both 'conservative' and 'optimistic' priors could be defined. Basing analyses on these consensus priors would allow for a transparent means of drawing conclusions on the belief in the new physics. The consensus priors would be updated as the new data came along by a representative body of the community; e.g., represented by a subcommittee of the PDG.

References

[1] See, e.g, sec 33.2.2 of K. Nakamura et al. (Particle Data Group), J. Phys. **G 37**, 075021 (2010).

[2] A. Caldwell, K. Kröninger, Phys. Rev. **D 74** 092003 (2006).

[3] S. Pashapour, 'Bayesian Analysis Toolkit for searches', these Proceedings.

[4] F. Beaujean, A. Caldwell, D. Kollár, and K. Kröninger, Phys. Rev. **D 83**, 012004 (2011).

Statistical Searches in Astrophysics and Cosmology

Ofer Lahav
University College London, UK

Abstract
We illustrate some statistical challenges in Astrophysics and Cosmology, in particular noting the application of Bayesian methods and model selection criteria. We describe two examples where Bayesian methods have improved our inference: (i) photometric redshift estimation and (ii) orbital parameters of extra-solar planets. While sub-communities in Astrophysics, High Energy Physics and Statistics develop separately their specific techniques, it is beneficial to 'compare notes' and to exchange methods.

1 Introduction

The dramatic increase of data in Astronomy has renewed interest in the principles and applications of statistical inference methods. These methods can be viewed as a bridge between the data and the models. Common statistical problems in Astronomy fall broadly into the following tasks:

- Data compression (e.g. of galaxy images or spectra).
- Classification (e.g. of stars, galaxies or Gamma Ray Bursts).
- Reconstruction (e.g. of blurred galaxy images or mass distribution from gravitational lensing).
- Feature extraction (e.g. signatures feature of stars, galaxies or quasars).
- Parameter estimation (e.g. orbital parameters of extra-solar planets or cosmological parameters).
- Model selection (e.g. Are there 0,1,2,... planets around a star? Is a cosmological model with non-zero neutrino mass more favourable?).

It is possible for these tasks to be related. For example, estimation of cosmological parameters from the Cosmic Microwave Background (CMB) or galaxy redshift surveys are commonly deduced from a compressed information, usually in the form of the angular and 3D power spectra, respectively. A further example is classification of galaxy spectra. It can be achieved in a compressed space of the spectra, or in the space of astrophysical parameters estimated from the spectra.

The Astro-statistics community is fortunate to have these days excellent textbooks, among them (in chronological order): Lyons (1986), Lupton (1993), Babu & Feigelson (1996), Sivia (1996), Cowan (1998), Starck & Murtagh (2002), Martinez & Saar (2002), Press et al. (1992), Wall & Jenkins (2003), Saha (2003) and Gregory (2005). Useful reviews on Bayesian methods in Cosmology can be found in the book edited by Hobson et al. (2009) and in Trotta (2008).

2 Inference Methods

There is an ongoing debate between the 'Frequentist' approach and the 'Bayesian' methodology. The 'Frequentist' approach interprets probability as the frequency of the outcome of a repeatable experiment. In contrast, the 'Bayesian' methodology (first published in 1764) views the interpretation of probability more generally and it includes a degree of belief, formulated as:

$$P(model|data) = P(data|model)P(model)/P(data),$$

where on the right hand side the first term is the *likelihood*, the second is the *prior* and the third is the *evidence*.

In the Bayesian approach the choice of *priors* may strongly affect the inference. However it is an 'honest' approach in the sense that all the assumptions are explicitly spelled out in a logical manner.

2.1 Sources of Systematics

A major part of research in Astronomy is devoted to the effect of systematic errors. Consider the example of estimating a specific parameter, e.g. the Dark Energy equation of state parameter w from Baryon Acoustic Oscillations observed in galaxy clustering (e.g. Eisenstein et al. 2005; for review of Cosmological parameters see e.g. Lahav & Liddle 2010). We can distinguish three types of systematics:

- Cosmological uncertainty (due to the assumptions on the other $N - 1$ cosmological parameters' associated priors).
- Astrophysical uncertainty (e.g. what is the relation between the clustering of luminous galaxies and the matter fluctuations?).
- Observational uncertainty (e.g. selection effects in the galaxy sample).

Each of these contributes to the error budget of w in a different way and should be incorporated in the statistical analysis accordingly.

2.2 Justifying Priors

The choice of prior is crucial in the Bayesian framework, yet the justification of each prior is not always spelled out in research articles. To give an example, a prior on the curvature of the universe can be justified in a number of ways, some theoretical, some empirical:

- Theoretical prejudice (e.g. 'according to Inflation, the universe must be flat').
- Previous observations (e.g. 'we know from the CMB WMAP experiment the universe is flat to within 2%, under the assumption of other priors').
- Parameterized ignorance (e.g. 'a uniform prior' or 'a Jeffreys prior').

2.3 Recent Trends in Astro-statistics

Trends noted in recent conferences incude the following:

- Astro-statistics has become a 'respectable' discipline of its own.
- 'Bayesian' approaches are more commonly used, and in better co-existence with 'Frequentist' methods.
- There is more awareness of model selection methods, e.g. the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC), see e.g. Liddle et al. (2006).
- Computer intensive methods, e.g. Markov Chain Monte Carlo (MCMC) are more popular.
- Free software packages are more widely used.

It is beyond the scope of this short review to cover every topic. I shall focus on two examples; photometric redshifts and extrasolar planets. Both cases illustrate how Bayesian approaches have improved our inference on the science questions of interest.

3 An Example From Cosmology: Photometric Redshifts

Mapping the galaxy distribution in 3D requires the galaxy redshifts. In the absence of spectroscopic data, redshifts of galaxies may be estimated using multi-band photometry, which may be thought of as very low-resolution spectroscopy. While the redshift error per galaxy is relatively large, having a great number of galaxies could reduce the errors on measures of the galaxy clustering. Photometric surveys over large areas of the sky may compete well with spectroscopic surveys. Photo-z methods proved very

useful e.g. in recent analyses of the GOODS, COMBO-17 and SDSS Luminous Red Galaxies. Several wide-field photometric redshifts are planned, e.g. the Dark Energy Survey, PanSTARRS, LSST, Euclid and WFIRST. Understanding the photometric redshift errors is crucial for quantifying the errors on e.g. the Dark Energy equation of state parameter w from galaxy clustering and weak lensing.

In more detail, photometric redshift methods rely on measuring the signal in the photometric data arising from prominent "break" features present in galaxy spectra e.g. the 4000 A break in red, early-type galaxies, or the Lyman break at 912 A in blue, star-forming galaxies. There are two basic approaches to measuring a galaxy photometric redshift z (e.g. Csabai et al. 2003 and references therein). The first, template matching, relies on fitting model galaxy spectral energy distributions (SEDs) to the photometric data, where the models span a range of expected galaxy redshifts and spectral types. This is done via a simple χ^2 statistic, i.e. via the likelihood $P(colours|z)$, but it may lead to catastrophic errors. Benitez (2000) generalized the method by incorporating Bayesian priors. The prior $P(z|magnitude)$ for the redshift of a galaxy given its magnitude (apparent luminosity) then multiplies the likelihood to give the posterior

$$P(z|colours, magnitude) \propto P(colours|z, magnitudes) \times P(z|magnitude) \ .$$

This Bayesian chain, which can also be generalized to include galaxy type, greatly reduces the number of outliers.

Another approach utlilises an existing spectroscopic redshift sample as a training set to derive an empirical photometric redshift fitting relation. An example of a training-based method, ANNz, which is also Bayesian, utilizes Artificial Neural Networks (Collister & Lahav 2004). When applied to SDSS galaxies the rms error using ANNz is $\sigma_z = 0.02$, compared with $\sigma_z = 0.07$ using a template method.

4 An Example From Extra-solar Planets: Orbital Parameter Estimation

Astronomers have faced a growing number of free parameters in modelling astrophysical systems, for example cosmological parameters or extra-solar planet orbital parameters. In the case of a model with N free parameters marginalizing over N-1 parameters, it proves to be computationally expensive if the parameter space is mapped into a grid. An alternative method, the Markov Chain Monte Carlo (MCMC), has been known since the 1950's and a wide range of methods exists in the literature to implement it, e.g. the Metropolis-Hasting algorithm.

The key idea is to turn a probability distribution function in N dimensions into a cloud of points which represents the probability distribution function. The probability distribution function could incorporate the probabilities for the priors, in the Bayesian spirit. The MCMC algorithm constructs a random walk in the model parameter space with steps drawn from a multi-dimensional proposal distribution (e.g. a Gaussian). It is crucial to apply tests for convergence, i.e. to ensure that the parameter space is properly sampled, in particular if there are several peaks in a high dimensional space.

MCMC algorithms have been applied widely to parameter estimation from the CMB and other cosmological data sets (e.g. Lewis & Bridle 2003; Verde et al. 2003) and to both detecting and characterizing orbits of extrasolar planets (e.g. Gregory 2005; Ford 2005; Balan & Lahav 2009).

Nearly 2000 extrasolar planets have been discovered so far. Most of those were discovered using measurements of the radial velocity of the host star. The radial velocity curve can be modelled by approximately a dozen parameters, depending on the complexity of the assumed model. It is also important to allow for more than one planet around the star, hence for more free parameters. This leads to the challenging problem of non-linear minimization in a high dimensional parameter space. Deriving these parameters accurately is very important as this can then influence the interpretation for an individual object, as well as the statistics of orbital parameters for an ensemble of extra-solar planets.

For example, in many of the discovery papers the approach taken is to estimate first the period P and then for that fixed P to solve later for the orbital parameters. As there is degeneracy of parameters

and dependence on their priors this could lead to the wrong value of P. This was pointed out by Gregory (2005), who developed an MCMC Bayesian approach to cope with the multi-parameter estimation. He illustrated the method for the data for HD73526, where he found three possible solutions for P. In fact the previously reported one turned out to be the least probable orbit (but apparently the data for this system somewhat changed since the publication of the paper).

5 Future Work in Astro-statistics

The following topics represent current and further work in Astro-statistics:

- Model selection methodology (e.g. which criteria and the role of priors).
- MCMC machinery and extensions (e.g. nested sampling).
- Detection of non-Gaussianity and shape finders (e.g. for galaxy survey and CMB maps).
- Blind de-convolution (e.g. for recovering galaxy shapes from blurred images).
- Object classification (e.g. stars, galaxies and quasars).
- Comparing simulations with data (e.g. large galaxy surveys with N-body and hydrodynamic simulations).
- Visualization (of e.g. 3D galaxy surveys or multi-parameter space).
- Virtual Observatories (including both Real Data and Mock data).

Astronomy, High Energy Physics and Statistics independent communities and meetings like this provide great opportunities to 'compare notes' and exchange ideas. Fundamental issues in statistical inference from data will not go away. With the exponential growth of data in Astronomy there is a great need for further interaction of astronomers with experts in other fields.

Acknowledgements

I would like to thank my collaborators for numerous dicussions over the years on the topics mentioned briefly here, and the organizers for a stimulating meeting.

References

[1] Babu, G.J., Feigelson, E.D., 1996, *Astrostatistics*, Chapman & Hall

[2] Balan, S. & Lahav, O., 2009, MNRAS, 394, 1936

[3] Benitez, N., 2000, ApJ, 536, 571

[4] Collister, A. & Lahav, O., 2004, PASP 116, 345

[5] Cowan, G.,1998 *Statistical Data Analysis*, Oxford University Press

[6] Csabai, I., et al. 2003, AJ, 125, 580

[7] Eisenstein, D.J. et al., 2005, ApJ, 633, 560

[8] Ford, E.B., 2006, ApJ, 642, 505

[9] Gregory, P.C., 2005, *Bayesian Logical Data Analysis for the Physical Sciences*, Cambridge University Press

[10] Hobson et al. (eds.) , 2009, *Bayesian Methods in Cosmology*, Cambridge University Press

[11] Lahav, O.. & Liddle, A. R., 2010, in *Reviews of Particle Physics*, arXiv:1002.3488

[12] Lewis, A. & Bridle, S.L., 2002, PRD, 66, 103511

[13] Liddle, A.R., Mukherjee, P. & Parkinson, D., 2006, astro-ph/0608184

[14] Lupton, R., 1993, *Statistics in Theory and in Practice*, Princeton University Press

[15] Lyons, L., 1986, *Statistics for Nuclear and Particle Physicist*, Cambridge University Press.

[16] Martinez, V.J., Saar, E., 2002, *Statistics of the Galaxy Distribution*, Chapman & Hall/CRC
[17] Press, W.H. et al., 1992, *Numerical Recipes*, Cambridge University Press
[18] Saha, P., *Principles of Data Analysis*, 2006 Available free on the WWW.
[19] Sivia, D., 1996, *Data Analysis: A Bayesian Tutorial*, Oxford University Press
[20] Starck, J-L, Murtagh, F., 2002, *Astronomical Image and Data Analysis*,
[21] Trotta, R., 2008, Contem. Physics, 49 (2), 71
[22] Verde, L. & the WMAP Team, 2003, ApJS, 148, 196
[23] Wall, J.V, Jenkins, C.R., 2003, *Practical Statistics for Astronomers*, Cambridge University Press

Setting Limits, Computing Intervals, and Detection

David A. van Dyk
Department of Statistics
University of California, Irvine, CA, United States

Abstract
This article discusses a number of statistical aspects of source detection, the computation of intervals and upper limits for a source intensity, and accessing the sensitivity of a detection procedure. Emphasis is placed on model diagnostics, validation, and improvement as means of avoiding odd behaviors in these procedures such as over abundant short or empty intervals. Improved model specification is viewed as a better response to systematic uncertainties, the look elsewhere effect, and general model inadequacy than simply insisting on a significance level of 5σ for source detection. We advocate reporting *both* the upper limit and the sensitivity to better represent the strength of evidence for detection and the reported source intensity. Finally, we explore the use of decision theoretic analysis to derive detection procedures, intervals, and limits in order to focus attention on the statistical properties of primary interest.

1 Introduction

Over the past 10-15 years there has been much discussion in the high energy physics community as to how best to derive statistical criterion for source detection and how best to compute intervals and limits for source intensities, see e.g., [1–3]; and the proceedings for the Phystat Conference Series (`URL: phystat.org`). This paper picks up a number of threads in this discussion from a statistical point of view and with an emphasis on encouraging adequate model specification and proper reporting of results. From my point of view the discussion has been too focused on technical properties and somewhat superficial concerns pertaining to statistical procedures. Thus, this paper explores a decision theoretic approach with the aim of focusing attention on the statistical properties most pertinent to ultimate scientific goals.

The paper is organized into five sections. In Section 2 we review the basic statistical framework for source detections and setting intervals and upper limits for the source intensity. Important in this is the clarification of a difference in nomenclature used in high energy physics and in astrophysics. In Section 3 we discuss a number of concerns that have arisen with this framework. The use of decision theoretic analysis to derive new procedures for detection and computing intervals and limits is explored in Section 4. The paper is summarized in Section 5.

2 Detection, Intervals, and Upper Limits

2.1 A Simple Poisson Model

To focus attention on the statistical issues we frame our discussion in terms of a simple detection problem involving a contaminated Poisson count. The methods and issues described are general, but the salient points are evident in this simple example. Thus, we consider the Poisson model for a source count,[1]

$$n|(\lambda_S, \lambda_B, \tau_S) \sim \text{Poisson}\Big(\tau_S(\lambda_S + \lambda_B)\Big), \tag{1}$$

where n is the source count, λ_S is the source intensity, λ_B is the background intensity, and τ_S is the source exposure time. We typically have a second background-only exposure that we model as

$$n_B|(\lambda_B, r, \tau_B) \sim \text{Poisson}(r\tau_B\lambda_B), \tag{2}$$

[1]The notation $X|Y \sim \text{Distribution}(Y)$ describes the *conditional* distribution of X *given* Y. For example, $X|Y \sim \text{Poisson}(g(Y))$ means that the conditional probability mass function of X given Y is $\exp\{-g(Y)\}g(Y)^X/X!$.

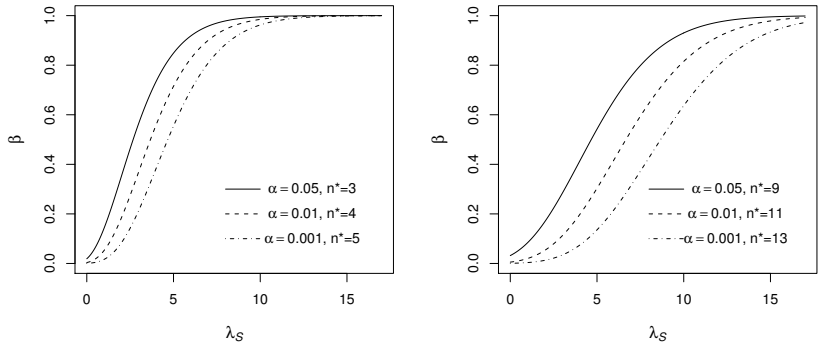


Fig. 1: The Power of the Detection Plotted as a Function of the Source Intensity, λ_S. The two panels correspond to $\lambda_B = 1$ and 5. In each panel the power is given for three values of α and their corresponding detection thresholds. The power of the detection increases with the source intensity and decreases with the background intensity. Insisting on a lower probability of a false detection (smaller α) decreases the power of the detection.

where n_B is the background count, τ_B is the background exposure time, and r is the relative area of the background and source exposures. For clarify, we sometimes assume λ_B is known. In any case, λ_S is of primary interest. We wish to determine if there is a source and if so how strong it is. Even if we cannot detect a source, we may wish to quantify how strong a possible source could be and go undetected.

A standard statistical hypothesis testing framework is used for source detection. In particular the default or *null hypothesis* states that there is no source. We assume this to be true unless we find this assumption to be at odds with the observed data, in which case we reject the null hypothesis in favor of the *alternative hypotheses* that a source is present. Formally, we write

$$H_0: \quad \text{There is no source, i.e., } \lambda_S = 0 \tag{3}$$

$$H_A: \quad \text{There is a source, i.e., } \lambda_S > 0. \tag{4}$$

2.2 Detection

To determine whether the observed data are at odds with the null hypothesis, we first identify a *test statistic* which is a function of the data for which larger (or smaller) values correspond to stronger evidence against the null hypothesis. In our simple Poisson example, the source count, n, is an obvious choice. Having identified a test statistic, we define the *detection threshold*, $n^\star$, as the smallest value such that

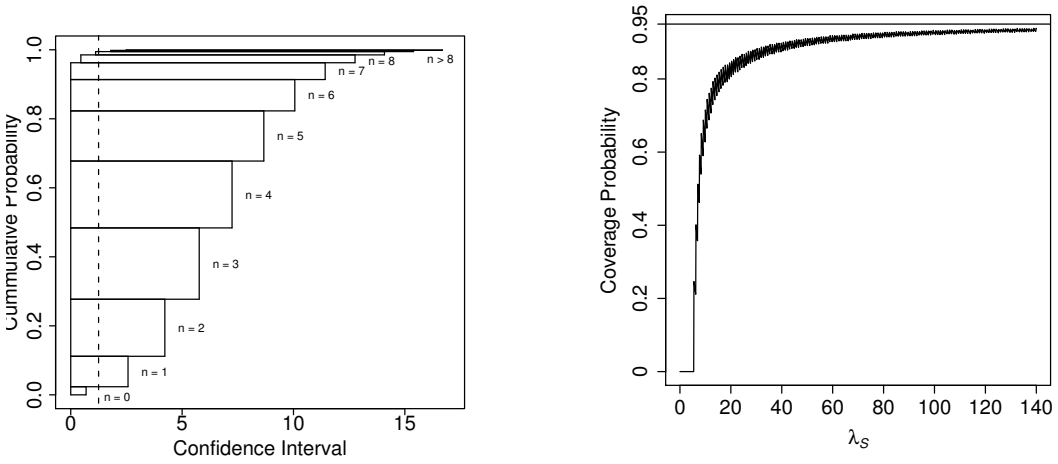
$$\Pr(n > n^\star | \lambda_S = 0, \lambda_B, \tau_S, \tau_B, r) \leq \alpha. \tag{5}$$

By conditioning on $\lambda_S = 0$ we are assuming there is no source. Under the null hypothesis the probability of a source count larger than $n^\star$ is less than or equal to the *significance level of the detection*, α. If α is set sufficiently small, and the source count is greater than $n^\star$, we conclude that there is sufficient evidence to *reject the null hypothesis* in favor of the alternative hypothesis that a source is present.

We choose a small value of α to minimizing the probability of a false detection. Of course, we can compute $\Pr(n > n^\star)$ for positive values of λ_S, in which case this becomes the probability of a true detection, which we would like to be as large as possible. The probability of a true detection depends on the the value of λ_S, is known as the *power of the detection*, and can be written

$$\beta(\lambda_S) = \Pr(n > n^\star | \lambda_S, \lambda_B, \tau_S, \tau_B, r)\,. \tag{6}$$

Note $\beta(\lambda_S = 0) \leq \alpha$ and $\beta(\lambda_S)$ is simply the probability of a detection, false or true depending on λ_S. The dependencies of the power on the source intensity and the level of the test are illustrated in Fig. 1.



(a) Sampling Distribution of a 95% Interval. The horizontal ranges of the rectangles give the confidence intervals for the given value of n, with $\lambda_B = 3.0$. Rectangle heights are the probabilities of each n and thus the probabilities of the intervals, see [2].

(b) Under Coverage. The plot shows the true coverage of 95% intervals that are only reported when a source is detected with significance level $\alpha = 0.05$ and with $\lambda_B = 5$. The true coverage is far below the nominal coverage for weak sources.

Fig. 2: Distribution and Under Coverage of Selectively Reported Confidence Intervals of [4].

2.3 Confidence Intervals, Sensitivity, Upper Limits, and Upper Bounds

A formal hypothesis test is only the first step in source detection. Whether or not there is a detection, we typically want to quantify the plausible values for the (possible) source intensity. This is certainly of interest in the event of a detection, but even in the absence of detection there is typically a non-zero probability of a *false negative*, that is, an undetected source. Formally, this probability that a source goes undetected is $1 - \beta(\lambda_S)$ and is generally expected to diminish as λ_S increases but to approach $1 - \alpha$ for λ_S near zero. (In principle $\beta(\lambda_S)$ may be discontinuous at zero or may not asymptotically approach one, but these are unusual cases.) Thus, even in the absence of a detection, a quantification of the plausible values of λ_S is of value. This quantification typically takes the form of an upper limit and/or an interval.

A frequentist *confidence interval* for λ_S aims to give the *plausible values of* λ_S. This is defined to be any interval that includes the true value of λ_S a given proportion of the time over the long run upon repetition of an experiment. Formally, we can derive an interval $\mathcal{I}(\lambda_S)$ for each value of λ_S, such that

$$\Pr(n \in \mathcal{I}(\lambda_S)|\lambda_S) \geq 95\%, \tag{7}$$

where 95% is the *confidence level* and can be replaced by any desired level. Upon observing a particular value of $n_{\rm obs}$ of n, a frequency confidence interval can be constructed as

$$\{\lambda_S : n \in \mathcal{I}(\lambda_S)\}. \tag{8}$$

Here we avoid the issue of *nuisance parameters*, such as λ_B. The probability in Equation 7 clearly depends on λ_B and thus so do the intervals $\mathcal{I}(\lambda_S)$ which complicates the construction of the confidence interval in Equation 8. Although this is an important issue, it is not central to our discussion, and we will simply fix λ_B at some known value when computing confidence intervals. Fig. 2(a), for example, illustrates the frequency properties of a Garwood's (1936) choice of interval for λ_S.

The upper end point of a one-sided confidence interval is called an *upper limit* by physicists (or an *upper bound* by astronomers). This is the largest plausible value of the source intensity consistent with the observation. Fig. 2(a) illustrates how one sided confidence intervals arise when n is relatively small.

In astronomy, an *upper limit* is used to quantify the source intensity of a possible, but undetected source. In particular, to an astronomer an upper limit is *the maximum intensity that a source can have*

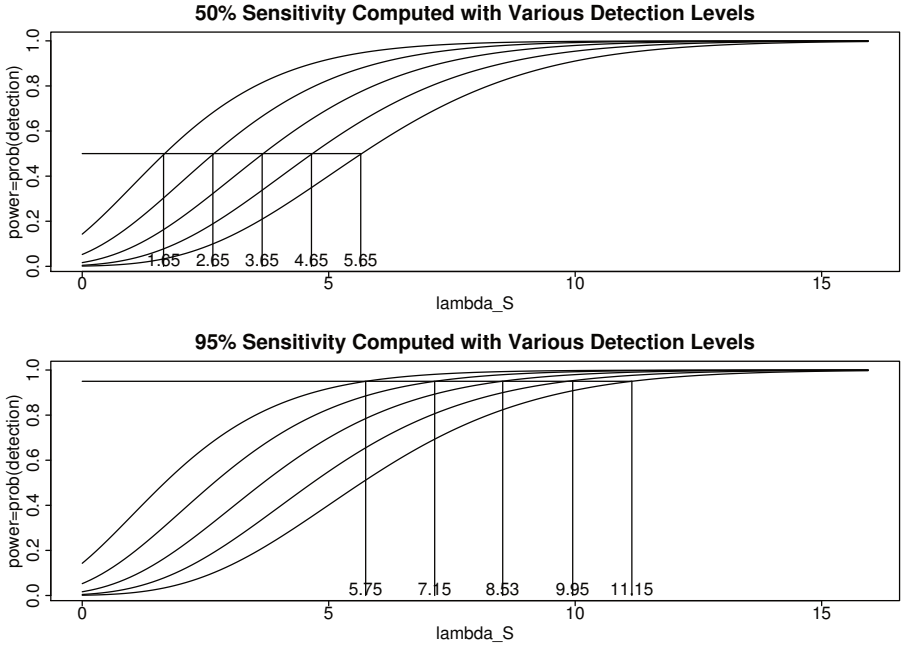


Fig. 3: Effect of α and $\beta_{\min}$ on the Upper Limit. The five curves in each panel give the probability of detection, $\beta(\lambda_S)$ for each of five values of the significance level, α, from left to right: 0.143, 0.053, 0.017, 0.005, and 0.001. When computing sensitivities we derive the minimum value of λ_S that has at least a probability of $\beta_{\min}$ of being detected. This is done for $\beta_{\min} = 0.50$ in the first panel and 0.95 in the second. The sensitivity of the detection increase as $\beta_{\min}$ increases and as α decreases.

without having at least a probability of $\beta_{\min}$ of being detected under an α-level detection threshold, or conversely, *the smallest intensity that a source can have with at least a probability of $\beta_{\min}$ of being detected under an α-level detection threshold*, see [5]. Physicists generally refer to this as the *sensitivity* of the detection. We will use the term "sensativity" from now on. Computing the sensitivity requires two probability calculations. The detection threshold is computed with the probability calculation in Inequality 5 and the probability of detection is computed using Equation 6. This is illustrated in Fig. 3.

The sensitivity of the detection is analogous to a sample size in that they both quantify the strength of an experiment. Larger sample sizes correspond to more powerful experiments that can detect weaker signals. Likewise smaller (i.e., better) sensitivities indicate a more powerful observation: any source with intensity greater than the sensitivity is expected to be detected (as calibrated by α and $\beta_{\min}$). The sensitivity directly quantifies the power in terms of the quantity of primary interest: the source intensity.

In a typical statistical power calculation, we find the minimum exposer time, τ_S, by solving Equation 6 so that the probability of detection achieves a minimum value for a given λ_S. For example, we might want to find the minimum exposure time so that $\beta(\lambda_S = 2) \geq 0.90$ if we want to be sure there is at least a 90% chance of detecting a source with intensity equal to two counts per unit time. The sensitivity of the detection is found by solving the same equation, but for λ_S with τ_S fixed. It is important to notice that all of these calculations can be done *before the observation is made*. Like power, the sensitivity does not depend on the data and can be computed in advance.

3 Addressing Concerns (Please Forgive my Soap Box!)

3.1 What Should be Reported?

A typical procedure for source detection in astronomy involves reporting different quantities depending on whether the source is detected [5]. When there is a detection astronomers often (i) report a detection and (ii) report a confidence interval for λ_S. When there is not a detection astronomers often (i) report no detection and (ii) report a detection sensitivity for λ_S. Similarly, with power-constrained limits, the data-dependent upper limit is only reported if it is greater than the sensitivity of the detection, otherwise the data-independent sensitivity is reported, see, e.g., [6]. Deciding whether or not to report an interval (or limit) *based on the data* alters its frequency properties [5,7]. This is illustrated in Fig. 2(b) which reports the frequency coverage of intervals that are only reported in the case of a detection. For small values of λ_S, the coverage can be far below its nominal value. Unfortunately, frequency properties depend on what you would have done, had you had a different data set.

To eliminate the coverage problems described in Fig. 2(b) and to provide a more complete summary of what was learned from the observation, [5] proposes that we *always report*

1. whether the source was detected,
2. a confidence interval for the source intensity (which may be a one-sided upper limit), and
3. the sensitivity of the detection, in order to quantify the strength of the experiment.

This is in contrast to both the power-constrained limit that report the larger of the sensitivity and the upper limit and to CL_S [8] that alters the upper limit in order to produce a smoothed version of the power-constrained limit [9]. Both of these procedures sacrifice frequency properties and lack a clear probabilistic interpretation. By reporting both the upper limit and the sensitivity, we provide both the largest value of λ_S consistent with the data (the upper limit) and the smallest value that we have sensitivity to detect. Reporting both the upper limit and the sensitivity is certainly more informative than reporting either max(upper limit, sensitivity) or a smoothed version of this maximum.

3.2 Short or Empty Confidence Intervals

One particular concern regarding available methods is the possibility that frequency-based intervals may be empty or very short. The former case is generally disconcerting and the later is interpreted by some users as implying an exaggerated experimental sensitivity. In my view this stems for a simple misunderstanding of the proper interpretation of the frequency-based intervals. Recall that a (say) 95% frequency-based interval is simply an interval constructed so that there is a 95% probability that an experiment conducted as formalized by the probabilistic model will result in an interval that contains the true value of λ_S. Fig. 2(a) illustrates that the same experiment sometimes produces relatively short and sometimes produces relatively long intervals. The sensitivity of an experiment, however, does not depend on the observed count. In the example in Fig. 2(a), the sensitivity is the same regardless of whether we observe $n = 0$ and obtain a short interval, or observe $n = 8$ and obtain a long interval.

Another difficulty is a tendency to interpret the *pre-data probabilities* associated with frequency intervals as *post-data probabilities*. A 95% interval will produce intervals that contain the true value of λ_S 95% of the time when observations are generated under the model, regardless of the true value of λ_S. Such a procure can produce empty intervals, so long as they are produced less that 5% of the time and overall at least 95% of the intervals contain the true value. (Of course the empty intervals may be wasteful!) This is not to say that an empty—or any other particular—interval has a 95% chance of containing the true value. An empty interval certainly does not contain the true value of λ_S, regardless of the frequency probability of the interval. Although our intuition leads us to interpret these probabilities in a post-data manner, frequency-based probabilities say nothing about the properties of a particular interval. Bayesian methods are better suited to quantifing post-data probabilities. The precise nature of frequency probabilities may be appealing, but precise probabilities are not necessary relevant probabilities.

Under the construction described in Section 2.3, we can further interpret the intervals as reporting values of λ_S that are *consistent with the observation*, where "consistent" is calibrated by the probability level associated with the interval. Short or empty intervals simply mean that there are few or no values of λ_S that are consistent with the observations. As illustrated in Fig. 2(a), very short intervals are possible, but are expected to be rare. Depending on how the interval is constructed, the same can be said for empty intervals. If empty or short intervals (relative to the sensitivity) are common, it is a clear indication that the probabilistic model used to describe the observation is inadequate—regardless of the strength of the *subjective prior belief in the underlying model*. Model checking, validation, and improvement are standard components of any statistical analysis. I expect far more would be gained by focusing on model improvement rather than on statistical properties of a particular statistical procedure.

3.3 5σ

It has become standard to require $\alpha = 1/1.7 \times 10^6$ for a detection in high energy physics, corresponding to the probability that a standard normal variable exceeds five standard deviations from its mean. This corresponds to a false positive rate of one in 1.7million experiments. Of course, the motivation is not to keep the false detection rate this low, but to attempt to account for other concerns such as the look elsewhere effect [3, 10], calibration and/or systematic errors, and statistical error rates that are not well calibrated due to general model misspecification [3, 11]. Unfortunately, reducing α does not really address these concerns. We do not know the actual effects of systematics and the look elsewhere effect on the final analysis. They likely induce both increased bias and variance. Reducing α does not address bias at all and is a completely uncalibrated response to variance. Even in the absence of these problems, statistical procedures are not well calibrated at such extreme depths in the tails of the sampling distributions, which are typically based on asymptotic approximations. Computing extreme tail probabilities poses its own challenges in all but the simplest cases [12]. Taken together these concerns lead us to conclude that we have no idea what the probability of a false detection is—the procedure itself is wholly uncalibrated.

The difficulty here is similar to what leads to over-abundant empty or narrow confidence intervals: model misspecification. The solution is not to crank down the value of α, but rather to directly deal with systematics, calibration, the look elsewhere effect, and general model misspecification. Model checking and improvement are the key to better statistical properties of detection procedures, intervals, and limits [2, 13]. Hiding unrealistic assumptions and using *ad hoc* fixes (such as using a 5σ detection criterion) do not address the root problems, but do make evaluating their effects more difficult. Calibration, systematics, and the look elsewhere effect must be modeled directly. Reasonable model specification is far more important than the detailed properties of a statistical procedure or the choice of a Bayesian, Frequentist, or other procedure. The ultimate goal is honest frequency error rates and/or a calibrated Bayesian procedure, both of which depend absolutely on careful model specification.

4 A More Coherent Approach?

4.1 Hypothesis Testing in High Energy Physics

Source detection in high energy physics is often conducted using a more involved hypothesis-testing procedure than is described in Sections 2–3. In addition to testing the hypotheses in Equations 3–4, a second hypothesis test is often conducted in tandem that interchanges the roles of the null and alternative hypotheses, see [14]. Rather than under the default assumption of no source, a second "detection threshold" is computed under the assumption that there is a source and the significance test is conducted treating the original alternative hypothesis as the null hypothesis and treating the original null hypothesis as the alternative hypothesis. (For clarity, we continue to use H_0 for the hypothesis of no source and H_A for the hypothesis that there is a source. In the reversed formulation of the significance test, we assume H_A when computing the second detection threshold, $n_A^\star$, in analogy to Equation 5.)

This reversed formulation of the hypothesis test is motivated by a well-known challenge associated

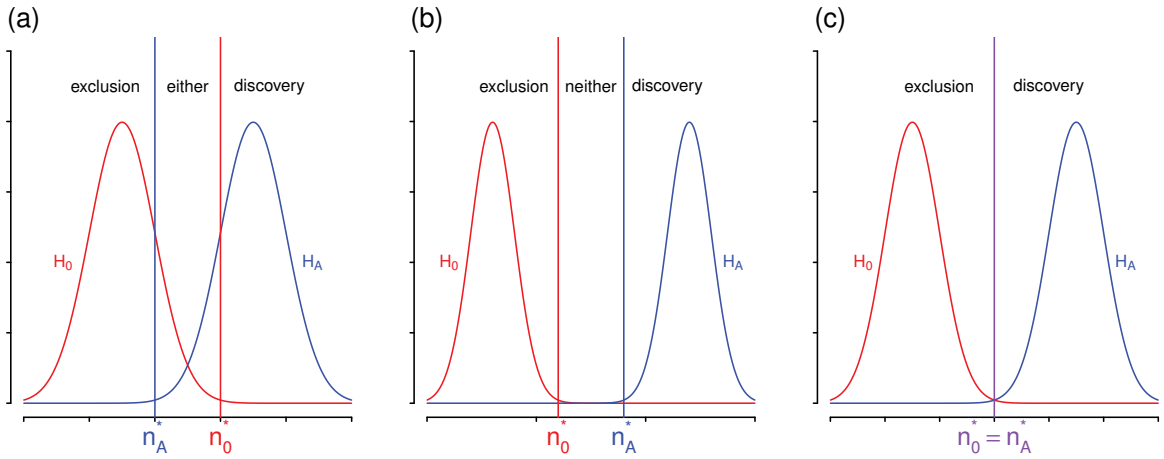


Fig. 4: Combining the Original and Reversed Significance Tests. The two curves in each panel depict the distribution of the test statistic under H_0 (left, red) and H_A (right, blue). Although we use the notation of our running example, here we assume that both distributions are fully specified, i.e., that they do not depend on any unknown parameters. The "detection" thresholds are denoted by $n_0^\star$ and $n_A^\star$, where $n_0^\star$ is the 1-α percentile of the distribution of the test statistic under H_0 and $n_A^\star$ is the α percentile under H_A. The three panels give the decision regions under three scenarios: (a) $n_0^\star > n_A^\star$, (b) $n_0^\star < n_A^\star$, and (c) $n_0^\star = n_A^\star$. We accept H_0, but reject H_A if $n < \min(n_0^\star, n_A^\star)$; reject H_0, but accept H_A if $n > \max(n_0^\star, n_A^\star)$; reject both H_0 and H_A if $n_0^\star < n < n_A^\star$; and accept both H_0 and H_A if $n_A^\star < n < n_0^\star$. Notice that in each of the scenarios, at most three of the four decisions is possible.

with *model selection*: a model being the better of two at explaining the data does not mean that it is an adequate model. In the context of hypothesis testing, rejecting the null hypothesis indicates that the hypothesis is inadequate for explaining the data, at least in the dimension quantified by the test statistic. This alone, however, is not enough for us to conclude that the alternative hypothesis *is* adequate. There are other possibilities besides the model given in Equations 1–2 with $\lambda_S = 0$ and with $\lambda_S > 0$. The reversed hypothesis test aims to identify evidence that $\lambda_S > 0$ is inadequate as well. Of course, all hypothesis tests look for evidence in the dimension specified by the test statistic, so the interplay of the original and the reversed hypothesis tests depends intimately on the two choices of test statistics.

The decision in the original hypothesis testing framework involves either *accepting* H_0 or *rejecting* H_0. When we conduct both the original and the reversed hypothesis test, each test has these two possible outcomes, leading to a total of four possibilities:

exclusion: accept H_0 and reject H_1,

discovery: reject H_0 and accept H_1,

no decision: accept both hypotheses (*either* is possible), or

excluding both: reject both hypotheses (*neither* is possible).

As illutrated in Figure 4, in any particular situation only one of "no decision" and "exclude both" is possible, depending on the ordering of the detection thresholds for the two hypothesis tests.

While it is completely standard to use model diagnostics and checking to evaluate the adequacy of any statistical model, formal symmetric testing of H_0 and H_A in this way is unusual, if not unique to high energy physics. Inverting a significance test to form confidence intervals or upper limits is a related and very common technique. This involves treating each possible value of the parameter as a null hypothesis and compiling the interval as the set of parameter values that are not rejected at a given α-level. An additional complication arrises in high energy physics in that different significance levels are used for the original and the reversed significance tests, typically 5σ and 2σ, respectively. In the following section we employ a decision theoretic approach to analyze the use of such symmetric testing.

Table 1: Loss Functions. Table (a) gives a detailed loss function for the six possible errors if we assume either H_0 or H_A is true. To simplify calculations, Table (b) gives a loss function where the cost of all errors except a false detection are equal.

(a)

	Decision			
Truth	exclusion	discovery	no decision	exclude both
H_0	0	C_{01}	C_{0e}	C_{0n}
H_A	C_{10}	0	C_{1e}	C_{1n}

(b)

	Decision			
Truth	exclusion	discovery	no decision	exclude both
H_0	0	C	c	c
H_A	c	0	c	c

4.2 A Decision Theoretic Approach: Loss, Risk, and Bayes Risk

Although concerns about detection procedures are often expressed in terms of detailed observations about the character of procedures under certain circumstances (e.g., the upper limit may increase as n decreases), a desire for strict adherence to frequency properties (e.g., the "Goldilocks effect": coverage should be above a minimum, but no more than the minimum); and apprehension about Bayesian methods and their prior distributions, e.g., [1], ultimately we are primarialy concerned with rates of detection errors and ensuring that intervals and limits do a good job of capturing the true source intensities. In this section, we discuss a *decision theoretic analysis* that allows us to directly optimize a detection procedure in terms of the quantities of ultimate interest.

We begin with a loss function that quantifies the cost of the possible errors in a significance test with the four possible decision: "exclusion", "discovery", "no decision", and "exclude both". With four possible decisions there are more possible errors than the "false detection" and "false negative" of a standard significance test, see Table 1(a). While it can be argued that "no decision" is not an "error" regardless of the truth, this decision is clearly less desirable than a true exclusion or a true discovery. In this regard it is appropriate to assign a non-zero loss to this decision, even if it is not an "error". A more complete table would include a third row, "Truth = Neither" to capture the possibility that neither H_0 nor H_A holds. We avoid this possibility because the necessary probability calculations are arbitrary when no true model is specified. In Table 1(a), C_{01} is the cost of the most troubling error, a false positive. The costs of the all other errors are likely significantly smaller than C_{01}. The loss function in Table 1(b) quantifies this by setting the cost of all other errors to $c \ll C = C_{01}$. That is, for simplicity we assume that all errors except a false detection have an equal cost that is dominated by the cost of a false detection. Finally we assume that $C + c = 1$; this is simply a choice of scale for the loss function.

Given detection thresholds, $n_0^\star$ and $n_A^\star$, we compute the *risk*, which is the expected loss, under H_0,

$$\text{Risk}(n_0^*, n_A^*|H_0) = C\,\Pr[n > \max(n_0^*, n_A^*)|H_0] + c\Big\{\Pr[n_0^* < n < n_A^*|H_0] + \Pr[n_A^* < n < n_0^*|H_0]\Big\}$$

and under H_A,

$$\text{Risk}(n_0^*, n_A^*|H_1) = c\,\Pr[n > \min(n_0^*, n_A^*)|H_1] + c\Big\{\Pr[n_0^* < n < n_A^*|H_1] + \Pr[n_A^* < n < n_0^*|H_1]\Big\}.$$

Our goal is to find $n_0^\star$ and $n_A^\star$ to minimize the risk. The *Bayes risk* averages $\text{Risk}(n_0^*, n_A^*|H_0)$ and $\text{Risk}(n_0^*, n_A^*|H_A)$ using a probability of H_A, denoted by π,

$$\text{Bayes Risk}(n_0^*, n_A^*|\pi) = (1-\pi)\ \text{Risk}(n_0^*, n_A^*|H_0) + \pi\ \text{Risk}(n_0^*, n_A^*|H_A).$$

To minimize the Bayes risk, we make a simplifying assumption that the test statistic has a continuous distribution with probability density function f_0 under H_0 and f_A under H_A. This is not the case in the Poisson model, where n is a count. Under this assumption the Bayes risk is minimized either when

$$C = \frac{(1-\pi)f_0(n_0^*) + \pi f_A(n_0^*)}{2(1-\pi)f_0(n_0^*) + \pi f_A(n_0^*)} = \frac{(1-\pi)f_0(n_A^*) + \pi f_A(n_A^*)}{2(1-\pi)f_0(n_A^*) + \pi f_A(n_A^*)}$$

or at a point where the Bayes risk is not differentiable, $n_0^* = n_A^*$. Thus the optimal choice of $n_0^\star$ and $n_A^\star$ occurs when $n_0^\star = n_A^\star$, with the particular optimal value of $n_0^\star = n_A^\star$ determined by C and c. This corresponds to the standard detection setup in that there are only two possible decisions, see Fig. 4(c). This result depends on the simple loss function given in Table 1(b) and would be different if different costs were assigned to a false exclusion and the "no decision" and "exclude both" decisions under H_0 and/or H_A. Of course quantifying the relative costs of the various errors in Table 1 is not an easy task.

The result can be understood by referring to Fig. 4(a). Suppose we fix $n_0^\star$ and adjust $n_A^\star$ with the aim of decreasing the risk under H_0. Increasing $n_A^\star$ increases the probability of the correct (zero cost) decision of "exclusion" and reduced the probability of the c-cost decision of "no decision" or "either". Thus, we should increase $n_A^\star$ to be at least as large as $n_0^\star$. Likewise, if we again fix $n_0^\star$ and increase $n_A^\star$ under H_A we increase the probability of "exclusion" at the expense of the probability of "either", both of which have cost c so the overall risk given H_A is unaffected. Similar reasoning can be used in the scenario illustrated in Fig. 4(b) to see that $n_0^\star$ must be at least as large as $n_A^\star$ to minimize the risk. Thus, under the loss function in Table 1(b) the Bayes risk is minimized for $n_0^\star = n_A^\star$, for any value of π.

4.3 Decision Analysis for Intervals and Limits

In Section 4.2 we illustrated how decision theoretic analysis can be used to derive a detection criterion. It is important to emphasize that this construction does not aim to control the probability of a false detection, as in Equation 5. Instead the goal is to control the overall expected loss of the procedure. Of course, if we specify $C \gg c$, false detections will be far less frequent than false negatives. Because we can always construct a confidence interval by inverting a test (as the set of values of λ_0 such that we cannot reject $H_0 : \lambda_S = \lambda_0$), the decision theoretic framework for detection leads to a confidence interval for the source intensity. The coverage of an interval derived from inverting a test is a function of the test's probability of a false positive: if the probability of a false positive is less than α the coverage of the resulting interval will be greater than $1 - \alpha$. Since the decision theoretic approach does not aim to control the probability of a false positive, however, the coverage of the resulting interval will vary.

A better strategy is to specify a loss function to directly quantify the desired properties of the interval or limit. For example, for a interval we might use

$$\text{Loss} = b \times \text{length}(\text{interval}) - I\{\text{interval contains } \theta\}$$

and for an upper limit we might use

$$\text{Loss} = b \times \text{limit} - I\{\theta < \text{limit}\},$$

where θ is a generic parameter of interest, $I\{\text{condition}\}$ is one if the condition is true and is zero otherwise, and b is a tuning parameter that specifies the relative importance of length and coverage. Let $[L(Y), U(Y)]$ be a generic interval computed from data Y. The risk of the interval can be written

$$\text{Risk}(\theta) = b \times \Big\{\text{E}(U(Y)|\theta) - \text{E}(L(Y)|\theta)\Big\} - \Pr\Big\{\theta \in [L(Y), U(Y)] \mid \theta\Big\},$$

where the second term on the right is the coverage. Notice that if we take b equal to zero the risk depends only on the coverage and the optimal interval is the entire parameter space (e.g., $(-\infty, +\infty)$). If we take b equal to ∞, the risk only depends on expected length and the optimal interval has $L(Y) = U(Y)$. Both the expected length and the coverage may depend on the value of θ. The Bayes risk computes the average of both quantities using a distribution on θ.[2] The goal is then to find functions L and U that minimize the Bayes risk. This is generally accomplished by parameterizing L and U. For example in a symmetric problem, we might consider intervals of the form $\hat{\theta} \pm e\hat{\sigma}$, where $\hat{\theta}$ and $\hat{\sigma}$ are estimates of θ and its error. This reduces minimization of the Bayes risk to a one dimensional minimization over e.

[2]Frequentist decision theoretic procedures are available that avoid the use of a distribution on θ by deriving the maximum risk over all values of θ. The interval that minimizes this maximum risk is considered optimal in the *minimax* sense.

5 Summary

The most important aspect of any statistical analysis is the specification of an adequate model. The choice of the specific procedure and/or the choice of statistical paradigm (i.e., frequency-based, Bayesian, or other) are typically far less critical to the properties of the procedure and the ultimate outcome of the analysis. Thus, when a statistical analysis exhibits odd behavior, the first remedy must be model diagnostics, validation, and improvement rather than questioning the choice of statistical procedure under the apparently inadequate model. Decision theoretic analysis allows us to directly specify the statistical properties that we hope for in a procedure and the relative importance that we place on these properties. This strategy is ideally suited to deriving detection procedures, intervals, and limits that exhibit properties that are viewed as best facilitating progress on the ultimate scientific goals.

Acknowledgements. The author thanks Louis Lyons for many helpful conversations on statistical issues in high energy physics, the organizing committees of the *Workshop on Statistical Issues Relevant to Significance of Discovery Claims* held at the Banff International Research Station in July 2010 and of *Phystat 2011* held at CERN in January 2011 for invitations to participate in these stimulating meetings, and the US National Science Foundation for financial support of this work (DMS-09-07522).

References

[1] M. Mandelkern, Statistical Science **17**, 149 (2002).

[2] D. A. van Dyk, Statistical Science **17**, 164 (2002).

[3] L. Lyons, Annals of Applied Statistics **2**, 887 (2008).

[4] F. Garwood, Biometrika **28**, 437 (1936).

[5] V. L. Kashyap *et al.*, The Astrophysical Journal **719**, 900 (2010).

[6] L. Demortier, Power-Constrained Upper Limits to Solve the Sensitivity Problem, Unpublised Manuscript, 2010.

[7] J. Feldman, Gary and R. D. Cousins, Physical Review D **57**, 3873 (1998).

[8] A. L. Read, **81** (2000).

[9] L. Demortier, Open Issues in the Wake of Banff 2010, Presentation at Phystat 2011 (CERN, Geneva, Switzerland), 2011.

[10] L. Lyons, Comments on 'Look Elsewhere Effect', Unpublished manuscript., 2010.

[11] D. Cox, Discovery: A Statistical Perspective, Presentation at Phystat 2011 (CERN, Geneva, Switzerland), 2011.

[12] M. Woodroofe, Importance Sampling and Error Probabilities, Presentation at "Statistical Issues Relevant to Significance of Discovery Claims (BIRS, Banff, Alberta, Canada, `http://www.birs.ca/events/2010/5-day-workshops/10w5068`), 2010.

[13] L. Wasserman, Statistical Science **17**, 163 (2002).

[14] L. Lyons, Statistical Issues in Particle Physics Analysis, Presentation at "Statistical Issues Relevant to Significance of Discovery Claims (BIRS, Banff, Alberta, Canada, `http://www.birs.ca/events/2010/5-day-workshops/10w5068`), 2010.

Bayesian versus frequentist upper limits

Christian Röver[1], *Chris Messenger*[1,2] *and Reinhard Prix*[1]
[1]Max-Planck-Institut für Gravitationsphysik (Albert-Einstein-Institut), Hannover, Germany
[2]School of Physics & Astronomy, Cardiff University, Cardiff, UK

Abstract
While gravitational waves have not yet been measured directly, data analysis from detection experiments commonly includes an upper limit statement. Such upper limits may be derived via a frequentist or Bayesian approach; the theoretical implications are very different, and on the technical side, one notable difference is that one case requires *maximization* of the likelihood function over parameter space, while the other requires *integration*. Using a simple example (detection of a sinusoidal signal in white Gaussian noise), we investigate the differences in performance and interpretation, and the effect of the "trials factor", or "look-elsewhere effect".

1 Introduction

1.1 Upper limits

In general, an upper limit is a probabilistic statement bounding one of several unknown parameters determining the observed data at hand. While it would be hard to derive general properties applicable in any possible data analysis context, we will for illustration consider a simple case here: a sinusoidal signal in white Gaussian noise. This example exhibits many similarities with commonly encountered real-world problems, including the use of Fourier methods, nuisance parameters, trials factors, partly analytical and numerical analysis, etc., and we believe is general enough to yield valuable insights.

1.2 The frequentist case

The frequentist detection approach is based on some *detection statistic* d, which for given data is then used to derive a significance statement along the lines of *"If the data were only noise (null hypothesis H_0), a detection statistic value $\geq d_0$ would have been observed with probability p."* ($\mathrm{P}(d \geq d_0 \,|\, H_0) = p$). The probability p here is the p-value, and a low p-value is associated with a great significance. In the case of a non-detection, the statement then may be reversed to an upper limit statement *"Had the signal amplitude been $\geq A^\star$, a larger detection statistic value ($\geq d_0$) would have been observed with at least 90% probability"* ($\mathrm{P}(d \geq d_0 \,|\, A \geq A^\star) \geq 90\%$), where $A^\star$ is the 90% confidence upper limit (e.g. [1,2]).

1.3 The Bayesian case

In the Bayesian framework, detection and parameter estimation are more separate problems; for detection purposes one would need to derive the *marginal likelihood*, or *Bayes factor*, which (in conjunction with the prior probabilities for the "signal" and "noise only" hypotheses H_1 and H_0) allows one to derive the probability for the presence of a signal. The detection statement would then be *"(Given the observed data y,) the probability for the presence of a signal is p."* ($\mathrm{P}(H_1|y) = p$). The upper limit statement on the other hand is a matter of parameter estimation; given the joint posterior distribution of all unknowns in the model, one would need to marginalize to get the posterior distribution of the parameter of interest alone. The upper limit statement would then be *"(Given the observed data and the presence of a signal,) the amplitude is $\leq A^\star$ with 90% probability."* ($\mathrm{P}(A \leq A^\star \,|\, y, H_1) = 90\%$) [3,4].

2 The data model

We assume the data y to be a time series given by a parameterized signal s and additive noise n:

$$y(t_i) \;=\; s(t_i) + n(t_i), \tag{1}$$

where $i = 1, \ldots, N$ and $t_i = i\Delta_t$. The (sinusoidal) signal is given by

$$s(t) \;=\; A\,\sin(2\pi f t + \phi), \tag{2}$$

where $A \geq 0$ is the amplitude, $0 \leq \phi < 2\pi$ is the phase, and $f \in \{\frac{j_1}{N\Delta_t}, \ldots, \frac{j_k}{N\Delta_t}\}$ is the frequency, where $1 \leq j_1, \ldots, j_k \leq \frac{N}{2} - 1$ defines the range of possible (Fourier) frequencies. The number k of frequency bins may be varied and constitutes the so-called "trials factor" here. The noise n is assumed to be white and Gaussian with variance σ^2.

3 Frequentist approach

If there were no unknown parameters in the signal model, then, following from the Neyman-Pearson lemma, the optimal detection statistic would be given by the *likelihood ratio* of the two hypotheses. In the case that the hypotheses include unknowns (composite hypotheses) as in our case, this is commonly treated using the *generalized likelihood ratio* framework, that is, by considering the ratio of *maximized* likelihoods, where maximization is done over the unknown parameters [5].

In our case, we have a 3-dimensional parameter space under the signal model. The conditional likelihood for a given frequency may be maximized analytically over phase and amplitude. The *profile likelihood* (maximized conditional likelihood for given frequency, as a function of frequency) is eventually proportional to the time series' periodogram. The generalized likelihood ratio detection statistic then is given as the periodogram maximized over the frequency range of interest:

$$d^2 \;:=\; \max_j \tfrac{2}{N\sigma^2}\left|\tilde{y}_j\right|^2 \tag{3}$$

where $\tilde{y}_j$ is the (complex valued) jth element of the discretely Fourier transformed time series y. The "$\frac{2}{N\sigma^2}\left|\tilde{y}_j\right|^2$" term (the periodogram) maximized over in (3) is in fact also the *matched filter* for a sinusoidal signal [6], and the maximum d^2 is commonly referred to as the "loudest event" [2].

The detection statistic's distribution may be derived analytically under both hypotheses H_0 and H_1, as this is a particular case of an *extreme value statistic* [5]. Under the null hypothesis, d^2 is the maximum of k independently χ^2_2-distributed random variables; the cumulative distribution function (CDF) of d^2 is given by

$$F_{d^2;H_0}(x) \;=\; \mathrm{P}(d^2 \leq x \,|\, H_0) \;=\; \left(F_{\chi^2_2}(x)\right)^k \tag{4}$$

where $F_{\chi^2_2}$ is the CDF of a χ^2_2 distribution, and k again is the number of independent frequency bins, or "trials". This is essentially the "background distribution" of d^2. Under the signal hypothesis H_1, d^2 is the maximum of $(k–1)$ independently χ^2_2-distributed random variables *and* one noncentral-$\chi^2_2(\lambda)$-distributed variable with noncentrality parameter $\lambda = \frac{N}{2\sigma^2}A^2$. The corresponding CDF under H_1 then is

$$F_{d^2;H_1}(x) \;=\; \left(F_{\chi^2_2}(x)\right)^{(k-1)} \times F_{\chi^2_{2,\lambda}}(x) \tag{5}$$

where $F_{\chi^2_{2,\lambda}}$ is the CDF of a noncentral χ^2_2 distribution with parameter λ.

For some observed detection statistic value d_0^2, the (detection) significance is determined by the p-value $\mathrm{P}(d^2 \geq d_0^2 \,|\, H_0) = \int_{d_0^2}^{\infty} p(d^2|H_0)\,\mathrm{d}d^2$. The 90% loudest-event upper limit is given by the smallest amplitude value $A^\star$ for which $\int_{d_0^2}^{\infty} p(d^2 \,|\, A, H_1)\,\mathrm{d}d^2 \geq 90\%$, so that $\mathrm{P}(d^2 \geq d_0^2 \,|\, A \geq A^\star, H_1) \geq 90\%$.

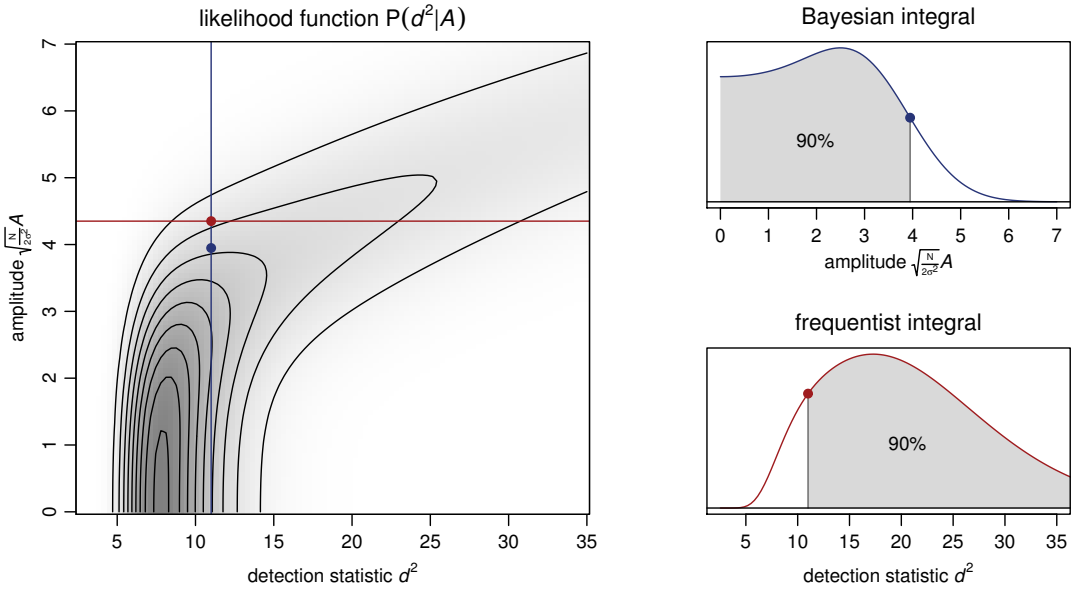


Fig. 1: The integrals to be computed for a frequentist and a Bayesian 90% upper limit are very different. The Bayesian integral is computed along the vertical amplitude axis, conditioning on the observed detection statistic value $d^2 = d_0^2$. The frequentist integral goes along the horizontal axis of possible realisations of d^2 for any given amplitude. (Example values here: $N = 100$, $\Delta_t = 1$, $\sigma^2 = 1$, $k = 49$, $d_0^2 = 11$.)

4 Bayesian approach

We assume uniform prior distributions on phase, frequency, and amplitude. Given the (3-dimensional) likelihood function [7], one can then derive joint and marginal posterior distributions $\mathrm{P}(A, \phi, f \,|\, y)$ and $\mathrm{P}(A|y)$. However, Monte Carlo simulations show that — in this particular model — the amplitude's marginal posterior distribution is virtually unaffected by whether one considers the complete data y, or only the "loudest event" d^2. The essential information about the signal amplitude is contained in that loudest event, and the marginal amplitude posterior is dominated by the conditional distribution of the loudest frequency bin. We find that the main difference between the two kinds of limits in this model is *not* due to maximization vs. integration of the posterior; in the following we will therefore consider only the simpler, directly comparable, and more illustrative case of a Bayesian loudest event limit based on $\mathrm{P}(A|d^2)$ instead of $\mathrm{P}(A|y)$.

Our relevant observable now is the "loudest event" d^2. The likelihood function $\mathrm{P}(d^2|A)$ was defined through (5) in the previous section. The 90% upper limit on the amplitude is given by the amplitude $A^\star$ for which $\int_0^{A^\star} p(A \,|\, d^2, H_1)\,\mathrm{d}A = 90\%$, so that $\mathrm{P}(A < A^\star \,|\, d^2, H_1) = 90\%$.

5 Comparison

The likelihood function here is a function of two parameters: the observable d^2 and the amplitude parameter A. Since the amplitude prior is assumed uniform, the posterior distribution is simply proportional to the likelihood, which allows for a nice comparison of both approaches. Fig. 1 illustrates the integrations performed for both the frequentist and the Bayesian upper limits for some particular realisation $d^2 = d_0^2$.

Since the data y are reduced to a single observable d^2, there also is a one-to-one mapping from d^2 to the upper limit $A^\star$. Fig. 2 shows both resulting upper limits as a function of the "loudest event" d^2. An important feature to note is that the frequentist limit will be *zero* for certain values of d^2. The point at (and below) which this happens is the lower 10% quantile of the "background" distribution of d^2 under H_0 (4) — at this point the probability of observing a larger d^2 value is (by definition) 90% for

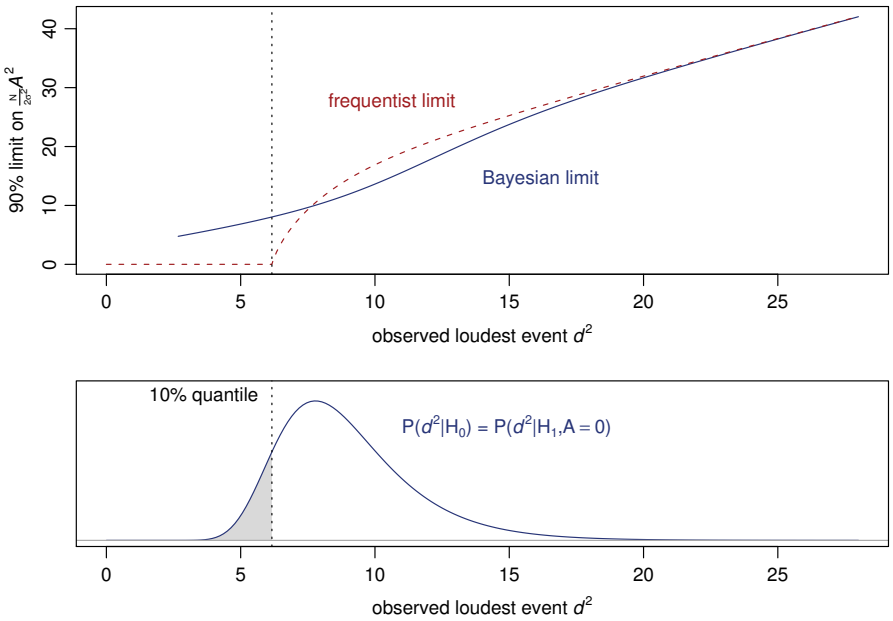


Fig. 2: The mapping from observable d^2 to the upper limit on amplitude. The bottom panel shows the "background" distribution of d^2 under H_0. (Example values here: $N = 100$, $\Delta_t = 1$, $\sigma^2 = 1$, $k = 49$.)

zero-amplitude signals already, which makes zero the 90% upper limit. Note that this implies that if H_0 in fact is true, 10% of all 90% upper limits will be zero. Note also that this is consistent with the intended 90% *coverage* of frequentist confidence bounds — if the upper limit is supposed to fall above and below the true amplitude value with 90% and 10% probabilities respectively, then 10% of the upper limits *must* be zero under H_0.

Having the distribution of the detection statistic (equations (4), (5)) and the mapping from d^2 to upper limit (Fig. 2) allows us to derive the distribution of upper limits for given parameters. Figure 3 illustrates the behaviour of the resulting upper limits for different values of amplitude A and trials factor k. The left panel shows that for large amplitudes the two limits behave roughly the same, as one could already see from Fig. 2, while for low amplitudes the posterior upper limit will level off and will not rule out amplitude values below a certain noise level. The frequentist limit's distribution on the other hand reaches all the way down to zero, and in particular the 90% limit's 10% quantile follows a straight line of slope 1 and intercept 0 — the frequentist 90% limit is (by construction) essentially a statistic that has its 10% quantile at the true amplitude value.

The right panel of Fig. 3 shows the differing behaviour of both limits as a function of the trials factor k when the true amplitude is zero. The frequentist limit's 10% quantile remains at zero (the true value), while the posterior limit is bounded away from zero but otherwise tends to yield tighter constraints on the amplitude, especially for large k.

6 Conclusions

The most obvious technical difference between frequentist vs. Bayesian upper limits is in maximization vs. integration over parameter space. This, however, is not — at least in the example discussed here — the primary origin of discrepancies between the two. When basing *both* limits on maximization (i.e., the "loudest event"), the behaviour of the Bayesian limit is affected very little; so the crucial information about the signal amplitude is in fact contained in the loudest event. Both kinds of upper limits behave very similarly for "loud" signals, i.e., a large signal-to-noise ratio (SNR), but their differences become apparent in the interesting case of (near-) zero amplitude signals. While the Bayesian upper limit expresses

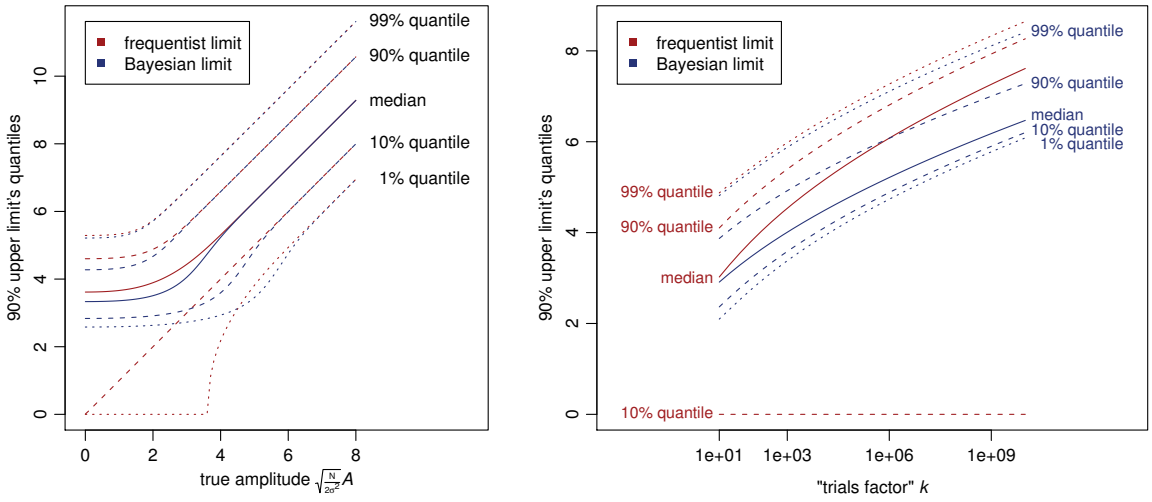


Fig. 3: The distribution of upper limits as a function of amplitude (left panel) and trials factor (for zero amplitude; right panel). Note that the frequentist 90% limit is essentially a statistic that is designed to have its 10% quantile at the true amplitude value.

what amplitude values may be ruled out with 90% certainty based on the data (and model assumptions), the frequentist upper confidence limit is defined solely through its "coverage" property. The frequentist 90% limit needs to end up above and below the true amplitude value with 90% and 10% probability respectively, which simply means that the frequentist limit may be any random variable that has its 10% quantile at the true amplitude. This in particular implies that for a true amplitude of $A = 0$ the limit has a 10% chance of being zero as well, and it makes the frequentist limit very hard to actually interpret, not only if it actually happens to turn out as zero. When considering the effect of the trials factor (or look-elsewhere effect) in the low-SNR regime where both limits behave differently, the posterior-based limit will usually yield tighter constraints especially for large trials factors, but it will never be zero.

The Bayesian upper limit based on the amplitude's posterior distribution will of course change with changing prior assumptions. For simplicity, we assumed an (improper) uniform amplitude prior here, but this should actually be a conservative choice in some sense, for a realistic prior in the continuous gravitational-wave context would in general be much more concentrated towards low amplitude values (something like the — also improper — prior with density $p(A) \propto \frac{1}{A^4}$).

Another question is how exactly one would do the actual computations for a Bayesian upper limit in practice — the frequentist upper limits are usually not computed via direct analytical or numerical integration of the likelihood, but the integral (see Fig. 1) is determined in a nonparametric fashion via Monte Carlo integration and bootstrapping of the data. While the frequentist limit requires finding the amplitude $A^\star$ at which the integral $(\mathrm{P}(d^2 > d_0^2 \,|\, A = A^\star))$ yields the desired confidence level, an analogous procedure to derive the Bayesian upper limit would probably require Monte Carlo sampling of $\mathrm{P}(d^2|A)$ across the range of all amplitudes A in order to then do the integral in the orthogonal direction.

Further complications arise especially for the frequentist limit when the signal model gets more complex. The general procedure required for the Bayesian upper limit is rather obvious — determine the marginal posterior distribution of amplitude $\mathrm{P}(A|y)$, then determine the 90% quantile. The frequentist procedure on the other hand may run into major problems. For example, if there are multiple parameters affecting the signal's SNR, a "loudest event" might be hard to define, or to translate into a constraint on the amplitude. As there may not be a simple one-to-one connection between SNR and amplitude parameter as in the present case, the "loudest event" may not be the only relevant figure to constrain the signal amplitude. The consideration of nuisance parameters is generally tricky in a frequentist framework and may effectively suggest the use of a Bayesian procedure instead [8]. Computation also becomes more

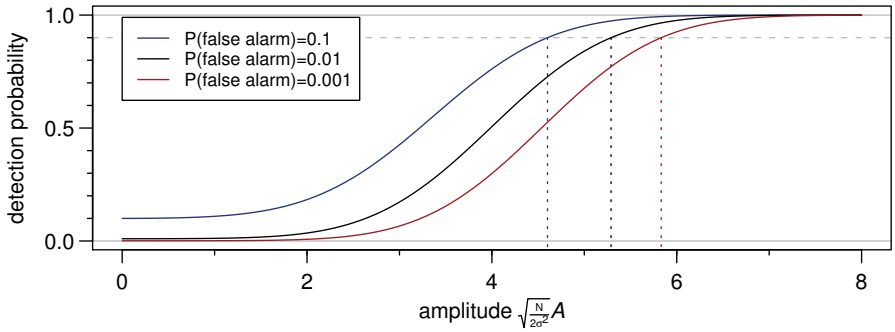


Fig. 4: Illustration of the determination of a 90% *detection sensitivity* threshold. Such a statement would be independent of the observed data, and it requires the specification of an additional parameter: the corresponding false alarm rate defining the threshold of what is considered a "detection". (Here: $N = 100$, $\Delta_t = 1$, $\sigma^2 = 1$, $k = 49$.)

complicated if the frequency parameter is not restricted to ("independent") Fourier frequencies. Note that the reasoning behind the generalized likelihood ratio approach (see Sec. 3) leading to the "loudest event" concept was very much an ad-hoc construction in the first place.

Another notable related concept is that of a *power constrained upper limit*. In search experiments, these may be based on the *sensitivity* of the search procedure. In case the search yielded no detection, one can state the signal amplitude that would have been detected with 90% probability; this number may then also be used as a lower bound on the frequentist limit (*"don't rule out what you wouldn't be able to detect"*). However, this kind of statement requires the specification of another, additional parameter: the corresponding false alarm rate defining the threshold of what is considered a "detection", and as such is inseparably connected to the detection procedure (see also Fig. 4). In particle physics a different approach is commonly taken; there the sensitivity is usually specified as the expected upper limit for many repetitions of the experiment in the absence of a signal. This figure would correspond to the solid lines at zero amplitude in Fig. 3. An important point to note is that both these sensitivity statements do not depend on the observed data.

References

[1] B. Abbott et al. Setting upper limits on the strength of periodic gravitational waves from PSR J1939+2134 using the first science data from the GEO 600 and LIGO detectors. *Physical Review D*, 69(8):082004, April 2004.

[2] P. Brady, J. D. E. Creighton, and A. G. Wiseman. Upper limits on gravitational-wave signals based on loudest events. *Classical and Quantum Gravity*, 21(20):S1775–1781, October 2004.

[3] A. Gelman, J. B. Carlin, H. Stern, and D. B. Rubin. *Bayesian data analysis*. Chapman & Hall / CRC, Boca Raton, 1997.

[4] B. P. Abbott et al. Searches for gravitational waves from known pulsars with science run 5 LIGO data. *The Astrophysical Journal*, 713(1):671–685, April 2010.

[5] A. M. Mood, F. A. Graybill, and D. C. Boes. *Introduction to the theory of statistics*. McGraw-Hill, New York, 3rd edition, 1974.

[6] L. A. Wainstein and V. D. Zubakov. *Extraction of signals from noise*. Prentice-Hall, Englewood Cliffs, NJ, 1962.

[7] G. L. Bretthorst. *Bayesian spectrum analysis and parameter estimation*, volume 48 of *Lecture Notes in Statistics*. Springer-Verlag, Berlin, 1988.

[8] A. C. Searle. Monte-Carlo and Bayesian techniques in gravitational wave burst data analysis. *Arxiv preprint 0804.1161 [gr-qc]*, April 2008.

Multichannel number counting experiments

V. Zhukov[*], *M. Bonsch*
KIT, Institute für Experimentelle Kernphysik, Universität Karlsruhe

Abstract
The confidence intervals calculated with different statistical methods for the combined model of number counting experiments have been compared. The Bayesian approach with flat prior provides the most conservative limit in a vast range of model parameters. The limits calculated with the Feldman-Cousins method are sensitive to the systematics. The results of Hybrid CLs calculations for the combined model can be the most optimistic but tend to undercover.

1 Introduction

The search for new phenomena in HEP is often reduced to the combination of many number counting experiments with different observed statistics, background expectations, signal sensitivity and systematic uncertainties [1]. Although such combination in one global fit is attractive from the physics perspective, it is challenging from the statistical point of view [2]. The combination of very different channels becomes sensitive to the range of the parameter of interest, systematic uncertainties and correlations, hence resulting in different predictions from different statistical methods.

In this study, the upper confidence limits (95% CL) have been calculated with the Profile Likelihood Calculator(PLC), Feldman-Cousins(FC) [9], Hybrid(using LR statistic and CLs ratio) [10] and Bayesian (with flat prior) methods. The modeling is done with the `RooStats` package of `Root` v5.28 [3]. For Bayesian calculations, the Markov Chain MC implemented in `BAT` has been used [4].

2 Statistical model

The combination of many(N_{ch}) exclusive channels can be written as a product of individual Likelihood functions $L = \prod_k^{Nch} L_k$ where each function may have two components: 1) the statistical term $Poiss(n_k|\mu_k)$ and 2) the systematics term $\prod_i^{Nnuis} G(\delta_i|\delta_0^i, \sigma_\delta^i)$ which is a set of pdf's for the vector of nuisances parameters $\vec{\delta}$ with mean $\vec{\delta}_0$ and $\vec{\sigma}_\delta$ affecting the expectation value $\mu_k(\vec{\delta})$. There are usually a few common sources of systematic uncertainties for all channels which can be factorized as $(1 + f_k\delta_i)\mu$, where δ_i is the i-nuisance and f_k is the scaling factor for the k-channel depending on the amplitude of variations. For example the background part for k-channel can be written as: $B_k = b\prod_i^{Nnuis}(1 + f_k^i\delta_i)$ where b is the total background in this channel. Similarly the signal part is: $S_k = s\nu_k\prod_i^{Nnuis}(1 + f_k^i\delta_i)$, here s is a common signal strength for all search channels and ν_k is the signal yield for k-channel. Such factorization of systematics allows to account for correlations in a simple way, otherwise the full covariance matrix has to be evaluated and the limit calculation becomes difficult for complicated models.

Although the shape of systematics pdf's can almost always be transformed into the standard distributions by replacement of variables, it is common to keep natural observables in the statistical model. Then the distribution of systematics usually is a compromise between a simple analytical form ensuring good performance of statistical methods and the most realistic distribution of uncertainties describing the data which in turn depends on the nature of uncertainties and can be roughly divided into two main categories [5].

First are the systematic uncertainties originating from statistical errors in some auxiliary measurements. The second type of systematic uncertainties is related to missing or incomplete knowledge, for

[*]On leave from SINP MSU

example model uncertainties. With the only statistical errors from auxiliary measurements the combined Likelihood can be written as: $Poiss(n|s+\tilde{c}\tau)Poiss(n_c|\tilde{c})$. Here the systematics term is replaced by an extra Poisson for control measurement n_c with an expectation $\tilde{c}$ treated as a nuisance parameter. The relation between control region c and the background in the signal region b is defined by the $\tau = b/c$ factor which can have its own uncertainty, usually of the second type. Without this uncertainty the model with auxiliary measurements is equivalent to the simple Poisson model with the Gamma distributed systematics on the background, where the Gamma skewness and σ_Γ are defined by the τ factor: $Poiss(n|s+b')Gamma(b'|b,\tau)$ [7]. The uncertainty in the τ factor introduces yet another nuisance parameter which breaks this equivalence. However if this τ uncertainty can be described by Gaussian with σ_τ and convoluted with the Gamma distribution, the resulting distribution will still have Gamma like shape with $\sigma_b^2 \approx \sigma_\Gamma^2 + \sigma_\tau^2$. Such distribution can be used in the systematics term of the Likelihood function instead of the full form with two Poissons. Here different systematics pdf's have been used for the comparison.

The use of a Gamma distribution introduces a relation between the standard deviation and the mean value which complicates the factorization of correlated systematic uncertainties in multichannel case. Some simplification can be achieved with a Lognormal distribution which has a similar shape and allows factorization when it is written as: $(1+\sigma_k)^{\delta_i}$, where σ_k is the relative uncertainty for k-channel and δ_i is the normally distributed i- nuisance parameter.

The combination of identical channels with the same signal to background ratio (S/B) without systematics uncertainties is equivalent to the splitting of Poisson statistics in one search; this property was used to verify the accuracy for the combined model. For channels with different S/B or different systematics the results depend upon observed statistics, especially in the limiting case when some channels have zero observations or the range of systematics is very different.

The results also depend upon the internal accuracy of the method and its implementation. There is no unique algorithm for estimating these internal uncertainties. For Profile Likelihood the intrinsic errors are related to the maximization of likelihood ratio(LR). The method based on Neyman construction, like FC, depends upon binning in the scan of parameter of interest and on the treatment of nuisance parameters. Bayesian integration uses either numerical integration or Markov Chain MC, in both cases the accuracy degrading fast with increased dimensionality. The Hybrid limits are estimated from MC toy experiments and the accuracy depends on the number of these toys which is tuned to have 0.5% accuracy in limits. However this does not guarantee that the whole range of nuisance parameter is explored.

3 Confidence limits for single channel and the combined model

The single number counting experiment with the anticipated relative systematic on the background σ_b has been used as a reference. The 95% CL upper limits versus expected backgrounds (N_{bkg}) and observations (N_{obs}) calculated with different methods are shown in Fig. 1. On this and the following plots only relatively small values of $N_{obs} - N_{bkg}$ are important for upper limits, for larger values the confidence belt becomes two sided but study of the flip-flop problem [9] is beyond the scope of this paper. The Bayesian approach produces the most conservative limits for a single channel. The Hybrid CLs limit is not defined for $N_{bkg} >> N_{obs}$ where $CL_b \to 0$. It is important to notice the behavior at $N_{obs} \approx 0$. The Bayesian and Hybrid CLs limits are independent of the background while the FC limits improve with larger background even for larger N_{obs}. The PLC fails at zero observation when the Likelihood ratio is collapsed to $lnQ = s$ and Wilk's theorem is not valid. The Hybrid CLs limits should be similar to the Bayesian credible intervals for this simple model and the difference can be related to the accuracy of implementations.

The systematics pdf have different effect in different methods, see Fig. 2. The FC is the most sensitive to the shapes because the nuisance parameter minimization is performed for each value of the scanned parameter of interest.

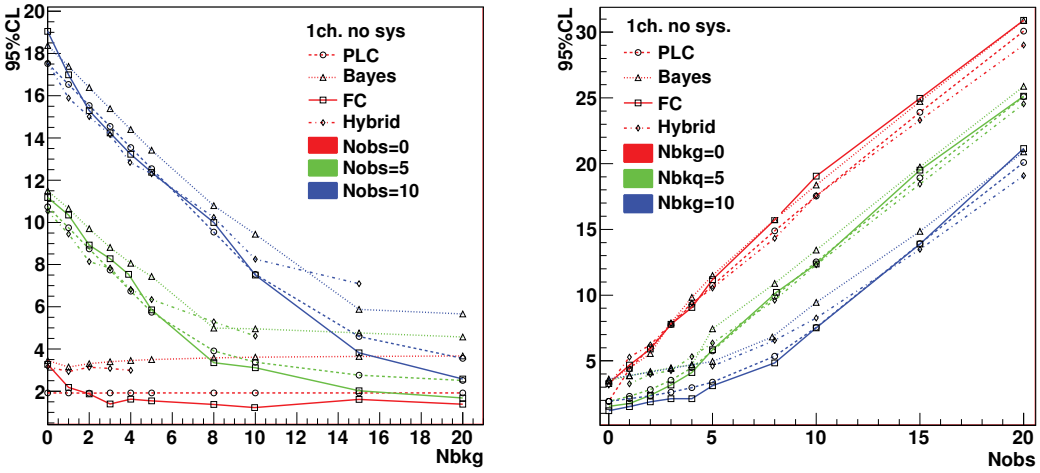


Fig. 1: 95% CL upper limits versus background and observations for single channel calculated with different methods.

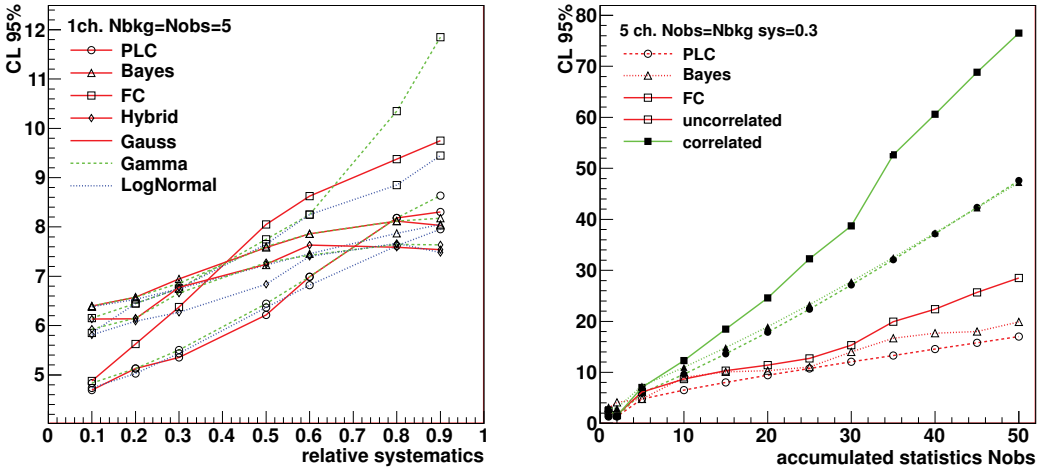


Fig. 2: Left: 95% CL upper limits for single channel versus relative systematics on background (σ_b) with different systematics pdf's. Right: limits versus accumulated statistics ($N_{obs} = N_{bkg}$) for the combined model of five channels with correlated and uncorrelated systematics $\sigma_b = 0.3$

There are some differences in limits even for this trivial case and the question is - which method to use? The obvious flaws in modeling can be spotted with the coverage test [9]. Figure 5 shows results of the coverage test for a single channel with relative systematics uncertainties on the background σ_b=0.3. All methods show no undercoverage with some overcoverage especially for the PLC. This coverage test can be done relatively easily for the simple model but becomes increasingly difficult or impossible for the multichannel search with many nuisance parameters when one has to guarantee no undercoverage for all possible combinations of nuisance parameters. The FC method should have correct coverage, or at least no undercoverage by construction. But for the PLC and Hybrid CLs methods coverage is not guaranteed and has to be checked. For the Bayesian credible limits the frequentist coverage test does not have a clear statistical interpretation and depends on the prior.

The combined model, here with five channels and uncorrelated Lognormal systematics with relative error $\sigma_b = 0.3$, is shown in Fig. 3. The Bayesian limit remains the most conservative while the

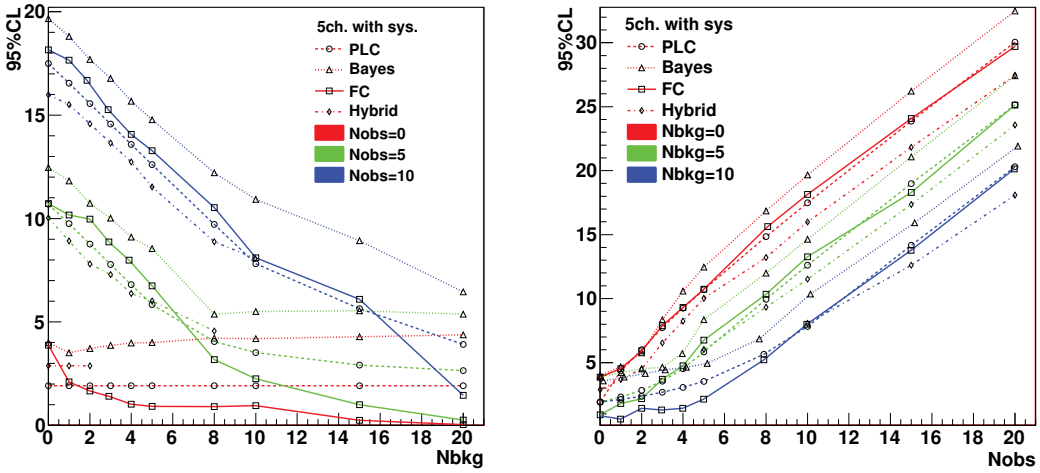


Fig. 3: 95% CL upper limits for the combined model of five identical channels with uncorrelated Lognormal systematics σ_b=0.3

Hybrid CLs becomes even more optimistic than PLC, which apparently is too low for small N_{obs}.

The effect of correlations is demonstrated in Fig. 2 at different statistics. The correlations reduce the number of independent nuisance parameters and usually degrade the limits. However in some cases the combination can be beneficial, that is, some channels with large systematics and low signal sensitivity can serve as a control measurement for the channels with higher sensitivity.

In reality the channels can have different S/B, different observed statistics and different ranges of systematics. Two typical cases are considered. The first model has one channel with five times higher signal sensitivity than the other four but zero observation. The second model with correlated uncertainties has one channel with three times larger systematics. The upper limits calculated for these two cases are shown in Fig. 4 versus difference N_{obs} minus N_{bkg}: $\Delta = N_{obs} - N_{bkg}$. The biggest difference is between the Bayesian and the PLC, FC and Hybrid limits for the first model. The Hybrid limits are the most optimistic at relatively large Δ. For smaller Δ the FC limits are very low, comparable with PLC.

In spite of large differences in the calculated limits, all methods, except the Hybrid CLs, show relatively good coverage, see Fig. 5.

4 Summary

The upper limits calculated with different statistical methods for the model comprised of several number counting experiments can have large variations depending on model parameters.

The Bayesian intervals with flat prior are the most conservative limits for all considered parameters, especially at high background expectation and combined models with very different channels.

The Feldman-Cousins upper limits are close to the Bayesian at small background but are getting more optimistic for larger background expectations. The results are also dependent on the steps in the scan of the parameter of interest and distribution of systematics which can result in empty intervals for some configurations.

The Profile Likelihood delivers the most optimistic limits for low observations and has to be avoided with zero or small observation in some of the channels.

The results of Hybrid CLs calculations are difficult to predict. While for a simple model it is rather similar to the conservative Bayesian limits, for the combined models the Hybrid limits become the most optimistic with some non zero observations. For low observation and downward fluctuation the Hybrid

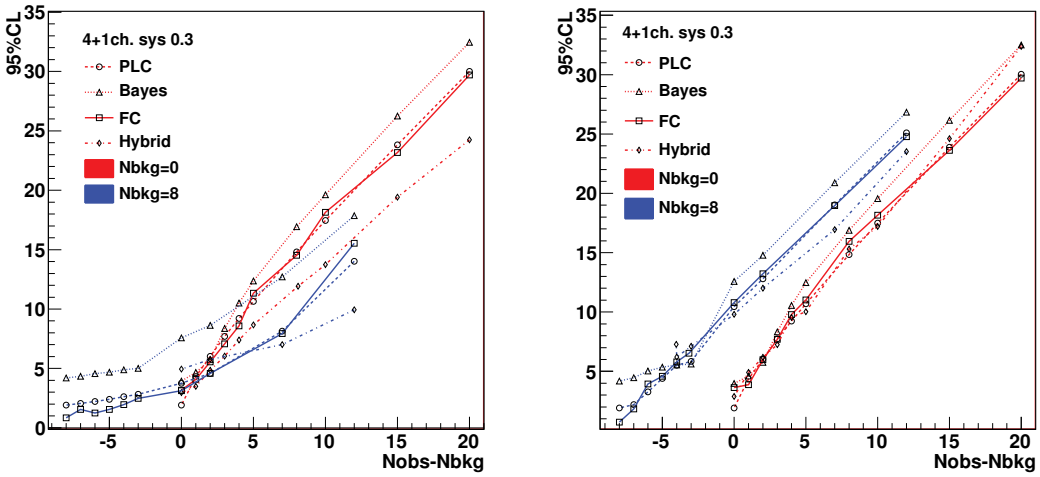


Fig. 4: Left: Upper limits versus $Nobs - Nbkg$ for the combined model of five channels with one five times more sensitive but with zero observation. All channels have Lognormal uncorrelated systematics with σ_b=0.3. Right: Combined model with correlated systematics σ_b=0.3 and one channel with σ_b=0.9

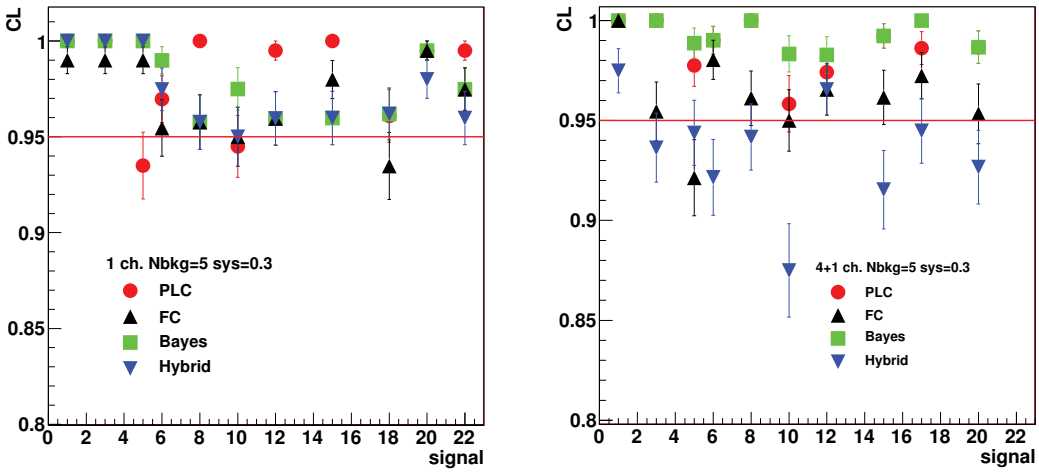


Fig. 5: Coverage of 95% CL upper limits calculated with different methods for single channel (σ_b=0.3) and combined model with five channels where one has five times higher signal sensitivity and zero observation (N_{MCtoys}=500)

limits are protected by the CLs ratio. Moreover the Hybrid method tends to undercover for combined models of very different channels.

The final choice of statistical method remains rather subjective. Apart from the conservative Bayesian limits, the usage of frequentist methods always have some drawbacks. For complicated models with many different channels the Feldman-Cousins limits looks the most attractive due to its intrinsic coverage.

In conclusion, the RooFit/RooStats package offers an excellent tool for the statistic modeling in LHC analysis.

5 Acknowledgments

The authors thank the PHYSTAT2011 Organizers. VZ thanks R.Cousins, J.Linnemann and L.Lyons for the inspiring discussions and help.

References

[1] CMS PAS SUS-10-008, 2011
[2] K.Cranmer, PhysStat2007, 261, 2007
[3] L.Moneta et al. arXiv:1009.1003.438, 2010
[4] A. Caldwell, D. Kollar, K. Kroeninger, Comp. Phys. Comm. 180, 2197, 2009
[5] P.Sinervo, PhyStat2003,122,2003
[6] J.Heinrich, PhyStat2003,52 , 2003
[7] J. Linnemann arXiv:ph/0312059, 2003
[8] T.Junk, arXiv:hep-ex/9902006, 1999
[9] G.J. Feldman and R.D. Cousins, Phys.Rev. D 57, 3873, 1998
[10] R.D. Cousins and V.L. Highland, NIM A320, 331, 1992

Statistical Challenges of Global SUSY Fits

Roberto Trotta[1] *& Kyle Cranmer*[2]
[1]Imperial College London. [2]New York University

Abstract
We present recent results aiming at assessing the coverage properties of Bayesian and frequentist inference methods, as applied to the reconstruction of supersymmetric parameters from simulated LHC data. We discuss the statistical challenges of the reconstruction procedure, and highlight the algorithmic difficulties of obtaining accurate profile likelihood estimates.

1 Introduction

Experiments at the Large Hadron Collider (LHC) have already started testing many models of particle physics beyond the Standard Model (SM), and particular attention is being paid to the Minimal Supersymmetric SM (MSSM) and to other scenarios involving softly-broken supersymmetry (SUSY).

In the last few years, parameter inference methodologies have been developed, applying both Frequentist and Bayesian statistics (see e.g., [1–6]). While the efficiency of Markov Chain Monte Carlo (MCMC) techniques has allowed for a full exploration of multidimensional models, the likelihood function from present data is multimodal with many narrow features, making the exploration task with conventional MCMC methods challenging. A powerful alternative to classical MCMC has emerged in the form of Nested Sampling [7], a Monte Carlo method whose primary aim is the efficient calculation of the Bayesian evidence, or model likelihood. As a by-product, the algorithm also produces samples from the posterior distribution. Those same samples can also be used to estimate the profile likelihood. MULTINEST [8], a publicly available implementation of the nested sampling algorithm, has been shown to reduce the computational cost of performing Bayesian analysis typically by two orders of magnitude as compared with basic MCMC techniques. MULTINEST has been integrated in the `SuperBayeS` code[1] for fast and efficient exploration of SUSY models.

Having implemented sophisticated statistical and scanning methods, several groups have turned their attention to evaluating the sensitivity to the choice of priors [4, 9, 10] and of scanning algorithms [11]. Those analyses indicate that current constraints are not strong enough to dominate the Bayesian posterior and that the choice of prior does influence the resulting inference. While confidence intervals derived from the profile likelihood or a chi-square have no formal dependence on a prior, there is a sampling artifact when the contours are extracted from samples produced from Bayesian sampling schemes, such as MCMC or MULTINEST [10].

Given the sensitivity to priors and the differences between the intervals obtained from different methods, it is natural to seek out a quantitative assessment of their performance, namely their *coverage*: the probability that an interval will contain (cover) the true value of a parameter. The defining property of a 95% confidence interval is that the procedure adopted for its estimation should produce intervals that cover the true value 95% of the time; thus, it is reasonable to check if the procedures have the properties they claim. While Bayesian techniques are not designed with coverage as a goal, it is still meaningful to investigate their coverage properties. Moreover, the intervals obtained from the profile likelihood or chi-square functions are based on asymptotic approximations and are not guaranteed to have the claimed coverage properties.

Here we report on recent studies investigating the coverage properties of both Bayesian and Frequentist procedures commonly used in the literature. We also highlight the numerical and sampling challenges that have to be met in order to obtain a sufficienlty high-resolution mapping of the profile

[1]Available from: `www.superbayes.org`

likelihood when adopting Bayesian algorithms (which are typically designed to map out the posterior mass, instead).

For the sake of example, we consider in the following the so-called mSUGRA or Constrained Minimal Supersymmetric Standard Model (CMSSM), a model with fairly strong universality assumptions regarding the SUSY breaking parameters, which reduce the number of free parameters to be estimated to just five, denoted by the symbol $\mathbf{\Theta}$: common scalar (m_0), gaugino ($m_{1/2}$) and tri–linear (A_0) mass parameters (all specified at the GUT scale) plus the ratio of Higgs vacuum expectation values $\tan\beta$ and sign(μ), where μ is the Higgs/higgsino mass parameter whose square is computed from the conditions of radiative electroweak symmetry breaking (EWSB).

2 Coverage study of the CMSSM

2.1 Accelerated inference from neural networks

Coverage studies require extensive computational expenditure, which would be unfeasible with standard analysis techniques. Therefore, in Ref. [12] a class of machine learning devices called Artificial Neural Networks (ANNs) was used to approximate the most computationally intensive sections of the analysis pipeline.

Inference on the parameters of interest $\mathbf{\Theta}$ requires relating them to observable quantities, such as the sparticle mass spectrum at the LHC, denoted by $\mathbf{m}$, over which the likelihood is defined. This is achieved by evolving numerically the Renormalization Group Equations (RGEs) using publicly available codes, which is however a computationally intensive procedure. One can view the RGEs simply as a mapping from $\mathbf{\Theta} \rightarrow \mathbf{m}$, and attempt to engineer a computationally efficient representation of the function. In [12], an adequate solution was provided by a three-layer perceptron, a type of feed-forward neural network consisting of an input layer (identified with $\mathbf{\Theta}$), a hidden layer and an output layer (identified with the value of $\mathbf{m}(\mathbf{\Theta})$ that we are trying to approximate). The weight and biases defining the network were determined via an appropriate training procedure, involving the minimization of a loss function (here, the discrepancy between the value of $\mathbf{m}(\mathbf{\Theta})$ predicted by the network and its correct value obtained by solving the RGEs) defined over a set of 4000 training samples. A number of tests on the accuracy and noise of the network were performed, showing a correlation in excess of 0.9999 between the approximated value of $\mathbf{m}(\mathbf{\Theta})$ and the value obtained by solving the RGEs for an independent sample. A second classification network was employed to distinguish between physical and un-physical points in parameter space (i.e., values of $\mathbf{\Theta}$ that do not lead to physically viable solutions to the RGEs). The final result of replacing the computationally expensive RGEs with the ANNs is presented in Fig. 1, which shows that the agreement between the two methods is excellent, within numerical noise. By using the neural network, a speed-up factor of about 3×10^4 compared with scans using the explicit spectrum calculator was observed.

2.2 Coverage results for the ATLAS benchmark

We studied the coverage properties of intervals obtained for the so-called "SU3" benchmark point. To this end, we need the ability to generate pseudo-experiments with $\mathbf{\Theta}$ fixed at the value of the benchmark. We adopted a parabolic approximation of the log-likelihood function (as reported in Ref. [13]), based on the measurement of edges and thresholds in the invariant mass distributions for various combinations of leptons and jets in final state of the selected candidate SUSY events, assuming an integrated luminosity of 1 fb^{-1} for ATLAS. Note that the relationship between the sparticle masses and the directly observable mass edges is highly non-linear, so a Gaussian is likely to be a poor approximation to the actual likelihood function. Furthermore, these edges share several sources of systematic uncertainties, such as jet and lepton energy scale uncertainties, which are only approximately communicated in Ref. [13]. Finally, we introduce the additional simplification that the likelihood is also a multivariate Gaussian with the same covariance structure. We constructed 10^4 pseudo-experiments and analyzed them with both MCMC

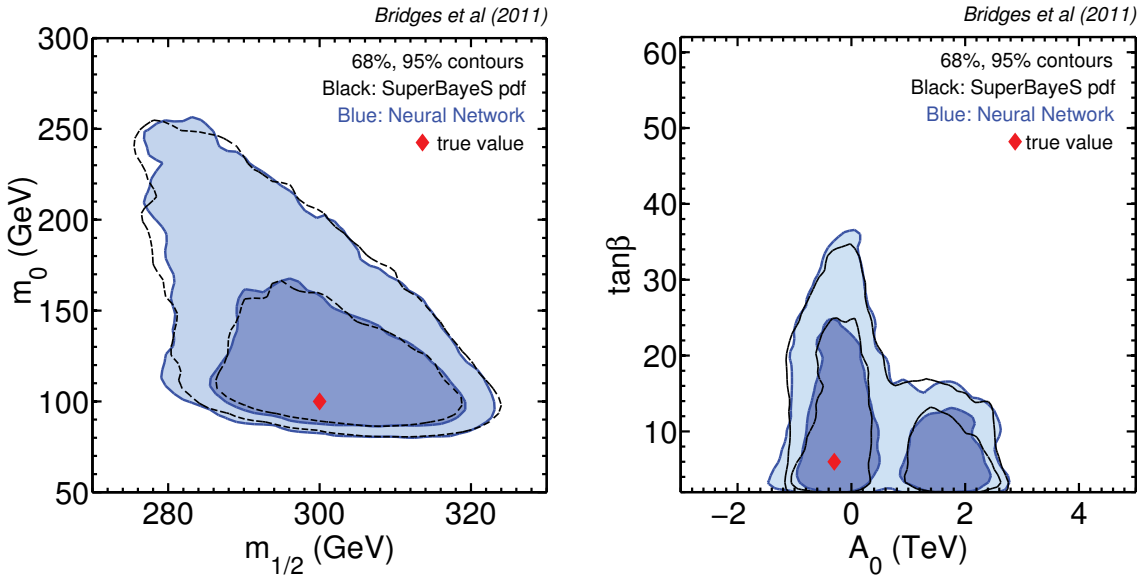


Fig. 1: Comparison of Bayesian posteriors obtained by solving the RGEs fully numerically (black lines, giving 68% and 95% regions) and neural networks (blue lines and corresponding filled regions), from simulated ATLAS data. The red diamond gives the true value for the benchmark point adopted. From [12].

(using a Metropolis-Hastings algorithm) and MULTINEST. Altogether, our neural network MCMC runs have performed a total of 4×10^{10} likelihood evaluations, in a total computational effort of approximately 2×10^4 CPU-minutes. We estimate that the solving the RGEs fully numerically would have taken about 1100-CPU years, which is at the boundary of what is feasible today, even with a massive parallel computing effort.

The results are shown in Fig. 2, where it can be seen that the methods have substantial over-coverage for 1-d intervals, which means that the resulting inferences are conservative. While it is difficult to unambiguously attribute the over-coverage to a specific cause, the most likely cause is the effect of boundary conditions imposed by the CMSSM. When $\mathbf{\Theta}$ is composed of parameters of interest, θ, and nuisance parameters, ψ, the profile likelihood ratio is defined as

$$\lambda(\theta) \equiv \frac{\mathcal{L}(\theta, \hat{\hat{\psi}})}{\mathcal{L}(\hat{\theta}, \hat{\psi})}. \tag{1}$$

where $\hat{\hat{\psi}}$ is the conditional maximum likelihood estimate (MLE) of ψ with θ fixed and $\hat{\theta}, \hat{\psi}$ are the unconditional MLEs. When the fit is performed directly in the space of the weak-scale masses (i.e., without invoking a specific SUSY model and hence bypassing the mapping $\mathbf{\Theta} \rightarrow \mathbf{m}$), there are no boundary effects, and the distribution of $-2 \ln \lambda(\mathbf{m})$ (when $\mathbf{m}$ is true) is distributed as a chi-square with a number of degrees of freedom given by the dimensionality of $\mathbf{m}$. Since the likelihood is invariant under reparametrizations, we expect $-2 \ln \lambda(\theta)$ to also be distributed as a chi-square. If the boundary is such that $\mathbf{m}(\hat{\theta}, \hat{\psi}) \neq \hat{\mathbf{m}}$ or $\mathbf{m}(\theta, \hat{\hat{\psi}}) \neq \hat{\hat{\mathbf{m}}}$, then the resulting interval will modified. More importantly, one expects the denominator $\mathcal{L}(\hat{\theta}, \hat{\psi}) < \mathcal{L}(\hat{\mathbf{m}})$ since $\mathbf{m}$ is unconstrained, which will lead to $-2 \ln \lambda(\theta) < -2 \ln \lambda(\mathbf{m})$. In turn, this means more parameter points being included in any given contour, which leads to over-coverage.

The impact of the boundary on the distribution of the profile likelihood ratio is not insurmountable. It is not fundamentally different than several common examples in high-energy physics where an unconstrained MLE would lie outside of the physical parameter space. Examples include downward

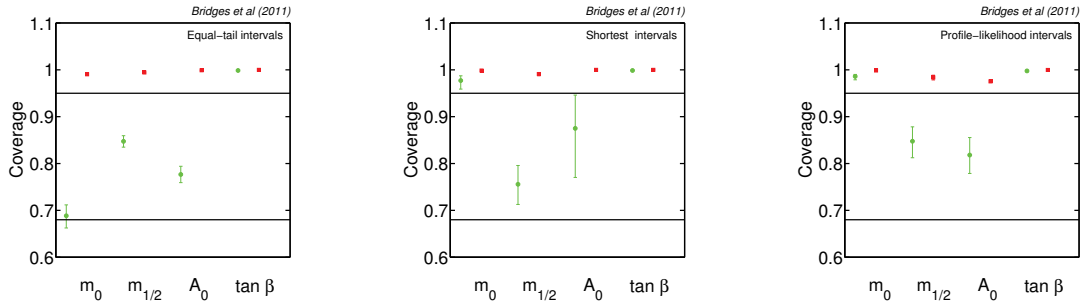


Fig. 2: Coverage for various types of intervals for the CMSSM parameters, from 10^4 realizations, employing MCMC for the reconstruction (each pseudo-experiment is reconstructed with 10^6 samples). Green/circles (red/squares) is for the 68% (95%) error. From [12].

fluctuations in event-counting experiments when the signal rate is bounded to be non-negative. Another common example is the measurement of sines and cosines of mixing angles that are physically bounded between $[-1, 1]$, though an unphysical MLE may lie outside this region. The size of this effect is related to the probability that the MLE is pushed to a physical boundary. If this probability can be estimated, it is possible to estimate a corrected threshold on $-2 \ln \lambda$. For a precise threshold with guaranteed coverage, one must resort to a fully frequentist Neyman Construction. A similar coverage study (but without the computational advantage provided by ANNs) has been carried out for a few CMSSM benchmark points for simulated data from future direct detection experiments [14]. Their findings indicate substantial under-coverage for the resulting intervals, especially for certain choices of Bayesian priors. Both works clearly show the timeliness and importance of evaluating the coverage properties of the reconstructed intervals for future data sets.

3 Challenges of profile likelihood evaluation

For highly non-Gaussian problems like supersymmetric parameter determination, inference can depend strongly on whether one chooses to work with the posterior distribution (Bayesian) or profile likelihood (frequentist) [4, 10, 15]. There is a growing consensus that both the posterior and the profile likelihood ought to be explored in order to obtain a fuller picture of the statistical constraints from present-day and future data. This begs the question of the algorithmic solutions available to reliably explore both the posterior and the profile likelihood in the context of SUSY phenomenology.

The profile likelihood ratio defined in Eq. (1) is an attractive choice as a test statistic, for under certain regularity conditions, Wilks [16] showed that the distribution of $-2 \ln \lambda(\theta)$ converges to a chi-square distribution with a number of degrees of freedom given by the dimensionality of θ. Clearly, for any given value of θ, evaluation of the profile likelihood requires solving a maximisation problem in many dimensions to determine the conditional MLE $\hat{\hat{\psi}}$. While posterior samples obtained with MULTINEST have been used to estimate the profile likelihood, the accuracy of such an estimate has been questioned [11]. As mentioned above, evaluating profile likelihoods is much more challenging than evaluating posterior distributions. Therefore, one should not expect that a vanilla setup for MULTINEST (which is adequate for an accurate exploration of the posterior distribution) will automatically be optimal for profile likelihoods evaluation. In Ref. [17] the question of the accuracy of profile likelihood evaluation from

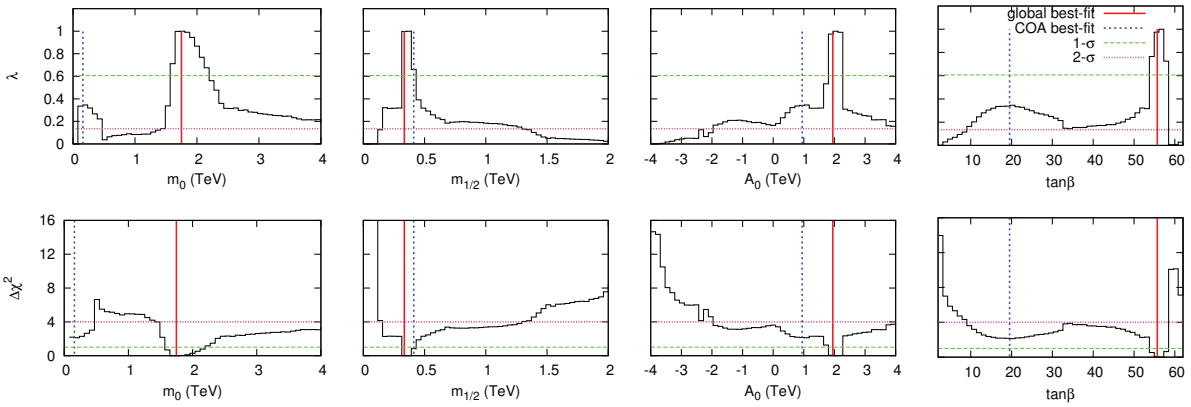


Fig. 3: 1-D profile likelihoods from present-day data for the CMSSM parameters normalized to the global best-fit point. The red solid and blue dotted vertical lines represent the global best-fit point ($\chi^2 = 9.26$, located in the focus point region) and the best-fit point found in the stau co-annihilation region ($\chi^2 = 11.38$) respectively. The upper and lower panel show the profile likelihood and $\Delta\chi^2$ values, respectively. Green (magenta) horizontal lines represent the 1σ (2σ) approximate confidence intervals. MULTINEST was run with 20,000 live points and $\mathrm{tol} = 1 \times 10^{-4}$ (a configuration deemed appropriate for profile likelihood estimation), requiring approximately 11 million likelihood evaluations. From [17].

MULTINEST was investigated in detail. We report below the main results.

The two most important parameters that control the parameter space exploration in MULTINEST are the number of live points n_{live} – which determines the resolution at which the parameter space is explored – and a tolerance parameter tol, which defines the termination criterion based on the accuracy of the evidence. Generally, a larger number of live points is necessary to explore profile likelihoods accurately. Moreover, setting tol to a smaller value results in MULTINEST gathering a larger number of samples in the high likelihood regions (as termination is delayed). This is usually not necessary for the posterior distributions, as the prior volume occupied by high likelihood regions is usually very small and therefore these regions have relatively small probability mass. For profile likelihoods, however, getting as close as possible to the true global maximum is crucial and therefore one should set tol to a relatively smaller value. In Ref. [17] it was found that $n_{\mathrm{live}} = 20,000$ and $\mathrm{tol} = 1 \times 10^{-4}$ produce a sufficiently accurate exploration of the profile likelihood in toy models that reproduce the most important features of the CMSSM parameter space.

In principle, the profile likelihood does not depend on the choice of priors. However, in order to explore the parameter space using any Monte Carlo technique, a set of priors needs to be defined. Different choices of priors will generally lead to different regions of the parameter space to be explored in greater or lesser detail, according to their posterior density. As a consequence, the resulting profile likelihoods might be slightly different, purely on numerical grounds. We can obtain more robust profile likelihoods by simply merging samples obtained from scans with different choices of Bayesian priors. This does not come at a greater computational cost, given that a responsible Bayesian analysis would estimate sensitivity to the choice of prior as well. The results of such a scan are shown in Fig. 3, which was obtained by tuning MULTINEST with the above configuration, appropriate for an accurate profile likelihood exploration, and by merging the posterior samples from two different choices of priors (see [17] for details). This high-resolution profile likelihood scan using MULTINEST compares favourably with the results obtained by adopting a dedicated Genetic Algorithm technique [11], although at a slightly higher computational cost (a factor of ~ 4). In general, an accurate profile likelihood evaluation was about an order of magnitude more computationally expensive than mapping out the Bayesian posterior.

4 Conclusions

As the LHC impinges on the most anticipated regions of SUSY parameter space, the need for statistical techniques that will be able to cope with the complexity of SUSY phenomenology is greater than ever. An intense effort is underway to test the accuracy of parameter inference methods, both in the Frequentist and the Bayesian framework. Coverage studies such as the one presented here require highly-accelerated inference techniques, and neural networks have been demonstrated to provide a speed-up factor of up to $30,000$ with respect to conventional methods. A crucial improvement required for future coverage investigations is the ability to generate pseudo-experiments from an accurate description of the likelihood. Both the representation of the likelihood function and the ability to generate pseudo-experiments are now possible with the workspace technology in RooFit/RooStats [18]. We encourage future experiments to publish their likelihoods using this technology. Finally, an accurate evaluation of the profile likelihood remains a numerically challenging task, much more so than the mapping out of the Bayesian posterior. Particular care needs to be taken in tuning algorithms appropriately, depending on whether one is interested in posterior mass or likelihood maximization. We have demonstrated that the MULTINEST algorithm can be successfully employed for approximating the profile likelihood functions, even though it was primarily designed for Bayesian analyses. In particular, it is important to use a termination criterion that allows MULTINEST to explore high-likelihood regions to sufficient resolution.

Acknowledgements: We would like to thank the organizers of PHYSTAT11 for a very interesting workshop. We are grateful to Yashar Akrami, Jan Conrad, Joakim Edsjö, Louis Lyons and Pat Scott for many useful discussions.

References

[1] E. A. Baltz and P. Gondolo, *JHEP* **10** (2004) 052.

[2] B. C. Allanach and C. G. Lester, *Phys. Rev.* **D73** (2006) 015013.

[3] R. Ruiz de Austri, R. Trotta, and L. Roszkowski, *JHEP* **05** (2006) 002.

[4] B. C. Allanach, K. Cranmer, C. G. Lester, and A. M. Weber, *JHEP* **0708** (2007) 023.

[5] O. Buchmueller, R. Cavanaugh, A. De Roeck, J. R. Ellis, H. Flacher, S. Heinemeyer, G. Isidori, K. A. Olive *et al.*, Eur. Phys. J. **C64**, 391-415 (2009).

[6] L. Roszkowski, R. Ruiz de Austri, and R. Trotta, *JHEP* **07** (2007) 075.

[7] J. Skilling, *Nested Sampling*, in *American Institute of Physics Conference Series* (R. Fischer, R. Preuss, and U. V. Toussaint, eds.), pp. 395–405, (2004).

[8] F. Feroz, M. P. Hobson, and M. Bridges, *Mon. Not. R. Astron. Soc.* **398** (2009) 1601–1614.

[9] R. Lafaye, T. Plehn, M. Rauch, and D. Zerwas, *European Physical Journal C* **54** (2008) 617–644.

[10] R. Trotta, F. Feroz, M. Hobson, L. Roszkowski, and R. Ruiz de Austri, *JHEP* **12** (2008) 24.

[11] Y. Akrami, P. Scott, J. Edsjo, J. Conrad, and L. Bergstrom, *JHEP* **04** (2010) 057.

[12] M. Bridges, K. Cranmer, F. Feroz, M. Hobson, R. R. de Austri, R. Trotta, JHEP **1103**, 012 (2011).

[13] The ATLAS Collaboration: G. Aad, E. Abat, B. Abbott, J. Abdallah, A. A. Abdelalim, A. Abdesselam, O. Abdinov, B. Abi, M. Abolins, H. Abramowicz, and et al., *ArXiv e-prints* (Dec., 2009) [http://xxx.lanl.gov/abs/0901.0512].

[14] Y. Akrami, C. Savage, P. Scott, J. Conrad, and J. Edsjö, (2010), pre-print: [http://xxx.lanl.gov/abs/1011.4297].

[15] P. Scott, J. Conrad, J. Edsjö, L. Bergström, C. Farnier, and Y. Akrami, JCAP **1001**, 031 (2010).

[16] S. Wilks, *Ann. Math. Statist.* **9** (1938) 60–2.

[17] F. Feroz, K. Cranmer, M. Hobson, R. Ruiz de Austri, R. Trotta, (2011), pre-print: [http://xxx.lanl.gov/abs/1101.3296].

[18] L. Moneta, K. Belasco, K. Cranmer, A. Lazzaro, D. Piparo, *et al.*, *The RooStats Project*, *Proceed-*

ings of Science (2010) Proceedings of the 13th International Workshop on Advanced Computing and Analysis Techniques in Physics Research, India, [http://xxx.lanl.gov/abs/arXiv:1009.1003].

p-values for Model Evaluation

Frederik Beaujean[1*], *Allen Caldwell*[1], *D. Kollár*[2], *K. Kröninger*[3]
[1]Max-Planck-Institut für Physik, München, Germany
[2]CERN, Geneva, Switzerland
[3]II Physikalisches Institut, Universität Göttingen, Germany
* Corresponding author

Abstract
A quantitative procedure to decide whether a model provides a good description of data is often based on a specific test statistic and a p-value summarizing both the data and the statistic's sampling distribution. We provide a Bayesian motivation for using p-values in the goodness-of-fit problem with no explicit alternative models considered. Some typical pitfalls encountered with common statistics are reviewed for Poisson and Gaussian uncertainties. Finally, we present a new test statistic for ordered Gaussian data, the *runs* statistic.

1 Introduction

Progress in science is the result of an interplay between model building and the testing of models with experimental data. In this paper, we discuss model evaluation and focus primarily on situations where a statement is desired on the validity of a model without explicit reference to other models. We introduce different *discrepancy variables* [1] (an extension of classical test statistics to allow possible dependence on unknown (nuisance) parameters) for this purpose and define p-values based on these. p-values have been discussed extensively in the literature [2, 3], in particular also at previous PHYSTAT conferences [4, 5].

Following a Bayesian motivation for p-values in Section 2, we introduce an example fit problem in Section 3. Next, we explore some common pitfalls in p-value calculations with Gaussian uncertainties in Section 4 and study the usefulness of p-values despite approximations for the Poisson case in Section 5. Finally, we present a new discrepancy variable based on runs for ordered Gaussian data in Section 6.

In general, any discrepancy variable which can be calculated for the observations can be used to define a p-value. We use $R(\vec{x}|\vec{\theta}, M)$ and $R(\vec{D}|\vec{\theta}, M)$ to denote discrepancy variables evaluated with a possible set of observations $\vec{x}$ for given model M and parameter values $\vec{\theta}$, and for the observed data, $\vec{x} = \vec{D}$, respectively. To simplify the notation, we will occasionally drop the arguments on R and use R^D to denote the value of the discrepancy variable found from the data set at hand. R can be interpreted as a random variable, whereas R^D has a fixed value.

Assuming that smaller values of R imply better agreement between the data and model predictions, the definition of p (for continuous frequency distribution of R) is written as:

$$p = \int_{R>R^D} P(R|\vec{\theta}, M)\mathrm{d}R \ . \tag{1}$$

The basic fact used in interpreting p-values is the following: under M with the value of $\vec{\theta}$ fixed before data is analyzed, p is a random variable with uniform distribution on $[0, 1]$; i.e. $p \sim U[0, 1]$.

However, in most practical examples the value of $\vec{\theta}$ is not fixed a-priori. The choice of parameter values from fitting the data set, $\vec{\theta}_{fit}$, affects the distribution $P(R|\vec{\theta}, M)$ typically in an unknown way[1]. Hence, a p-value based on the distribution $P(R|\vec{\theta} = \vec{\theta}_{fit}, M)$ assuming fixed $\vec{\theta}$ is in general not $U[0, 1]$. This introduces confusion in interpreting p, as Berger put it: "being $U[0, 1]$ defines a proper p-value, allowing for its common interpretation across problems. Statistical measures that lack a common interpretation across problems are simply not very useful" [2].

[1]The one notable exception, χ^2, is discussed in Sec. 4

2 Bayesian motivation for p-values

When using a p-value to claim discovery of, say, a new particle at a high energy physics experiment, it is indispensable to take systematic effects of the detector into account. However, the correct distribution of the data fluctuations, including systematic effects, is often not known and best guesses are used. These guesses introduce a degree of subjectivity that affect p, no matter if the distribution of R is estimated from simulating data sets or approximated using simple closed-form expressions as in Sections 4, 5. This inherent vagueness should always be remembered when interpreting p-values.

We contend that the (frequentist) use of p-values for evaluation of models is essentially Bayesian in character. Assume that the p-value probability density for a good model, M_0, is uniform, $P(p|M_0) = 1$, and for poor models, $M_i \;\; (i = 1, k)$, can be represented by

$$P(p|M_i) \approx \lambda_i e^{-\lambda_i p} \tag{2}$$

where $\lambda_i \gg 1$ so that the distribution is strongly peaked at 0 and approximately normalized to 1. Using Bayes' theorem, we update the prior *degree-of-belief* (DoB) in model M_0, $P_0(M_0)$, to the posterior DoB, $P(M_0|p)$, after finding a particular p-value

$$P(M_0|p) = \frac{P(p|M_0)P_0(M_0)}{P(p|M_0)P_0(M_0) + \sum_{i=1}^{k} P(p|M_i)P_0(M_i)} \;\; . \tag{3}$$

If we take all models to have similar prior DoBs, $P_0(M_0) \approx P_0(M_i)$, then

$$P(M_0|p) \approx \frac{P(p|M_0)}{P(p|M_0) + \sum_{i=1}^{k} P(p|M_i)} \;\; . \tag{4}$$

In the limit $p \to 0$, we have

$$P(M_0|p) \approx \frac{1}{1 + \sum_{i=1}^{k} \lambda_i} \ll 1 \;\; , \tag{5}$$

while for $\lambda_i p \gg 1 \;\; \forall i$ we have $P(M_0|p) \approx 1$, ruling out any alternative to M_0.

Although this formulation in principle allows for a ranking of models, the vague nature of this procedure indicates that any model which can be constructed to yield a reasonable p-value should be retained. Effectively, the posterior $P(M_0|p)$ depends on the data only indirectly through $p = p(D)$. Clearly, if p is not a sufficient statistic, valuable information is not used.

3 Example fit problem

In the following, we test the usefulness of different discrepancy variables R by looking at the respective p-value distributions for an example typical of high energy physics. We first consider a data set which consists of a background known to be smoothly rising and, in addition to the background, a possible signal. This could correspond for example to an enhancement in a mass spectrum from the presence of a new resonance. The width of the resonance is not known, so that a wide range of widths must be allowed for. Also, the shape of the background is not well known. We do not have an exhaustive set of models to compare and want to look at GoF's for models individually to make decisions; direct model comparison is outside the scope of this paper. In Sections 4 and 6 we model fluctuations of the data relative to expectations with Gaussian distributions. We also consider the same problem in Section 5 with small event numbers, so that Poisson statistics are appropriate. These examples are discussed in more detail in [6, 7]. Typical data sets are shown in Fig. 1 for $N = 25$ data points (Poisson: bin contents), generated from the function

$$f(x_i) = A + B\,x_i + C\,x_i^2 + \frac{D}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \;\; , \tag{6}$$

with parameter values ($A = 0$, $B = 0.5$, $C = 0.02$, $D = 15$, $\sigma = 0.5$, $\mu = 5.0$). The y_i are generated from $f(x_i)$ as $y_i = f(x_i) + z_i$ where z_i is sampled according to $\mathcal{N}(0, 4)$. We fit the following four models to the data:

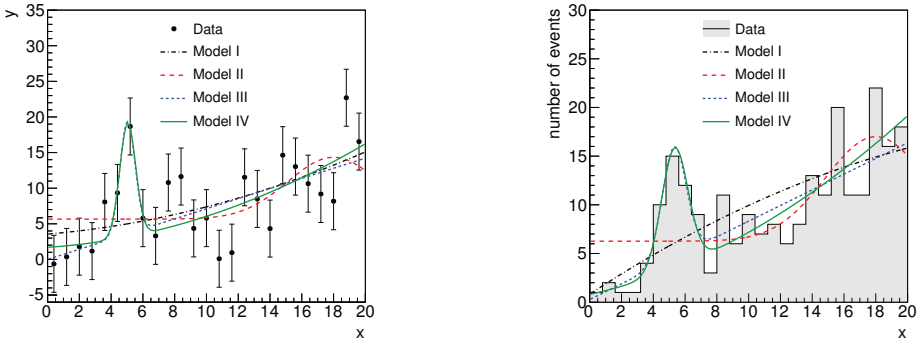


Fig. 1: Example data set for the case $N = 25$ with Gaussian (left) and Poissonian (right) fluctuations. The fits of the four models are superimposed on the data.

I. quadratic: $\vec{\theta} = (A, B, C)$ corresponding to the "standard model";

II. constant + Gaussian: $\vec{\theta} = (A, D, \mu, \sigma)$;

III. linear + Gaussian: $\vec{\theta} = (A, B, D, \mu, \sigma)$;

IV. quadratic + Gaussian: $\vec{\theta} = (A, B, C, D, \mu, \sigma)$ corresponding to the true function (6).

4 Revisiting the Gaussian case

For uncorrelated data assumed to follow Gaussian probability distributions relative to the model predictions, the discrepancy variable considered most often in high energy physics is the classic χ^2

$$R_G = \chi^2 = \sum_{i=1}^{N} \frac{\Big(y_i - f(x_i|\vec{\theta}, M)\Big)^2}{\sigma_i^2} \tag{7}$$

R_G is both fast to evaluate and, at first sight, easy to turn into a p-value (using ROOT's `TMath::Prob(...)`). However, in practical examples the conditions to do so are usually not satisfied. The frequency distribution of R_G is the celebrated χ^2-distribution with $(N - \dim\vec{\theta})$ degrees-of-freedom (DoF) if [8]

- the data fluctuations are Gaussian and the σ_i's are independent of the parameters,
- the function to be compared to the data depends linearly on the parameters, and
- the parameters are chosen such that R_G is at its *global* minimum.

In our example, the above conditions may be violated in two ways:

1. by construction: the predictions $f(x_i|\vec{\theta}, M)$ from (6) are non-linear in $\vec{\theta}$, or
2. for numerical reasons: the likelihood $P(\vec{x}|\vec{\theta}, M) \propto \exp(-R_G/2)$ has several modes.

Multimodality gives rise to technical issues: when using a gradient-based optimization algorithm like MIGRAD from the MINUIT package [9], it is critical to choose a good starting point in parameter space. If best-fit parameter values $\vec{\theta}_{loc}$ are chosen at a local minimum rather than at the global minimum $\vec{\theta}_{glob}$ such that $R_G(\vec{D}|\vec{\theta}_{loc}, M) > R_G(\vec{D}|\vec{\theta}_{glob}, M)$, then using the χ^2-distribution to turn $R_G(\vec{D}|\vec{\theta}_{loc}, M)$ into a p-value yields a p-value distribution that peaks at $p = 0$, significantly deviating from $p \sim U[0, 1]$. The physicist performing a fit often believes to have a good idea "where the best-fit parameters ought to be"

and takes that as a starting point. However, we have seen in our fits that even when we know the true value of $\vec{\theta}$, $\vec{\theta}_{true}$, starting MIGRAD there for some data sets doesn't lead to the global maximum. We thus recommend a different procedure: if there is any concern that several modes may exist, one should use a Monte Carlo sampling method (we used the implementation of the Metropolis-Hastings algorithm in BAT [7]) to explore the parameter space, and take the best-fit parameters encountered in the sampling to seed MIGRAD.

A further complication may arise when a Bayesian fit with non-uniform priors on $\vec{\theta}$ is performed; the maximum of the posterior doesn't coincide with the minimum of R_G. Choosing parameter ranges in a maximum-likelihood fit is (at least at the numerical level) equivalent to performing a Bayesian fit with uniform priors with compact support. Obviously different priors can lead to a different resulting p-value distribution. In our example we have used hypercubes in parameter space of a different size. Using the smaller volume, which contains $\vec{\theta}_{true}$, the distribution of p is biased towards $p = 0$ with a maximum deviation from uniformity of about 20%. On the other hand, with a much larger volume the distribution is now biased towards $p = 1$, again with a maximum deviation from uniformity of about 20%. The discrepancy between the two stems from the fact that the global optimum is in some cases outside of the smaller volume. For the larger volume, $p \nsim U[0, 1]$ is expected, since the fit function is non-linear in $\vec{\theta}$. For plots and further details see Ref. [6], Chapter 4.

5 Revisiting the Poisson case

Similar to the previous section, we now fit the models I-IV to a histogram . We proceed in analogy to Baker&Cousins [10] and limit the discussion to three common discrepancy variables to judge the GoF. Suppose N is the number of bins, $\nu_i = \nu_i(\vec{\theta}, M)$ is the expected number of events in bin i, and n_i is the observed number of events. Then we define

$$R_P = \text{Pearson's } \chi^2 = \sum_{i=1}^{N} \frac{(n_i - \nu_i)^2}{\nu_i}, \quad R_N = \text{Neyman's } \chi^2 = \sum_{i=1}^{N} \frac{(n_i - \nu_i)^2}{n_i} \quad . \tag{8}$$

In cases where $n_i = 0$, practitioners of this approach set $n_i = 1$ in R_N's denominator to avoid divergence. Sometimes bins with $n_i = 0$ are ignored, which can lead to very misleading results since finding $n_i = 0$ is valuable information. Finally, we have the log likelihood ratio (sometimes called Cash statistic [11])

$$R_C = 2 \log \frac{P(\vec{x}|\nu_i = n_i)}{P(\vec{x}|\nu_i = \nu_i(\vec{\theta}))} = 2 \sum_{i=1}^{N_b} \left[\nu_i - n_i + n_i \log \frac{n_i}{\nu_i} \right] \quad , \tag{9}$$

where $P(\vec{x}|\nu_i)$ is the product of Poisson probabilities for each bin. Asymptotically, i.e. for $n_i \gg 1 \; \forall i$, R_P, R_N and R_C are χ^2-distributed with $(N - \dim \vec{\theta})$ DoF. But for "finite sample size ... general results are lacking" [10], and that is precisely the case of interest in physics. The situation is aggravated in our example, as some bins typically have few or even no events, see Figure 1. We have generated 10000 data sets with Poissonian fluctuations from (6) to estimate the p-value frequency distributions across 20 bins in Figure 2. For each data set and discrepancy variable R, we calculated a p-value using the χ^2-distribution with the value of $\vec{\theta}$ chosen to minimize the respective R.

Models I and II are ruled out by each R. By construction, models III and IV are very similar. R_P doesn't distinguish well between the two, while R_N's and R_C's distributions peak for III, but look more uniform for IV. If one is interested in setting a frequentist limit at the 95% confidence level, then the first bin of each distribution in Figure 2 is the relevant one. For model IV, the densities (R_P: 0.58, R_N: 1.85, R_C: 1.35) differ significantly from the desired value of 1, given a statistical uncertainty of $\mathcal{O}(5\%)$ obtained from a binomial model with uniform prior on the chance of ending up in this first bin. We also display model IV (true model) with the true parameter values to get a feeling for the quality of the approximation in using the asymptotic χ^2 distribution; here, only N DoF are used. R_N has a worrisome peak at $p = 0$, while R_P and R_C are fairly uniform. Based on this numerical study, we discourage the use

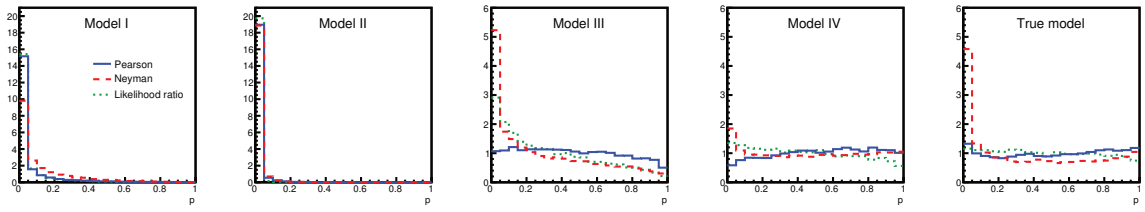


Fig. 2: p-value distributions based on R_P, R_N and R_C using the χ^2-distribution with $N - n$ degrees of freedom, where n is the number of fitted parameters.

of Neyman's χ^2 and recommend Pearson's χ^2 and the likelihood ratio, bearing in mind the inaccuracy due to finite sample size.

6 Runs statistic

When defining the p-value based on a statistic T, the usually multi-dimensional data is compressed into a single number T. Often, T is, by construction, insensitive to certain features of the data. While this is beneficial in some cases, it often represents a short coming. Returning to the Gaussian example of Section 4, R_G is merely a measure of the average distance of a single observation to its predicted value. But what if one is interested in checking that the *sequence* of data points agrees with model predictions? In high energy physics, data is often available in a 1-D ordering, e.g., the cross section y for a set of energies $x_i, i = 1 \ldots N$. Suppose there is a peak in the distribution $y(x)$ that is not predicted by the standard model. If the peak is localized in the sense that only a few y_i exceed the standard model predictions, then even for moderately large N, R_G will not detect a mismatch as the average deviation to the standard model is typically within accepted levels.

Recently, we have proposed the runs statistic [6,11] as a companion to R_G in order to gain sensitivity to local clustering of observations in the case of independent, Gaussian distributed samples. Assume the ordered set of N observations $\{(x_i, y_i)\}$ is partitioned into subsets containing the success and failure runs (defined as sequences of consecutive y_i above or below the expectation from the model, $f(x_i|\vec{\theta}, M)$, respectively).

Let A_j denote the subset of the observations of the j^{th} success run. The weight of the j^{th} success run is then taken to be

$$\chi^2_{\mathrm{run},j} = \sum_{i=j_1}^{j_1+N_j-1} \frac{\Big(y_i - f(x_i|\vec{\theta}, M)\Big)^2}{\sigma_i^2} \tag{10}$$

where the sum over i covers the $(x_i, y_i) \in A_j$ and N_j is the length of the run. The discrepancy variable is then the largest weight of any success run: $R_{sr} \equiv \max_j \chi^2_{\mathrm{run},j}$.

The exact frequency distribution of R_{sr}, used to define the p-value, $p = P(R_{sr} > R^D_{sr}|N)$, is given in [11] for the case when $(\vec{\theta}, M)$ are fully specified (no fitting). A similar discrepancy variable can be defined for failure measurements, R_{fr}.

To illustrate the definition we present a simple example. Suppose $N = 5$ observations at x positions $(1, 2, 3, 4, 5)$ with standardized residuals $(y_i - f(x_i|\vec{\theta}, M))/\sigma_i$ given by $(0.3, -0.1, -0.8, 0.4, 0.2)$. Then there are two success runs $A_1 = \{(1, 0.3)\}$, $A_2 = \{(4, 0.4), (5, 0.2)\}$ and we find $R_{sr} = 0.16 + 0.04 = 0.2$ due to the second run. Similarly, for the single failure run, $R_{fr} = 0.65$.

For the example (6) from Section 3, the joint distribution of p-values for success and failure runs based on $P(R_{sr} > R^D_{sr}|N)$ for models I and IV is shown in 3. A cut in two dimensions allows for a clean separation, while from the marginal 1-D distributions the different models are much harder to separate.

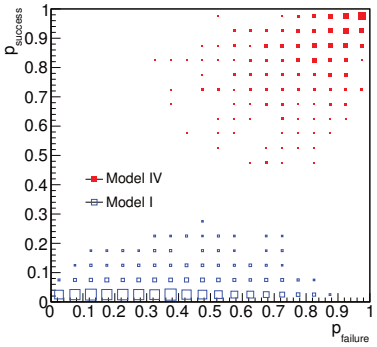


Fig. 3: Joint distribution of the p-values for success and failure runs. Bins with probability less than $3.5 \cdot 10^{-3}$ have been excluded from the plot for the purpose of clarity.

7 Discussion

In the examples which we studied it has become apparent that it is difficult to construct a p-value which is $U[0,1]$ when parameters are fitted. On the one hand, this is due to approximations of a discrepancy variable's frequency distribution. On the other hand, the numerical fitting procedure may have an impact if it finds only a local optimum. In the discussion at PHYSTAT2011, Kyle Cranmer stressed that he would prefer that a quantity defined as in (1) with non-uniform distribution should not be called p-value at all to avoid confusion in its interpretation. However, in our opinion $p \nsim U[0,1]$ is tolerable, as p-values should not be used in a simple accept/reject fashion, but merely as guidance as to whether a better model has to be constructed to explain the data. After all, the p-values displayed in Figure 2 serve that purpose: a physicist starting with model II only would be well advised to look further, and hopefully arrive at model III or IV.

References

[1] A. Gelman, X. L Meng, and H. Stern. Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 6:733–759, 1996.

[2] M. J. Bayarri and James O. Berger. P values for composite null models. *Journal of the American Statistical Association*, 95(452):1127–1142, 2000.

[3] Mark J. Schervish. P values: What they are and what they are not. *The American Statistician*, 50(3):203–206, 1996.

[4] Luc Demortier. P values and nuisance parameters. In *PHYSTAT-LHC Workshop on Statistical Issues for LHC Physics" (2007)*, pages 23–33, 2007.

[5] I. V. Narsky. Goodness of fit. *PHYSTAT2003*, pages 70–74, 2003.

[6] F. Beaujean, A. Caldwell, D. Kollár, and K. Kröninger. p-values for model evaluation. *Physical Review D*, 83(1):012004, 2011.

[7] A. Caldwell, D. Kollár, and K. Kröninger. BAT- the Bayesian Analysis Toolkit. *Computer Physics Communications*, 180(11):2197–2209, 2009.

[8] W. T. Eadie, D. Drijard, F. E. James, M. Roos, and B. Sadoulet. *Statistical Methods in Experimental Physics*. North-Holland, Amsterdam, 1971.

[9] F. James. MINUIT - a system for function minimization and analysis of the parameter errors and correlations. *Computer Physics Communications*, 10:343–367, 1975.

[10] Steve Baker and Robert D. Cousins. Clarification of the use of χ^2 and likelihood functions in fits to histograms. *Nuclear Instruments and Methods in Physics Research*, 221(2):437–442, 1984.

[11] Frederik Beaujean and Allen Caldwell. A test statistic for weighted runs. *arxiv:1005.3233*, 2010.

Estimating the "look elsewhere effect" when searching for a signal

Ofer Vitells
Weizmann Institute of Science, Rehovot 76100, Israel

Abstract
The "look elsewhere effect" refers to a common situation where one searches for a signal in some space of parameters - for example, a resonance search with unknown mass, or a search for astrophysical point sources with unknown location in the sky. Since Wilks' theorem does not apply in such cases, one usually has to resort to computationally expansive Monte-Carlo simulations in order to correctly estimate the significance of a given observation. Recent results from the theory of random fields provide powerful tools which may be used to alleviate this difficulty, in a wide range of applications. We review those results and discuss their implementation in problems of practical interest.

1 Introduction

Experiments that aim at discovering new physical phenomena often involve a search for a signal over some space of continuous parameters. One such example is the search for the Higgs boson at particle colliders, where one searches for a peak within some range of an invariant mass distribution. Another example is the search for astrophysical neutrino sources that can be located at any direction in the sky. To assess the significance of a local deviation from the background hypothesis in terms of a p-value, one needs to take into account the probability of such a deviation to occur anywhere within the search range. This is the so called "look elsewhere effect". Estimation of the p-value could be performed by repeated Monte Carlo simulations of the experiment's outcome under the background-only hypothesis, but this approach could be highly time consuming since for each of those simulations the entire search procedure needs to be applied to the data, and to establish a discovery claim at the 5σ level (p-value=2.87×10^{-7}) the simulation needs to be repeated at least $\mathcal{O}(10^7)$ times. Fortunately, recent advances in the theory of random fields provide analytical tools that can be used to address exactly such problems, in a wide range of experimental settings. Such methods could be highly valuable for experiments searching for signals over large parameter spaces, as the reduction in necessary computation time can be dramatic. Random field theoretic methods were first applied to the statistical hypothesis testing problem in [1], for some special case of a one dimensional problem. A practical implementation of this result, aimed at the high-energy physics community, was made in [2]. Similar results for some cases of multi-dimensional problems [3] [4] were applied to statistical tests in the context of brain imaging [5]. More recently, a generalized result dealing with random fields over arbitrary Riemannian manifolds was obtained [6], openning the door for a plethora of new possible applications. Here we discuss the implementation of these results in the context of physics experiments, taking two representative cases as specific examples. In section 2 the general framework of an hypothesis test is briefly presented with connection to random fields. In section 3 the main theoretical result is presented, and two examples are treated in detail in sections 4 and 5.

2 Formalism of a search as a statistical test

Consider the problem of an hypothesis testing, where one tests the background (null) hypothesis $H_0 : \mu = 0$, against a signal hypothesis $H_1 : \mu > 0$, where μ represents the signal strength. Suppose that θ are some nuisance parameters describing other properties of the signal (such as location), which are therefore not present under the null. Additional nuisance parameters, denoted by θ', may be present under both hypotheses. Denote by $\mathscr{L}(\mu, \theta, \theta')$ the likelihood function. One may then construct the profile likelihood ratio test statistic [7]

$$q = -2\log \frac{\max\limits_{\theta'} \mathscr{L}(\mu = 0, \theta')}{\max\limits_{\mu,\theta,\theta'} \mathscr{L}(\mu, \theta, \theta')} \tag{1}$$

and reject the null hypothesis if the test statistic is larger then some critical value. Note that when the signal strength is set to zero the likelihood by definition does not depend on θ, and the test statistic (1) can therefore be written as

$$q = \max_{\theta \in \mathscr{M}} q(\theta) \tag{2}$$

where $q(\theta)$ is the profile likelihood ratio with the signal nuisance parameters fixed to the point θ, and we have denoted by $\mathscr{M}$ the D-dimensional manifold to which the parameters θ belong. Under the conditions of Wilks' theorem [8], $q(\theta)$ follows a χ^2 distribution with one degree of freedom when the null hypothesis is true. When viewed as a function over the manifold $\mathscr{M}$, $q(\theta)$ is therefore a χ^2 *random field*, namely a set of random variables that are continuously mapped to the manifold $\mathscr{M}$. To quantify the significance of a given observation in terms of a p-value, one is required to calculate the probability of the maximum of the field to be above some level, that is, the excursion probability of the field:

$$p\text{-value} = \mathbb{P}[\max_{\theta \in \mathscr{M}} q(\theta) > u]. \tag{3}$$

In most cases, direct calculation of the above quantity will probably be too difficult to be of any practical use. However, other closely related quantities exist for which surprisingly simple closed-form expressions have been derived under general conditions. Those will allow to estimate the excursion probability (3) when the level u is large, which is the main region of interest.

3 The excursion sets of random fields

The excursion set of a field above a level u, denoted by A_u, is defined as the set of points θ for which the value of the field $q(\theta)$ is larger than u,

$$A_u = \{\theta \in \mathscr{M} : q(\theta) > u\} \tag{4}$$

and we will denote by $\phi(A_u)$ the Euler characteristic of the excursion set A_u. A fundamental result of [6] states that the expectation of the Euler characteristic $\phi(A_u)$ is given by the following expression:

$$\mathbb{E}[\phi(A_u)] = \sum_{d=0}^{D} \mathscr{N}_d \rho_d(u). \tag{5}$$

The coefficients $\mathscr{N}_d$ are related to some geometrical properties of the manifold and the covariance structure of the field, for our purposes however they will be just a set of unknown constants. The functions $\rho_d(u)$ are 'universal' in the sense that they are determined only by the distribution type of the field $q(\theta)$, and their analytic expressions are known for a large class of 'Gaussian related' fields, such as χ^2 with arbitrary degrees of freedom. The zeroth order term of eq. (5) is a special case for which $\mathscr{N}_0$ and $\rho_0(u)$ are generally given by

$$\mathscr{N}_0 = \phi(\mathscr{M}), \qquad \rho_0(u) = \mathbb{P}[q(\theta) > u] \tag{6}$$

Namely, $\mathscr{N}_0$ is the Euler characteristic of the entire manifold and $\rho_0(u)$ is the tail probability of the distribution of the field. (Note that when the manifold is reduced to a point, this result becomes trivial).

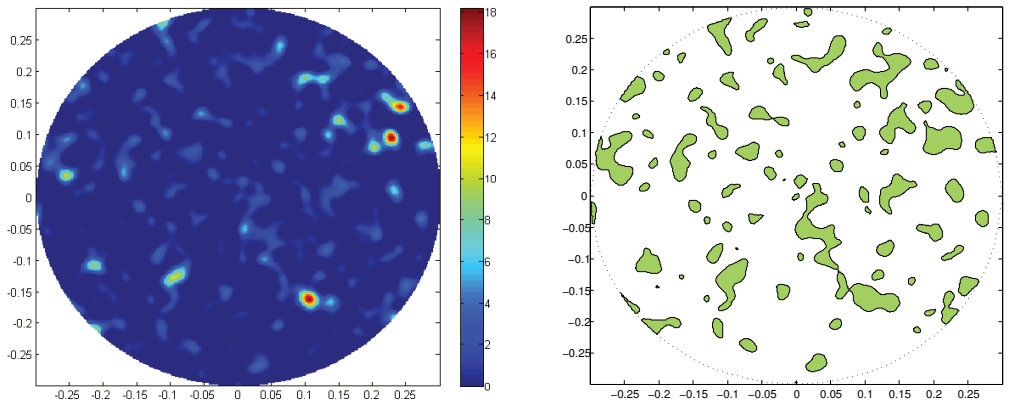
When the level u is high enough, excursions above u become rare and the excursion set becomes a few disconnected hyper-ellipses. In that case the Euler characteristic $\phi(A_u)$ simply counts the number of disconnected components that make up A_u. For even higher levels this number is mostly zero and rarely one, and Its expectation therefore converges asymptotically to the excursion probability. We can thus use it as an approximation to the excursion probability for large enough u [9]

$$\mathbb{E}[\phi(A_u)] \approx \mathbb{P}[\max_{\theta \in \mathscr{M}} q(\theta) > u]. \tag{7}$$

The practical importance of Eq. (5) now becomes clear, as it allows to estimate the excursion probabilities above high levels. Furthermore, the problem is reduced to finding the constants $\mathscr{N}_d, d > 0$. Since Eq. (5) holds for any level u, this could be achieved simply by calculating the average of $\phi(A_u)$ at some low levels, which can be done using a small set of Monte Carlo simulations. We shall now turn to a few examples where this procedure is demonstrated.

4 Example 1: detecting neutrino sources

The IceCube experiment [10] is a neutrino telescope located at the south pole and aimed at detecting astrophysical neutrino sources. The detector measures the energy and angular direction of incoming neutrinos, trying to distinguish an astrophysical point-like signal from a large background of atmospheric neutrinos spread across the sky. The nuisance parameters over which the search is performed are therefore the angular coordinates (θ, φ)[1]. We follow [11] for the definitions of the signal and background distributions and calculate a profile likelihood ratio as described in the previous section. Figure 1 shows a "significance map" of the sky, namely the values of the test statistic $q(\theta, \varphi)$ as well as the corresponding excursion set above $q = 1$. To reduce computation time we restrict here the search space to the portion of the sky at declination angle 27° below the zenith, however all the features of a full sky search are maintained. Note that the most significance point has a value of the test statistic above 16, which would correspond to a significance exceeding 4σ if this point would have been analyzed alone, that is without the "look elsewhere" effect.



(a) (b)

Fig. 1: (a) A significance map showing the test statistic $q(\theta, \varphi)$ for a background simulation (b) The corresponding excursion set above $q = 1$.

[1]The signal model may include additional parameters such as spectral index and time, which we do not consider here for simplicity.

For a χ^2 random field with one degree of freedom and for two search dimensions, Eq. (5) reads [6]

$$\mathbb{E}[\phi(A_u)] = \mathbb{P}[\chi^2 > u] + e^{-u/2}(\mathcal{N}_1 + \sqrt{u}\mathcal{N}_2). \tag{8}$$

To estimate the the coefficients $\mathcal{N}_1, \mathcal{N}_2$ we use a set of 20 background simulations, and calculate the Euler characteristic of the excursion set corresponding to the levels $u = 0, 1$. This gives the estimates $\mathbb{E}[\phi(A_0)] = 33.5 \pm 2$ and $\mathbb{E}[\phi(A_1)] = 94.6 \pm 1.3$. By solving for the unknown coefficients we obtain $\mathcal{N}_1 = 33 \pm 2$ and $\mathcal{N}_2 = 123 \pm 3$. The prediction of Eq. (8) is then compared against a set of approx. 200,000 background simulations, where for each one the maximum of $q(\theta, \varphi)$ is found by scanning the entire range. The results are shown in Figure 2. As expected, the approximation becomes better as the p-value becomes smaller. The agreement between Eq. (8) and the observed p-value is maintained up to the smallest p-value that the available statistics allows us to estimate.

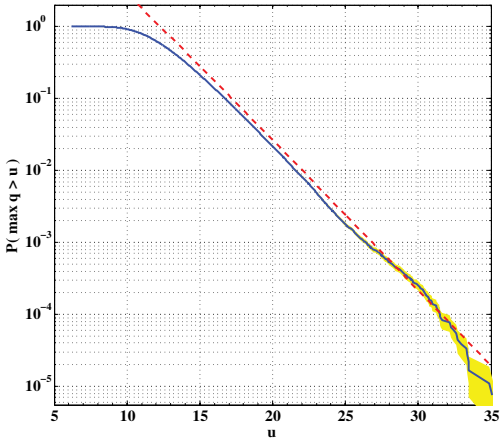


Fig. 2: The prediction of Eq. (8) (dashed red) against the observed p-value (solid blue) from a set of 200,000 background simulations. The yellow band represents the statistical uncertainty due to the available number of background simulations.

4.1 Slicing the parameter space

A useful property of Eq. (5) that can be illustrated by this example, is the ability to consider only a small 'slice' of the parameter space from which the expected Euler characteristic (and hence p-value) of the entire space can be estimated, if a symmetry is present in the problem. This can be done using the 'inclusion-exclusion' property of the Euler characteristic:

$$\phi(A \cup B) = \phi(A) + \phi(B) - \phi(A \cap B). \tag{9}$$

Since the neutrino background distribution is assumed to be uniform in azimuthal angle (φ), we can divide the sky to N identical slices of azimuthal angle, as illustrated in Figure 3. Applying (9) to this case, the expected Euler characteristic is given by

$$\mathbb{E}[\phi(A_u)] = N \times (\mathbb{E}[\phi(slice)] - \mathbb{E}[\phi(edge)]) + \mathbb{E}[\phi(0)] \tag{10}$$

where an 'edge' is the line common to two adjacent slices, and $\phi(0)$ is the Euler characteristic of the point at the origin (see Figure 3).

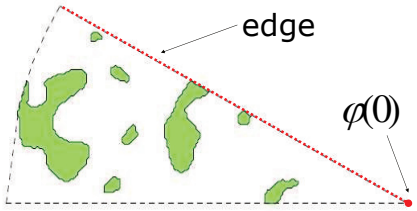


Fig. 3: Illustration of the excursion set in a slice of a sky, showing also an edge (dashed red) and the origin as defined in Eq. (10).

We can now apply Eq. (5) to both $\phi(slice)$ and $\phi(edge)$ and estimate the corresponding coefficients as was done before, using only simulations of a single slice of the sky. Following this procedure we obtain for this example with $N = 18$ slices from 40 background simulations, $\mathscr{N}_1^{slice} = 6 \pm 0.5, \mathscr{N}_2^{slice} = 6.7 \pm 0.8$ and $\mathscr{N}_1^{edge} = 4.4 \pm 0.2$. Using (10) this leads to the full sky coefficients $\mathscr{N}_1 = 28 \pm 9$ and $\mathscr{N}_2 = 120 \pm 14$, a result which is consistent with the full sky simulation procedure.

5 Example 2: A 'mass bump' with unknown width

As a second example we consider the common problem in high energy physics of detecting a 'mass bump' on top of a continuous background, and we assume that the width of the bump is also a-priori unknown (within some range). We assume an exponential background distribution and a gaussian signal distribution, with the signal location and width being the free nuisance parameters. The search space is therefore two dimensional in this problem as well. Figure 4 shows an example histogram of background events with a signal best fit. Figure 5 shows the values of the test statistic q as a function of the mass and the width, and the corresponding excursion set above $q = 1$.

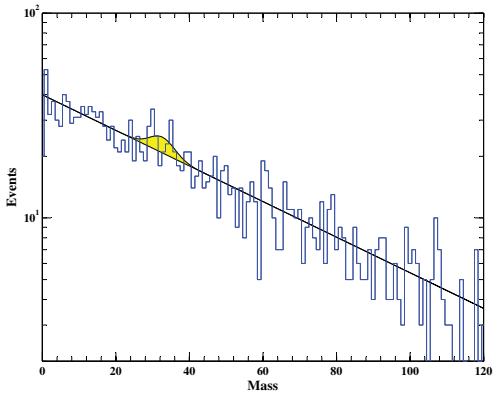
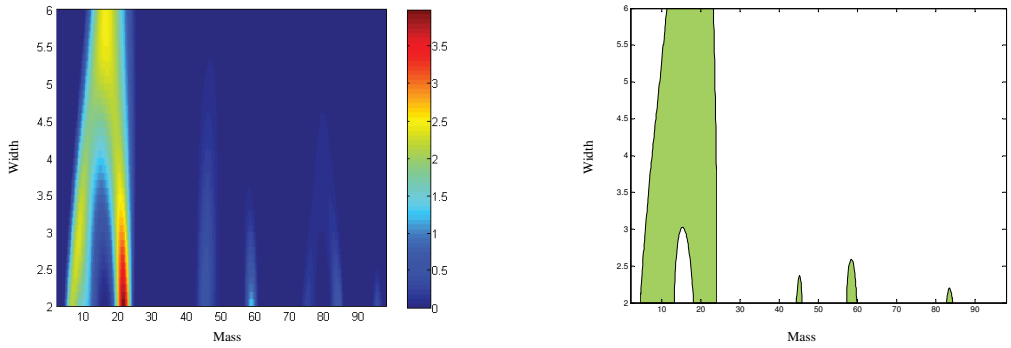


Fig. 4: An example histogram of background events with the signal best fit shown in yellow.

The search space (mass,width) has clearly a non trivial correlation structure, which is evident in Figure 5. The procedure for estimating the p-value is nevertheless identical to that of the previous



(a) (b)

Fig. 5: (a) The values of the test statistic q as a function of the signal mass and the width (b) The corresponding excursion set above $q = 1$.

example, and the expected Euler characteristic is similarly given by Eq. (8), where the difference is only in the numerical values of the coefficients $\mathcal{N}_1, \mathcal{N}_2$. Here we find $\mathcal{N}_1 = 4 \pm 0.2$ and $\mathcal{N}_2 = 0.7 \pm 0.3$, and the predicted p-value is shown in Figure 6 compared to the observed p-value from a set of 200,000 background simulations. Again we find an excellent agreement between the Euler characteristic formula and the observed p-value, demonstrating the usefulness of this result.

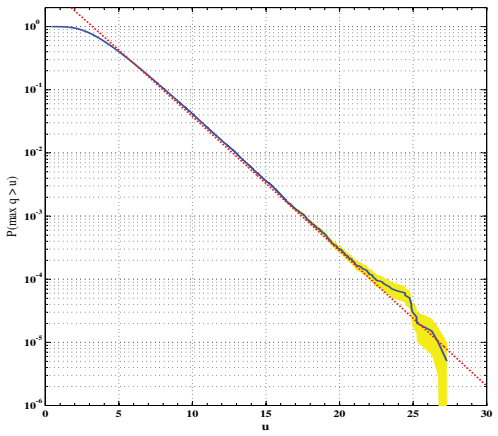


Fig. 6: The prediction of Eq. (8) (dashed red) against the observed p-value (solid blue) from a set of 200,000 background simulations. The yellow band represents the statistical uncertainty due to the available number of background simulations.

6 Summary

The Euler characteristic formula, a fundamental result from the theory of random fields, provides a practical mean of estimating a p-value while taking into account the "look elsewhere effect". This result is valid under general conditions and is therefore applicable to a wide range of problems, as we have demonstrated in the two representative cases studied in this work. This could greatly ease the compu-

tational burden of having to perform large number of Monte Carlo simulations, required to establish a discovery claim.

Acknowledgements

We are grateful to Michael Woodroofe and Luc Demortier, for their helpful comments and discussions during the 2010 Banff workshop on statistical issues [12]. We thank Jim Braun and Teresa Montaruli for their help in providing us the background simulation data of IceCube which was used to perform this analysis.

References

[1] R.B. Davies, *Hypothesis testing when a nuisance parameter is present only under the alternative.*, Biometrika **74** (1987), 33-43.

[2] E. Gross and O. Vitells, *Trial factors for the look elsewhere effect in high energy physics* , Eur. Phys. J. C, **70** (2010), 525-530.

[3] R.J. Adler and A.M. Hasofer, *Level Crossings for Random Fields*,Ann. Probab. **4**, Number 1 (1976), 1-12.

[4] R.J. Adler, *The Geometry of Random Fields*, New York (1981), Wiley, ISBN: 0471278440.

[5] K.J. Worsley, S. Marrett, P. Neelin, A.C. Vandal, K.J. Friston and A.C. Evans, *A Unified Statistical Approach for Determining Significant Signals in Location and Scale Space Images of Cerebral Activation*, Human Brain Mapping **4** (1996) 58-73.

[6] R.J. Adler and J.E. Taylor, *Random Fields and Geometry* , Springer Monographs in Mathematics (2007). ISBN: 978-0-387-48112-8.

[7] G. Cowan, K. Cranmer, E. Gross and O. Vitells, *Asymptotic formulae for likelihood-based tests of new physics*, Eur. Phys. J. C **71** (2011) 1544, [arXiv:1007.1727].

[8] S.S. Wilks, *The large-sample distribution of the likelihood ratio for testing composite hypotheses*, Ann. Math. Statist. **9** (1938) 60-62.

[9] J. Taylor, A. Takemura and R.J. Adler, *Validity of the expected Euler characteristic heuristic*, Ann. Probab. 33 (2005) 1362-1396.

[10] J. Ahrens et al. and The IceCube Collaboration, Astropart. Phys. **20** (2004), 507.

[11] J. Braun, J. Dumma, F. De Palmaa, C. Finleya, A. Karlea and T. Montaruli, *Methods for point source analysis in high energy neutrino telescopes*, Astropart. Phys. **29** (2008) 299-305 [arXiv:0801.1604].

[12] Banff International Research Station Workshop on Statistical issues relevant to significance of discovery claims, http://www.birs.ca/events/2010/5-day-workshops/10w5068.

An alternative view of the Look Elsewhere Effect

G. Ranucci
Istituto Nazionale di Fisica Nucleare, 20133 Milano, Italy

Abstract
The Look Elsewhere Effect, which influences the significance of a potential signal of a particle with *a priori* unknown mass, has a striking counterpart in searches for the presence of a modulation of unknown frequency in a time series of experimental data. In this work, the formulation of the problem in the frequency domain is outlined and the methodology is illustrated using the time series data of the Super-Kamiokande solar neutrino experiment. The parallelism between the mass and frequency domains will be highlighted.

1 Introduction

It has become common practice in high energy physics to denote as the Look Elsewhere Effect (LEE) [1] the statistical implications, especially for significance, of the search for a new signal when a parameter such as a particle mass is unknown. The search for new signals is at the core of the quest for new physics, especially at the dawn of the exciting LHC era. Where to look for new signals is also an aspect of the searches themselves. In this case, the standard statistical treatment for discovery needs to be suitably modified to account for the multiplicity, in term of possible locations, inherent in a given search. In particular, what is decisively affected is the statistical significance of a claimed detection, which changes drastically from the situation in which the mass of the putative particle is known. In this standard case, assuming the background to be reliably estimated, via auxiliary measurements or through detailed Monte Carlo modeling, the usual statistical procedure of comparing the number of counts with the expected background (typically based on the Poisson distribution for counting experiments) holds.

On the other hand, the *a priori* unknown location of the signal complicates this simple picture, because it effectively enhances the background. In effect, the fixed mass background is replaced with some kind of extreme value distribution in which the background process populates the whole search range. The statistical description of the Look Elsewhere Effect encompasses the mathematical tools and procedures needed to describe and quantify numerically such an occurrence. However, it is not frequently appreciated that the mathematics underling all the aspects of this phenomenon has an immediate correspondence in the methodologies that are exploited in the frequency domain while searching for a modulation of unknown period possibly embedded in a noisy data time series. I fully exploit this analogy here to provide an alternative view of the LEE in the frequency domain, not only specifically unraveling its peculiarities and consequences in the time series scanning and analysis, but also underlining the correspondence and parallelism between the description of the effect in the frequency domain and in the usual context we are interested in of a putative signal in a unknown mass range.

2 General framework for this illustration of the LEE: search for modulations embedded in experimental time series

For the purpose of the present discussion it is enough to recall that a popular method to unravel the presence of possible modulations hidden in noisy time series is the computation of the power density spectrum in the frequency domain of the data under study, which are regarded as originated from the sampling of a random process. By denoting with $H(f)$ the Fourier transform of the series, the corresponding power spectrum is simply defined as $|H(f)|^2$. Periodicities hidden in the data would produce sharp, distinct peaks in the power spectrum, which are in principle easily identified. The calculation of the spectrum depends on the nature of the sampling, which may be regular, when the data points

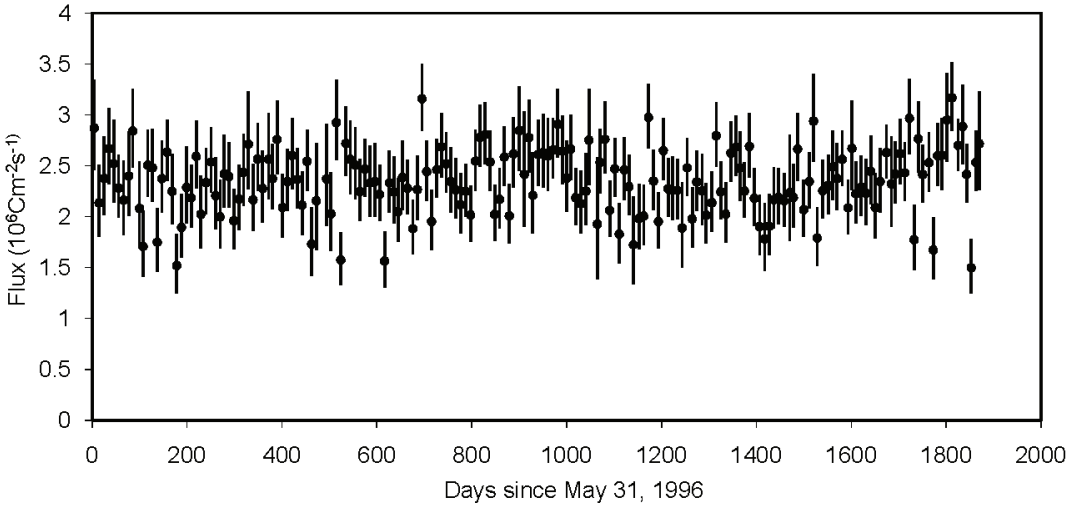


Fig. 1: Time series data of the solar neutrino measurements released by Super-Kamiokande.

are spaced at regular intervals, or as in most practical cases, irregular, i.e. with an unevenly sampled time series. The spectrum implementation in the latter occurrence is performed through the very popular Lomb-Scargle methodology, which produces a power spectrum commonly termed the Lomb-Scargle periodogram [2] [3] (the use of the word periodogram reflects the major emphasis that in this kind of searches is given to the period rather than to the frequency of the searched modulation, especially in astronomical studies of variable phenomena). The Lomb-Scargle periodogram can be viewed as a generalization of the power spectrum estimated via the direct application of the Fourier transform to evenly sampled time series, sometime called Schuster periodogram [3][4][5], not frequently used in practice, but extremely useful to understand the basic statistical features of the spectrum of an experimental time series.

3 A concrete example

To illustrate the method and its statistical implications I use the analysis of the time series data of the Super-Kamiokande solar neutrino experiment [6]. The data officially released by the Collaboration, spanning a 5 year range from April 1996 to July 2001, the so called phase 1 of the experiment, were packed in 10 and 5 day bin format, but I consider here only the 10 day bin series, which for reference is shown in Fig. 1. The analysis of all the released datasets can be found in Ref. [7].

The noise affecting the data can mask a potential low amplitude modulation embedded in the series. The purpose of the frequency analysis is to identify a potential sinusoidal signal, despite the blurring effect of the noise itself.

How a Fourier-related algorithm maps the time series in Fig. 1 into the frequency domain is shown in Fig. 2, specifically illustrating the power spectrum of the series produced by the Lomb-Scargle algorithm (the series indeed is unevenly sampled). The spectrum, naively, is simple to interpret: an unusual high peak should indicate a modulation at that frequency embedded in the data. However, the noise surely present in the time domain has an impact in the frequency domain as well, producing random noise peaks. Therefore, given a high peak in the spectrum, the claim of the detection of a modulation signal at the corresponding frequency has to be confronted with the probability that we are actually dealing with a noise fluctuation. In this context, the Look Elsewhere Effect comes into play because of the a-priori unknown frequency of the sought modulation, therefore the significance of the detection

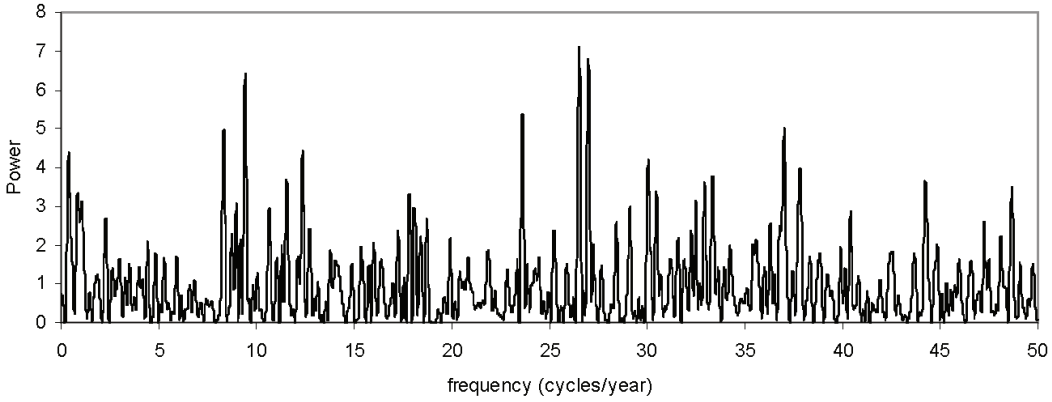


Fig. 2: Lomb-Scargle spectrum of the Super-Kamiokande solar neutrino time series.

is expressed by the probability that a peak as high or higher than the highest peak found in the actual spectrum, can be generated by chance noise fluctuations *at any of the scanned frequencies*. How this can be computed is the subject of the next paragraphs.

4 Formulation and statistical properties of the periodogram

For the sake of brevity the complete expressions of the Lomb-Scargle and Schuster periodograms are not given here, rather the interest reader can refer to the explicit formulation in [2][3][8]. A fundamental statistical property valid for both periodograms is that under the null hypothesis and for gaussian noise affecting the data, the ordinate z of the spectrum at a generic frequency is distributed simply according to e^{-z}, the absence of extra factors in this rather simple noise distribution being due to a normalization term in the periodogram expression containing σ^2, the data variance. It should be highlighted that σ^2 is inferred from the scatter of the data themselves, therefore the errors on the individual data points, even if available, are not taken into account in the standard periodogram framework (a likelihood ratio approach to the spectrum computation, not considered here, would overcome this limitation). Before moving on to the implications of the LEE effect, letÕs consider the search of a signal at a predefined frequency, which occurs when there are reasons to presume the existence of some effect at that specific frequency, with the Look Elsewhere Effect turned off. In this case the ingredients of the detection problem are the usual ones, e.g. the distribution of the ordinate at that frequency in case of absence of the signal (null hypothesis) and the ordinate distribution for the same frequency in case of presence of a signal (alternative hypothesis). The former is simply e^{-z}, while it can be shown that the latter belongs to the family of the non central χ^2 distributions with two degrees of freedom, but is, however, completely defined only if also the presumed amplitude of the sought signal is known. The detection scenario is shown in Fig. 3, where it is possible to distinguish the single frequency exponential noise distribution from three examples of signal distributions for three different signal amplitudes a, the relative amplitude with respect to the average value of the series. Given a threshold, e.g. the straight vertical line in the figure, the significance level of a detection is unambiguously defined as the integral ($p - content$) of the simple exponential noise distribution above it.

5 The Look Elsewhere Effect in action

Now I modify the previous framework moving to a situation in which, lacking the knowledge of where to expect the signal, a whole frequency range is spanned for the search. The usual criterion to decide about the presence of a signal is based upon the height of the largest peak detected in the searched frequency interval, which acts as the test statistics.

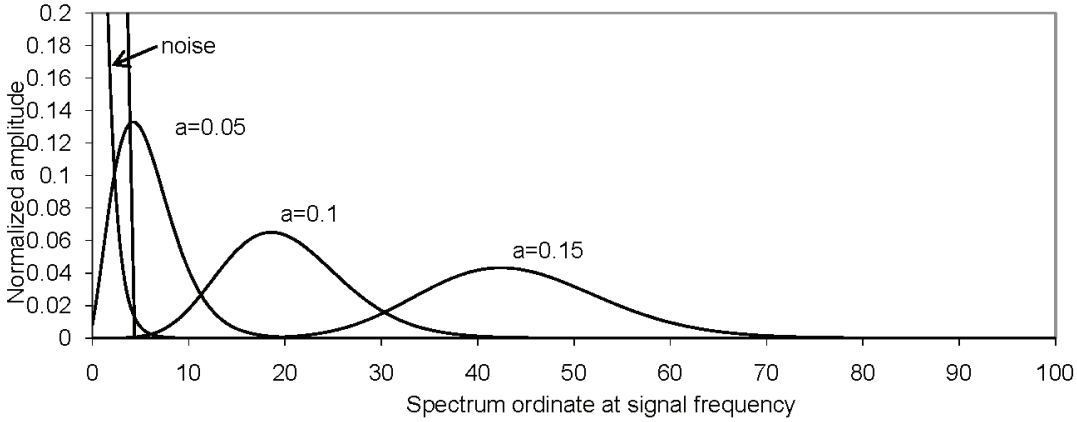


Fig. 3: Single frequency noise distribution together with three different amplitude signal distributions.

The LEE implies the need to modify the detection scenario depicted in Fig. 3 by replacing the single frequency exponential noise PDF with the PDF of the highest peak generated over the entire search band by a pure noisy series (null hypothesis).

The paradigmatic Schuster periodogram is very useful to describe the derivation of this PDF, by exploiting the important frequency analysis result that the direct application of the DFT (Discrete Fourier Transform) to an evenly sampled series with N number of samples generates a spectrum, i.e. the Schuster periodogram, meaningfully computed up to the Nyquist frequency $1/(2T)$, where T is the sampling interval, and which comprises only $N/2$ independent frequencies [5]. Therefore, the spectrum is made of $M = N/2$ independent ordinates each distributed according to e^{-z} under the null hypothesis.

As a consequence, the desired PDF of the height z of the largest among M peaks is given by

$$M\left(1-e^{-z}\right)^{M-1}e^{-z}, \tag{1}$$

where the third factor represents the highest peak, the second factor the occurrence that the remaining $M-1$ peaks do not exceed z and the first factor M the number of possible choices of the highest peak among the M spectral ordinates.

Hence, if we denote with H the ordinate of the largest peak actually detected in the experimental spectrum, the probability to have a chance noise fluctuation as high or higher than H is

$$\int_H^\infty M\left(1-e^{-z}\right)^{M-1}e^{-z}\,dz = 1-\left(1-e^{-H}\right)^M, \tag{2}$$

which can be immediately recognized as a Šidàk-Bonferroni-type [9] formula, typical of Multiple Hypothesis Testing problems.

Following the standard terminology, the above formula gives the $p-value$ of the highest detected peak. If, however, the threshold th for the detection is chosen in advance, then the same formula, with H replaced by th, is more precisely interpreted as the significance (or equivalently the *type I error*) of the detection procedure.

It should be highlighted that this formalism is extendable to peaks of any rank, exploiting the explicit formulation of the PDF of a generic peak of rank i, given by [7]

$$p_i(z|M) = \frac{M!}{(i-1)!(M-i)!}[1-F(z)]^{M-i}\,[F(z)]^{-1}\,p(z), \tag{3}$$

where i is the order of the peak, $i = 1$ being the lowest peak and so on, up to $i = M$, i.e. the highest peak, and

$$F(z) = \int_0^\infty p(\lambda)\, d\lambda. \tag{4}$$

Expression (3) is valid for any form of $p(z)$ and not only for $p(z) = e^{-z}$ as in the problem under study. Furthermore, it is easily verified that Eq. (3) reduces to the Eq. (4) for $i = M$.

6 Application to the Super-Kamiokande time series

The validation of the model in Section 5 and its application to real data time series proceed usually via Monte Carlo (MC) techniques. Indeed, the previous formalism is valid not only for the reference Schuster periodogram, but also in the cases of real practical interest of unevenly sampled data when the Lomb-Scargle periodogram is used. Nevertheless, in such an occurrence the M parameter in the formulae is not derivable by the number of sampling points, but rather can be inferred via simulation. A thorough account of how the MC methodology is applied, via toy models examples, is given in [7]. Here the Monte Carlo estimation is directly employed for the Lomb-Scargle analysis of the measured Super-Kamiokande series; the final goal is to perform a quantitatively statistical inference about the possible presence of a modulation, through the assessment of the p-values of the largest peak(s) in the spectrum.

In general, the Monte Carlo approach envisages the simulation of many synthetic time series generated reproducing the characteristics of the specific time series under study, in particular the data scatter (i.e. the variance). For each simulated series the Lomb-Scargle periodogram is computed and the height of the four highest peaks recorded. At the end of the simulation cycles (10000) the resulting histograms are compared with the respective PDF from Eq. (3).

The MC evaluated distributions of the four highest peaks in a search frequency interval ranging from 0 to 50 cycles/year are shown in Fig. 4. They follow very well the model for a parameter M equal to 529, that in analogy with the paradigmatic Schuster periodogram we can identify as the ÒeffectiveÓ number of independently scanned frequencies in this specific case (see [10] for a clear definition of this concept). Remarkably, not only the highest peak, but also the others are in fairly good agreement with the model.

The MC inferred distributions just obtained are the basis to assess the significance of the signal search: indeed their integral above the height of the peak of corresponding rank (the highest, the second largest, the third largest peak, and so on; the example here is limited to the first four largest peaks) in the spectrum in Fig. 2 gives the desired p-values. The numerical values of the ordinates of the four peaks of interest are such that these integrals can be conveniently evaluated both directly by the MC distributions or through the corresponding model with the M parameter value, 529, inferred from the fit. Anyhow, the latter method would be the more convenient in the occurrence of a very high peak, since in that case in order to get a meaningful p-value from the MC distribution one should run much more than the 10000 simulations used here. The resulting p-values are summarized in the following Table 1; collectively they demonstrate that there is no hint whatsoever of a signal embedded in the series, which thus appears to be perfectly compatible with a pure noise series.

Table 1: Significance of the 4 highest peaks in the spectrum of the Super-Kamiokande time series.

Rank	Frequency (Hz)	Ordinate	Significance
1^{st}	26.51	7.1	34.7%
2^{nd}	26.99	6.8	15%
3^{rd}	9.4	6.41	8%
4^{th}	23.6	5.36	27%

The proper account of the Look Elsewhere Effect via the described MC procedure is essential

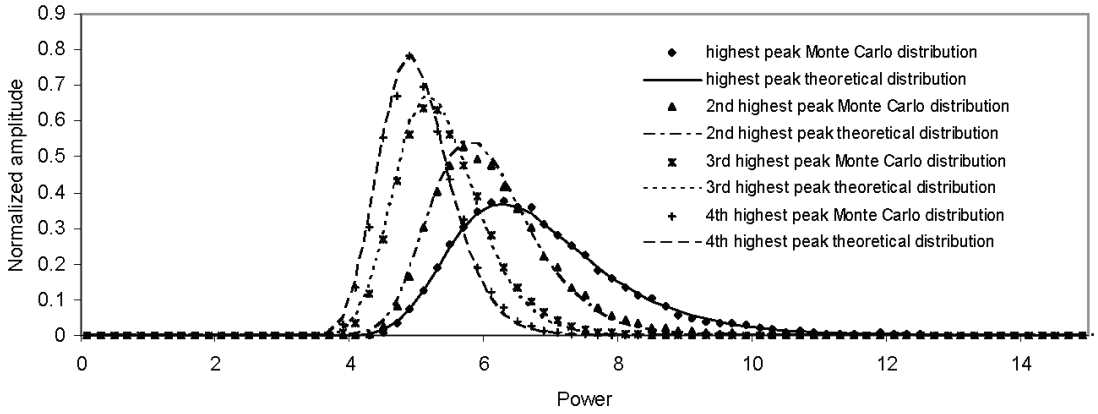


Fig. 4: Comparison of model and Monte Carlo concerning the distributions of the four highest peaks in the Lomb-Scargle periodogram of the Super-Kamiokande data series.

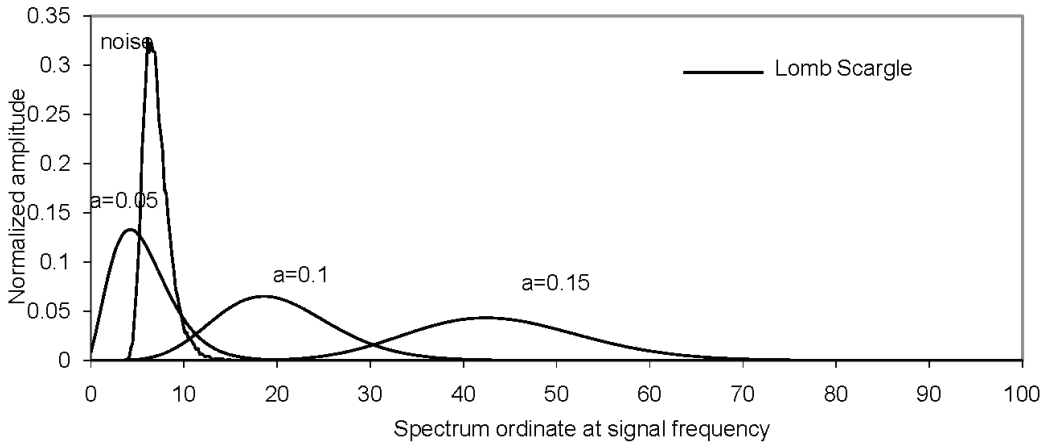


Fig. 5: Effective noise distribution generated by the LEE together with three different amplitude signal distributions.

to reach such a conclusion: for example for the highest 7.1 peak ordinate of the spectrum the p-value for a specific-frequency signal detection from the e^{-z} noise distribution would be 0.08251%, leading to a suspect that something is going on at that frequency. The incorporation of the LEE washes out completely this potential erroneous interpretation.

The overall picture of the impact of the Look Elsewhere Effect can be appreciated in Fig. 5, which reproduced the detection scenario of Fig. 3, but with the single frequency noise distribution now replaced by the effective noise distribution. i.e. the PDF of the highest peak over the search band. We can consider it as the effective noise distribution since it is such PDF that in practice limits the capability to detect small periodic signals embedded in the series. This fact is especially clear in the figure, by noting that the PDF of the highest peak overlaps almost completely the signal induced distribution when the signal amplitude a is small (5% in the figure), hence severely hindering its detection. A detailed description of the methodology and results can be found in Ref. [7], and a similar analysis for the time series data of the solar neutrino experiment SNO is reported in Ref. [11].

7 Parallelism with the search for a unknown mass particle standard problem

The procedure illustrated in the previous Sections to address the LEE in the frequency domain can be easily adapted to describe the LEE effect in the more usual scenario of search for a particle when its mass is unknown. To illustrate this I depict a simple prototype example, described by a toy model consisting of an experimental mass range from 0 to 100, affected by a Poisson background distributed with mean value 500, uniform over the entire mass range. The range is explored for a signal through a set of windows covering the whole interval.

The very basic case is that of a set of W non overlapped, contiguous windows of equal width, depending upon the presumed resolution of the sought signal; in each of them the noise is independently Poisson distributed with mean value $B = 500/W$.

The identification of a signal is denoted by a "significantÓ excess of events above background in any of the search windows. But, how to compute the significance? It is just in this calculation that the parallelism with the frequency search problem appears, establishing a correspondence between the number of explored windows W and the number of independent scanned frequencies M, as shown in the following.

Given the Poisson count rate in each window,

$$p(n|B) = \frac{e^{-B}B^n}{n!}, \tag{5}$$

what limits the detection of a signal is actually the distribution of the largest detected count N over the whole set of searched windows, that stems from that Poisson process. Such a distribution can be inferred by considering all the configurations that produce each specific realization of the random variable N. In particular, a given value N is obtained when in all the windows but one the counts are less than N, while in the residual window the count is exactly N. The probability associated with such a configuration is clearly

$$W \left(\sum_{n=0}^{N-1} \frac{e^{-B}B^n}{n!} \right)^{W-1} \frac{e^{-B}B^N}{N!}, \tag{6}$$

where the factor W takes into account the number of combinations of one window out of the total number W. The parallelism with the frequency case is clearly manifest in Eq. (6), whose three terms are in perfect one to one correspondence with the three factors in Eq. (1). However, in the discrete case there are additional configurations that must be accounted for explicitly. How they modify expression Eq. (6) is not reported here, but a complete derivation can be found in Ref. [12], leading to

$$P_{\max}(N) = \sum_{k=1}^{W} \binom{W}{k} \left(\sum_{n=0}^{N-1} \frac{e^{-B}B^n}{n!} \right)^{W-k} \left(\frac{e^{-B}B^N}{N!} \right)^{k}. \tag{7}$$

Equation (7) is the complete transposition in this context of the Eq. (1).

A graphical representation of the toy model described above shows clearly the implication of this formula. Considering 25 non overlapping windows of width 4, Fig. 6 displays both the individual window background probability function, i.e. the simple Poisson distribution, and the probability function of the highest peak of the 25 scanned windows. As in the frequency case, it is the latter that acts as the effective noise distribution affecting the capability to detect very low level signals, and from which significance and p-values calculations can be inferred. It shows the striking effect of the LEE in action. As in the frequency case, the problem can also be fully addressed via MC, through which, in particular, it has been possible to cross check very accurately the validity of the Eq. (7). Incidentally, we note that this formulation describes also the statistics of the highest bin in a finely binned histogram (under the assumption of constant background). In a histogram like this when one of the bins is anomalously high its $p - value$ is usually given as the single bin $p - value$, from the Poisson function, multiplied by the

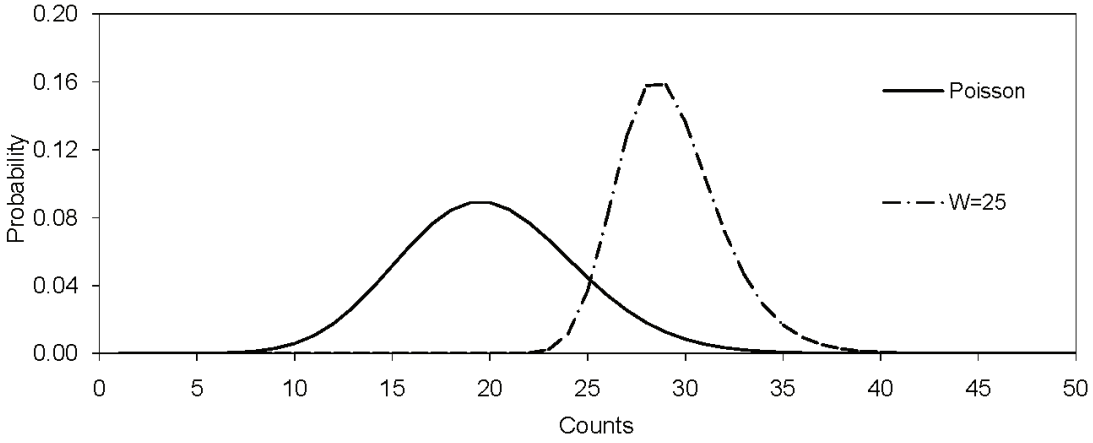


Fig. 6: Comparison of original Poisson distribution with the largest count distribution over 25 non overlapping observation windows, exploring a mass range 0-100 for an excess of events anywhere in that range. The plotted largest count distribution is the concrete manifestation of the LEE in this example.

number of bins. This approximation indeed can be demonstrated to be the asymptotic behavior of Eq. (7) when the single bin Poisson $p - value$ is very low. Finally, the parallelism with the frequency case can be further extended to the concept of effectively scanned windows and to the distributions of the second highest peak, the third, etc., in the explored mass range, as explained in detail in Ref. [12].

8 Conclusion

The search for a signal of unknown location, either in mass or frequency, is affected by noise fluctuations larger than those pertaining to a fixed location search, as described by the Look Elsewhere Effect. In this work a thorough illustration of the LEE in the frequency analysis has been given, showing its implications in the search for modulations embedded in an experimental time series, using as a concrete example the Super-Kamiokande solar neutrino data. Furthermore, a parallelism has been established between the approach in the frequency domain and a description of the search for particles of unknown mass, which led to the identification of interesting correspondences between the manifestation of the LEE in both cases, providing useful, additional insights to this rather peculiar effect.

Acknowledgement

I wish to thank the organizers who allowed me to contribute to such an enlightening workshop.

References

[1] E. Gross and O. Vitells, "Trial factors for the look elsewhere effect in high energy physics", The Eur. Phys. J. C, **70**, Issue 1-2, 525-530 (2010).

[2] N.R. Lomb, “Least-squares frequency analysis of unequally spaced data", Astrophysics and Space Science, **39**, 447-462 (1976).

[3] J.D. Scargle, “Studies in astronomical time series analysis. II - Statistical aspects of spectral analysis of unevenly spaced data", Astrophys. J. **263**, 835-853 (1982).

[4] A. Schuster, “On the investigation of hidden periodicities with application to a supposed 26 day period of meteorological phenomena", Terr. Mag. Atoms. Elect., **3**, 13-41,(1898).

[5] W.H. Press *et al.*, “Numerical recipes in Fortran ", Sect. 12.1 (Cambridge University Press,1991).

[6] J.Yoo *et al.*, “Search for periodic modulations of the solar neutrino flux in Super-Kamiokande-I", Phys. Rev. D **68**, 092002 (2003).

[7] G. Ranucci, “Likelihood scan of the Super-Kamiokande I time series data", Phys. Rev. D, **73**, 103003 (2006).

[8] W.H. Press *et al.*, “Numerical recipes in Fortran ", Sect. 13.8 (Cambridge University Press, 1991).

[9] Z. Šidàk, “Rectangular confidence region for the means of multivariate normal distributions", J. Am. Stat. Assoc. **62**, 626Ð633 (1967).

[10] J.H. Horne and S.L. Baliunas, “A prescription for period analysis of unevenly sampled time series", Astrophys. J., **302**, 757-763 (1986).

[11] G. Ranucci and M. Rovere, “Periodogram and likelihood periodicity search in the SNO solar neutrino data", Phys. Rev. D **75**, 013010 (2007).

[12] G. Ranucci, “On the significance of signal search through the ‘sliding window’ algorithm", Nucl. Instrum. Methods Phys. Res. A, **562**, 433-438 (2006).

RooStats for Searches

Grégory Schott, on behalf of the RooStats team
KIT, Institut für Experimentelle Kernphysik, Karlsruhe, Germany

Abstract
The RooStats toolkit, which is distributed with the ROOT software package, provides a large collection of software tools that implement statistical methods commonly used by the High Energy Physics community. The toolkit is based on RooFit, a high-level data analysis modeling package that implements various methods of statistical data analysis. RooStats enforces a clear mapping of statistical concepts to C++ classes and methods and emphasizes the ability to easily combine analyses within and across experiments. We present an overview of the RooStats toolkit, describe some of the methods used for hypothesis testing and estimation of confidence intervals and finally discuss some of the latest developments.

1 Introduction

The RooStats project [1, 2] is a collaborative open source project initiated by members of ATLAS, CMS and the CERN ROOT team. The RooStats toolkit — based on previously existing code used in ATLAS [3] and CMS [4], which has been extended and improved — has been distributed with ROOT since summer 2008. The toolkit provides and consolidates statistical tools needed for LHC analyses and allows one to apply and compare the most popular and well-established statistical approaches. Thanks to readily available well-known tools, results across experiments can be better understood and compared. This is not only a desirable feature but also a required one when it comes to combining analysis results as will be discussed later. Finally, the RooStats project aims to provide reasonably flexible, well-tested, documented tools. The RooStats developments benefit from scientific oversight from the statistics committees of both experiments.

In High Energy Physics, the goal of an analysis is usually to test a prediction or search for new physics, leading to the estimation of the statistical significance of a possible observation or the construction of confidence intervals — often expressed as upper or lower limits in case of a non-observation. The most common statistical procedures are:

- point estimation: i.e., the determination of the best estimate of parameters of the model,
- confidence or credible interval estimation: i.e., regions representing the range of parameters of interest compatible with the data,
- hypothesis tests: i.e., comparing the data to two or more hypotheses,
- goodness of fit: to quantify how well a given model describes the observed data.

RooStats aims to cover some of these common statistical procedures.

The RooStats package is built on top of RooFit [5], which is a data modeling toolkit developed originally within the BaBar collaboration and now integrated into ROOT. The most crucial element of RooFit is its ability to model probability densities, likelihood functions, and data, in a very flexible way that can deal with arbitrarily complex cases. Some recent developments in RooFit provide additional tools specifically needed by RooStats. The RooStats code is organized into three groups of classes: *calculators* that perform the statistical calculations, *results* and *utilities* that facilitate the RooStats work flow.

After a few generalities, given in Sect. 2, the classes implementing statistical inferences and results are discussed in Sect. 3. In Sect. 4, we describe RooStats utilities, while Sect. 5 will have a few words on some applications and perspectives.

2 Generalities

We begin by clarifying some of the terminology commonly used:

- *Observables*: quantities that are measured by an experiment (e.g., mass, helicity angle, output of a neural network) that form a *data set*.
- *Model*: the probability density function (PDF) — either parametric or non-parameteric — that describes one or multiple observables and normalized so that their integral over any observable is unity.
- *Parameters of interest*: parameters of the model whose value we wish to estimate or constrain (e.g., a particle mass or a cross-section).
- *Nuisance parameters*: uncertain parameters of the model other than the ones of interest (e.g., parameters associated with systematics, such as normalization or shape parameters). The treatment of nuisance parameters varies according to the statistical approach.

2.1 Likelihood Function

The modeling of the likelihood function is the principal task of RooFit. RooFit, which builds on ROOT, maps mathematical concepts to RooFit classes. For example, variables, functions, probability densities, integrals, a space point, or a list thereof, are handled by `RooRealVar`, `RooAbsReal`, `RooAbsPdf`, `RooRealIntegral`, `RooArgSet` and `RooAbsData`, respectively. A large collection of functions are available to describe the PDF. The functions are handled by classes inheriting from `RooAbsPdf` and can be easily combined to build arbitrarily complex models through addition, multiplication, and convolution. For both data and models there exist some binned and unbinned representations. For each model, integration and maximum likelihood fitting is supported and utilities are provided for the Monte Carlo generation of pseudo data, in order to perform "toy" studies, and for the visual inspection of results. The utilities and great modularity of RooFit are the principal factors that drove the choice of RooFit as the basis of RooStats. One can work with arbitrarily complex data and models and one can handle large sets of observables and parameters.

Most statistical methods usually start with a likelihood function. A rather general likelihood function, for use in our field, with multiple observables, can be written as:

$$L(\mathbf{x}|r,s,b,\theta_{\mathbf{s}},\theta_{\mathbf{b}}) = e^{-(rs+b)}\prod_{j=1}^{n}[rsf_s(\mathbf{x_j}|\theta_{\mathbf{s}}) + bf_b(\mathbf{x_j}|\theta_{\mathbf{b}})]. \quad (1)$$

The PDFs f_s and f_b represent the distributions of observables $\mathbf{x}$ for the signal and background, with parameters $\theta_{\mathbf{s}}$ and $\theta_{\mathbf{b}}$, respectively. The parameters s and b — typically, the expected signal and background counts, respectively — are constrained by the number n of observed events[1]. In this likelihood function a strength factor r multiplies the expected number of signal events[2].

2.2 Model Configuration

Before one can perform a statistical inference, it is necessary to specify the model: the PDF of possible observables, the actual observables, the parameters of interest, the nuisance parameters, possibly a Bayesian prior, etc. The RooStats calculators can be configured, via the constructor, either with the model specifications given as individual RooFit objects or with a `ModelConfig` object, in which the

[1] Sometimes described as an extended likelihood; it can also be viewed as the limit of a binned multi-Poisson likelihood function with arbitrarily small bins.

[2] This is sometimes done to redefine the parameter of interest such that r is the ratio of the signal production cross-section to the expected value of the cross-section. For example, in the search for the Standard Model Higgs boson, obtaining a 95% CL upper-limit for $r = 1$ means the Standard Model Higgs hypothesis can be excluded at 95% CL.

model specification is bundled. For most of the calculators both configuration mechanisms are available. The idea behind `ModelConfig` is to provide a uniform way to configure calculators. The downside is that it becomes less obvious what elements of the `ModelConfig` are necessary for a given calculator. For example, the prior probability will not be used in frequentist-based calculations while the list of observables, which is mainly used to generate pseudo-data, is not needed when computing Bayesian limits.

The model is often completed by a set of observed data. Moreover, the calculators can be configured for a number of options specific to the statistical algorithms (e.g., number of Monte Carlo iterations, size of the test, test statistic, etc.). Finally, the calculator is run and returns the result of a hypothesis test or a confidence interval.

3 RooStats Calculators

Below, we describe the RooStats calculators, which are based on the following conceptual approaches:

- *Classical or Frequentist*: this school of statistics restricts itself to statements of the form "probability of the data given the hypothesis". Probability is interpreted as a limit of relative frequencies of various outcomes.
- *Bayesian*: this school of statistics views probability more broadly, which permits statements of the form "probability of the hypothesis given the data". Typically, probability is interpreted as a "degree of belief" in the veracity of an hypothesis.
- *Likelihood*: this approach uses a frequentist notion of probability (e.g., it does not require the specification of a prior for the hypothesis), but inferences are not guaranteed to satisfy some frequentist properties (e.g., coverage). Like the Bayesian approach, this likelihood approach obeys the likelihood principle, while frequentist methods do not.

We give a brief description of the methods available in RooStats and refer the reader to textbook literature for details (see, for example [6, 7]).

As can be seen from Fig. 1, there are two general classes of calculators in RooStats: those performing hypothesis-tests and those computing confidence or credible intervals, which inherit, respectively, from the classes `HypoTestCalculator` and `IntervalCalculator` and return, respectively, objects inheriting from the classes `HypoTestResult` or `ConfInterval`.

The `IntervalCalculator` interface allows the user to provide the model, the data set, the parameters of interest, the nuisance parameters and the size α of the test ($\alpha = 1 - \mathrm{CL}$, where CL is the confidence/credible level). After configuring the calculator, a `ConfInterval` pointer is returned via the method `IntervalCalculator::GetInterval()`. Depending on the calculator used, a different type of `ConfInterval` will be returned (e.g., connected interval, multi-dimensional interval, etc.) but each shares the ability to test if a point lies within the interval using the method `ConfInterval::IsInInterval(p)`.

The `HypoTestCalculator` can be configured with the model, the data and parameter sets specifying the two hypotheses to be tested. Through `HypoTestCalculator::GetHypoTest()`, a pointer to the result can be retrieved and the result object can be queried for p-values and the corresponding significances, or Z-values, found by equating a p-value to a one-sided Gaussian tail probability and solving for the number of standard deviations. In this convention, a p-value of 2.87×10^{-7} corresponds to a Z-value of 5σ.

3.1 Profile-Likelihood Calculator

The `ProfileLikelihoodCalculator` class implements a likelihood-based method to estimate a confidence interval and to perform an hypothesis test for a given parameter value. To illustrate the method, let

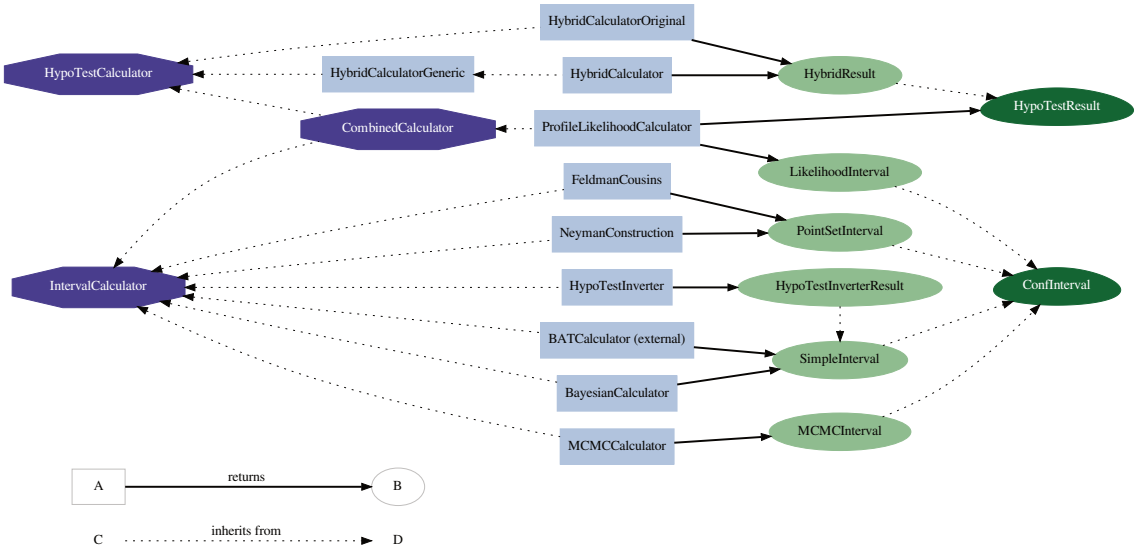


Fig. 1: Diagram of the interfaces for hypothesis testing and confidence interval calculations and classes used to return the results of these statistical tests.

us assume that the likelihood function depends on a set K parameters θ, one of which is the parameter of interest. From the likelihood function $L(\mathbf{x}|\theta_0, \theta_{i\neq 0})$, similar to the one of Eq. (1) but where the parameter of interest r has been renamed θ_0, for generality, the profile likelihood function is the numerator in the ratio:

$$\lambda(\theta_0) = \frac{L(\theta_0, \hat{\hat{\theta}}_{i\neq 0})}{L(\hat{\theta}_0, \hat{\theta}_{i\neq 0})}. \tag{2}$$

The denominator, $L(\hat{\theta})$ is the absolute maximum of the likelihood, while the numerator is the maximum value of the likelihood for a given value of θ_0.

Under certain regularity conditions, Wilks's theorem demonstrates that asymptotically $-2\ln\lambda(\theta_0)$ follows a χ^2 distribution. In the asymptotic limit, the likelihood ratio test statistic $\lambda(\theta_0)$ has a parabolic shape:

$$-2\ln\lambda(\theta_0) = -2(\ln L(\theta_0) - \ln L(\hat{\theta}_0)) = n_\sigma^2, \text{ with } n_\sigma = \frac{\theta_0 - \hat{\theta}_0}{\sigma}, \tag{3}$$

where n_σ represents the number of Gaussian standard deviations associated with the parameter θ_0. From this construction, it is possible to obtain the one- or two-sided confidence intervals (see Fig. 2). Owing to the invariance property of the likelihood ratios, it can be shown that this approach remains valid for non parabolic log-likelihood functions. This method is also known as MINOS in the physics community, since it is implemented by the MINOS algorithm of the Minuit program. Given the fact that asymptotically $-2\ln\lambda$ is distributed as a χ^2 variate, an hypothesis test can also be performed to distinguish between two hypotheses characterized by different values of θ_0.

In this approach, systematic uncertainties are taken into account by augmenting the likelihood function with terms that encode the knowledge we have of the systematic uncertainties and the profiling is now done over all nuisance parameters including those for the systematics.

This likelihood-based technique for estimating an interval and performing a hypothesis test is provided in RooStats by the `ProfileLikelihoodCalculator` class. The class implements both the `IntervalCalculator` and `HypoTestCalculator` interfaces. When estimating an interval, this calculator returns a `LikelihoodInterval` object, which, in the case of multiple parameters of interest, rep-

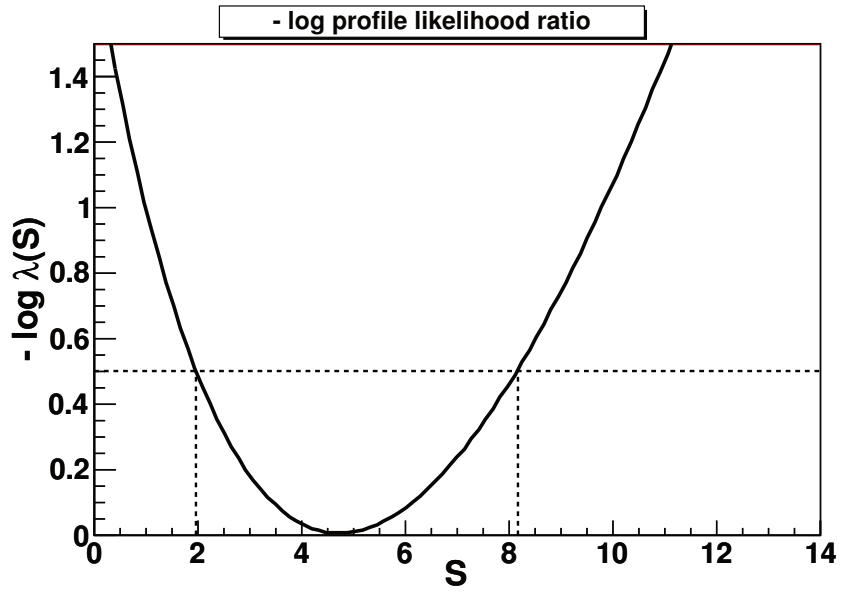


Fig. 2: Plot of the log profile likelihood curve as function of the parameter of interest, $\theta_0 \equiv S$. The 1σ interval (68% CL) is obtained from the intersect of the $-\log\lambda$ curve with the horizontal dashed line $-\log\lambda = 0.5$.

resents a multi-dimensional contour. When performing a hypothesis test, a `HypoTestResult` object is returned with the significance for the null hypothesis. Another class exists, `LikelihoodIntervalPlot`, to visualize the likelihood interval in the case of one or two parameters of interest (as shown in Fig. 2). A newly developed class, `ProfileInspector`, allows inspection of the value of the nuisance parameters for each value of the parameter of interest along the profile log-likelihood curve.

3.2 Bayesian Calculators

Bayes theorem relates the probability (density) of a hypothesis given data to the probability (density) of data given a hypothesis. The inversion of the probability is achieved by multiplying the likelihood function (the probability of the data given an hypothesis) by a prior probability for the model, which is characterized by parameters of interest and, typically, one or more nuisance parameters. This product is normalized so that the integral of the posterior density, over all parameters, is unity. The calculation of credible intervals, that is, Bayesian confidence intervals, requires the calculation of the cumulative posterior distribution. In the Bayesian approach, nuisance parameters are removed by marginalization, that is, by integrating over their possible values. RooStats provide two different types of Bayesian calculator, the `BayesianCalculator` and `MCMCCalculator` classes, depending on the method used for performing the required integrations.

The current implementation of the `BayesianCalculator` class works for a single parameter of interest and uses numerical integration to compute the posterior probability distribution. Various algorithms provided by ROOT for numerical integration can be used, including those based on Monte Carlo integration, such as implemented in the programs Vegas or Miser. The result of the class is a one-dimensional interval (`SimpleInterval`) obtained from the cumulative posterior distribution.

The `MCMCCalculator` uses a Markov-Chain Monte Carlo (MCMC) method to perform the integration. The calculator runs the Metropolis-Hastings algorithm, which can be configured by specifying parameters such as the number of iterations and burn-in-steps, to construct the Markov Chain. Moreover, it is possible to replace the default uniform proposal function with any other proposal function. The result of the `MCMCCalculator` is a `MCMCInterval`, which can compute the confidence interval for the desired

parameter of interest from the Markov Chain. The `MCMCInterval` integrates the posterior density from its mode downwards until the interval has a $1-\alpha$ probability content[3]. The `MCMCIntervalPlot` class can be used to visualize the interval and the Markov chain.

Users can also input the RooStats model into the Bayesian Analysis Toolkit (BAT) [8], a software package that implements Bayesian methods via Markov-Chain Monte Carlo. In the latest release, BAT provides a class, `BATCalculator`, which can be used with a similar interface to the RooStats `MCMCCalculator` class. Developments are foreseen that will further integrate BAT within RooStats.

3.3 Neyman Construction

The Neyman construction is a pure frequentist method to construct an interval at a given confidence level, $1-\alpha$, such that coverage is guaranteed for fully-specified probability models. A detailed description of the method is given in Ref. [6]. RooStats provides a class, `NeymanConstruction` that implements the construction. The class derives from `IntervalCalculator` and returns a `PointSetInterval`, a concrete implementation of `ConfInterval`.

The Neyman construction requires the specification of an ordering rule that defines the order in which potential observations are to be added to the interval in the space of observations until the desired confidence level is reached. The ordering rule is usually specified in terms of a specific test statistic. Consequently, the RooStats class must be configured with this information before it can produce an interval. More information can now be provided with the introduction of the interfaces `TestStatistic`, `TestStatSampler`, and `SamplingDistribution`. Different test statistics are available, including:

- Simple likelihood ratio: $Q = L_1(\theta_0 = 1)/L_0(\theta_0 = 0)$,
- Ratio of profiled likelihoods: $Q' = L_1(\theta_0 = 1, \hat{\hat{\theta}}_{i\neq 0})/L_0(\theta_0 = 0, \hat{\hat{\theta}}'_{i\neq 0})$,
- Profile likelihood ratio: $\lambda(\theta_0) = L_1(\theta_0, \hat{\hat{\theta}}_{i\neq 0})/L_0(\hat{\theta}_0, \hat{\theta}_{i\neq 0})$.

Another aspect to decide is how to sample it: assuming asymptotic distribution, generating toy-MC experiments with nuisance parameters fixed (used in `NeymanConstruction`) or with nuisance parameters sampled according to a prior distribution (used in `HybridCalculator`.

Common configurations, such as the Feldman-Cousins approach — where the ordering is based on the profile likelihood ratio as the test statistic [9], can be enforced by using the `FeldmanCousins` class. A generalization of the Feldman-Cousins procedure, when nuisance parameters are present, generating toy Monte Carlo experiments with nuisance parameters fixed as described in [3, 10], is also available.

The Neyman construction considers every point in the parameter space independently. Consequently, there is no requirement that the interval be connected nor that it have a particular structure. The result consists of a set of scanned points labeled according to whether they are inside or outside the interval (`PointSetInterval` class). The user either specifies points in the parameter space that are to be used to perform the construction or a range and a number of points within the range, which will be scanned uniformly in a grid. For each scanned point, the calculator will give the sampling distribution of the chosen test statistic. This is typically obtained by toy Monte Carlo sampling, but other techniques exist and can, in principle, be used. In particular, newly developed code may be helpful when testing hypotheses with very small p-values through the application of importance sampling techniques.

3.4 Hybrid Calculator

This calculator implements a Bayesian/frequentist hybrid approach for hypothesis testing. It consists of a frequentist toy Monte Carlo method, as in the Neyman construction, but with a Bayesian marginalization of nuisance parameters [11]. This technique is often referred to as a "Bayesian-Frequentist Hybrid".

[3] It should be noted that these *highest posterior density intervals* are not invariant under under one-to-one reparametrisation.

For example, let us define the null hypothesis, H_0, to be the background-only or no signal hypothesis, and H_1 to be the alternate hypothesis that a signal is present along with background. In order to quantify the degree to which each hypothesis is favoured or excluded by the experimental observation, one chooses a test statistic which ranks the possible experimental outcomes. Given the observed value of the test statistic, the p-values, $CL_{sb} \equiv p_1$ and $CL_b \equiv 1 - p_0$, can be computed. Since the functional forms of the test statistic distributions are typically not known a priori, a large number of toy Monte Carlo experiments are performed in order to approximate these distributions. Figure 3 provides an example of such distributions from the two pseudo data sets and where the observed value of the test statistic lies.

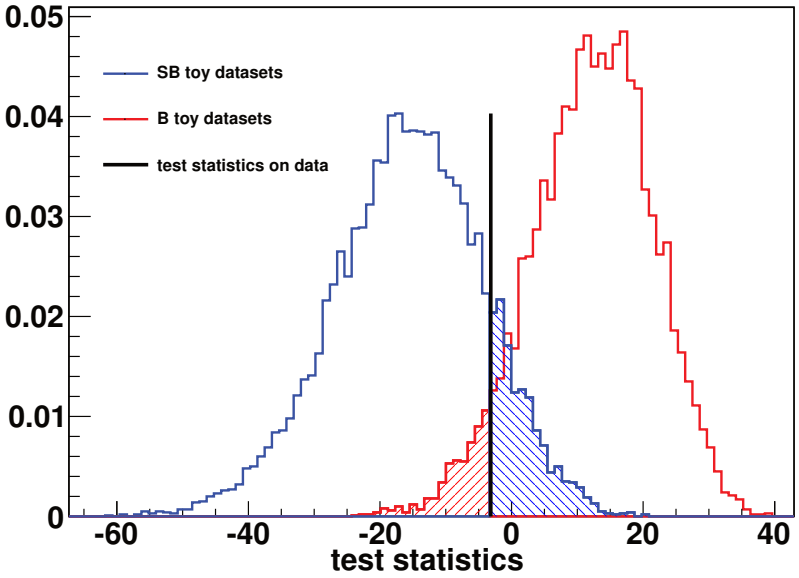


Fig. 3: Result from the hybrid calculator, the distributions of a test statistic in the background-only (red, on the right) and signal+background (blue, on the left) hypotheses. The vertical black line represents the value obtained on the tested data set. The shaded areas represent $1 - CL_b$ (red) and CL_{sb} (blue).

Systematics uncertainties are taken into account through Bayesian marginalization. For each toy Monte Carlo experiment, the values of the nuisance parameters are sampled from their prior distributions before generating the toy sample. The net effect it to broaden the distribution of the test statistic, as expected in the presence of systematic uncertainties, and thus degrade the separation of the hypotheses.

This procedure is implemented in RooStats by the `HybridCalculator` class. The input to the class are the models for the two hypotheses, the data set and, optionally, the prior distribution for the nuisance parameters, which is sampled during the toy generation process. As for the `NeymanConstruction`, the test statistic can be freely parameterized. The results of the `HybridCalculator` consists of the test statistic distribution for the two hypothesis, from which the hypothesis p-value and associated Z-value can be obtained. Since the simulation of the distributions could be computationally expensive, RooStats permits different results to be merged, which makes it possible to run the calculator in a distributed computing environment. The `HybridPlot` class provides a way of plotting the result, as shown for example in Fig. 3.

By varying the parameter of interest representing the hypothesis being tested (for example, the signal cross-section) one can obtain a one-sided confidence interval (e.g., an exclusion limit). RooStats provides a class, `HypoTestInverter`, which implements the interface `IntervalCalculator` and performs the scanning of the hypothesis test results of the `HybridCalculator` for various values of one parameter of interest. By finding where the confidence level curve of the result intersects the de-

sired confidence level, an upper limit can be derived, assuming the interval is connected. An estimate of the computational uncertainty is also provided. Finally, when defining exclusion limits, the condition that defines the upper bound can be chosen: either one can use the p-value p_1 of the alternate hypothesis (the pure-frequentist approach) or the ratio of p-values $CL_s = p_1/(1 - p_0)$ (modified-frequentist approach [12]).

4 RooFit and RooStats Utilities

4.1 RooFit's Workspace

One element of RooFit whose addition has been driven by the development of the RooStats project (although it would still be useful even without RooStats) is the `RooWorkspace` class. It is a container for RooFit objects that can be written to a ROOT file. When a RooFit object is imported from a file (e.g.,, a complex PDF with multiple parameters), all the other dependent objects are imported too. Later, it is very easy to rebuild and initialize all the parameters, to reconstitute the original PDF, via a single recall from the `RooWorkspace` (while still permitting adjustments to the imported object). These features make it possible to save the complete likelihood function, as well as the data, to a file in a well defined fashion, either as a technical convenience, as an intermediate step towards the combination of the results of multiple analyses or for the grander purpose of electronic publication of these results. In addition, the `RooWorkspace` interfaces to a newly developed utility, `RooFactoryWSTool`, which permits the building of a large class of RooFit objects in an interpreted mode with an intuitive syntax based on strings. Multiple dependent parameters are also defined, created and stored in the `RooWorkspace` on-the-fly, thereby allowing, for example, the creation of a Gaussian PDF in one line, instead of the four needed to create one (the PDF along with its observable and two parameters) using the RooFit classes directly. It will be discussed later how this factory tool is complemented by RooStats' `HLFactory` class.

4.2 User-Friendly Model Specification

Tools that simplify and automate the description of complex models in a user-friendly way are usually referred to as model factories. There are currently two such utilities provided within RooStats: `HLFactory` and `HistFactory`. Their use is optional. For more experienced users or in more complex cases, direct use of the lower level RooFit classes may be preferred.

`HLFactory` is a RooStats class whose aim is to disentangle the C++ code doing the calculations from the physics-driven and analysis-specific description of the probability models. The later can be written to a single text file describing all (and only) the physics inputs that are to be processed later in a single line of code. The fact that `HLFactory` is built as a simple wrapper around the `RooWorkspace` factory utility sidesteps the need to define yet another language that a user would have to learn, while not restricting the application to specific analyses since this model factory supports everything the `RooWorkspace` factory does. In addition, python-like instructions are added that allow better structuring of the description (through includes) and along with comments on the analysis model. Finally (and optionally), the `HLFactory` also allows the easy combination of multiple channels to form a combined model and combined data set.

`HistFactory` is a collection of classes to handle template histogram-based or binned analyses. It allows such analyses to use RooStats without requiring knowledge of the RooFit modeling language; instead, the likelihood function and elements of the statistical analysis are specified through an XML configuration file, which is used to produce the model. In this approach, the user provides histogram templates of one observable and of models for different contributing samples (e.g., of the signal and background processes). Then, the normalization in terms of number of events for each of these channels can be decomposed — for example, as a product of luminosity, efficiency, cross-section terms — each of which can be affected by systematic uncertainties. It supports Gaussian, gamma and log-normal distributions for nuisance parameters. Finally, histograms of variations can be provided that specify the

related systematic changes. Multiple channels can be given and combined and parameters which are identical across channels can be easily identified.

4.3 Other Utilities

Not all utilities are listed in this document. Here we mention briefly three more:

- `SPlot`, a class implementing a technique used to produce weighted plots of an observable distribution in a multi-dimensional likelihood-based analysis [13].
- `RooNonCentralChiSquare`, a class in RooFit that outlines the use of a generalization of Wilks' theorem called Wald's theorem which states that the asymptotic distribution of the test statistic $\lambda(\mu)$ for $\mu \neq \mu_{true}$ is a non-central χ^2 [14],
- `BernsteinCorrection`, a class that augments the nominal probability with a positive-defined polynomial given in the Bernstein basis, which can be used as an approach to incorporate systematic effects in a PDF.

5 Statistical Combinations and Perspective

The combination of results is a commonly used method for improving sensitivities or measurements of signals. With RooStats, the combination can be performed at the analysis level in contrast to combinations performed at the level of published results. This means that the global likelihood function for the ensemble of the analyses to be combined is explicitly written and the statistical analysis is performed on this combined likelihood. This approach has advantages, such as being able to account for known correlations consistently. But, it also has its inconvenience, such as making the likelihood function a quite complex object. One strong motivation for the RooStats project was to simplify the process of combining analyses by providing a tool that allows this to be done simply for arbitrarily complex models.

In December 2010, ATLAS and CMS created the LHC-HCG group mandated to prepare and produce a combined Higgs result from the LHC (with similar efforts also on-going in other analysis groups within the collaborations). RooStats will be used for the combination and one of the first tasks of the group has been to complement its validations with comparison to results obtained from independent software in specific analysis cases[4]. While the validations appear satisfactory so far, the RooStats team will keep improving interfaces and fix performance issues as well as develop new complementary tools based on users' experiences and feedback.

One aspect of statistical data analysis is left open by RooStats, namely that of the choice of statistical method. In that respect, it allows the implementation of one recommendation of the ATLAS and CMS statistics committees, which is that various methods be applied and compared (although different methods are not expected to give the same results since they have different properties and provide answers to different questions). A more specific method and statistical procedure to use when combining ATLAS and CMS analyses is a topic still under discussion and one of the focuses of this PHYSTAT conference.

Acknowledgements

The RooStats contributors are thankful to the members of the ATLAS and CMS statistics committees for the exchange of ideas, advice and encouragement. I also wish to thank L. Lyons and the rest of PHYSTAT committee for the organization of the very rich and useful conference and the invitation to present there progress on the development of the RooStats toolkit.

[4]For further insights on these activities see Ref. [15]

References

[1] L. Moneta *et al.*, *The RooStats project*, PoS ACAT2010, 057 (2010) [arXiv:1009.1003].

[2] RooStats homepage: https://twiki.cern.ch/twiki/bin/view/RooStats/WebHome.

[3] K. S. Cranmer, *Statistics for the LHC: Progress, challenges and future*, proceedings of PHYSTAT 2007, CERN-2008-001, 47 (2007).

[4] D. Piparo, G. Schott and G. Quast, *RooStatsCms: a tool for analysis modelling, combination and statistical studies*, J. Phys. Conf. Ser. 219, 032034 (2010) [arXiv:0905.4623].

[5] W. Verkerke, *Statistical software for the LHC*, proceedings of PHYSTAT 2007, CERN-2008-001, 169 (2007).

[6] F. James, *Statistical methods in experimental physics, 2nd edition*, Word Scientific (2006).

[7] K. Nakamura *et al. The Review of Particle Physics - Chapter 33*, J. Phys. G37, 075021 (2010).

[8] A. Caldwell, D. Kollar, K. Kröninger, *BAT: The Bayesian Analysis Toolkit*, Comput. Phys. Commun. 180, 2197 (2009) [arXiv:0808.2552].

[9] G. Feldman and R. D. Cousins, *Unified approach to the classical statistical analysis of small signals*, Phys. Rev. D57, 3873 (1998).

[10] K. S. Cranmer, *Frequentist hypothesis testing with background uncertainty*, proceedings of PHYSTAT 2003 [physics:0310108].

[11] R. D. Cousins and V. L. Highland, *Incorporating systematic uncertainties into an upper limit*, Nucl. Instrum. Meth. A320, 331 (1992).

[12] A. L. Read, *Modified frequentist analysis of search results (The CLs method)*, CERN OPEN-2000-205 (2000).

[13] M. Pivk and F. R. Le Diberder, *SPlot: A statistical tool to unfold data distributions*, Nucl. Inst. Meth. A555, 356 (2005) [physics:0402083].

[14] G. Cowan *et al.*, *Asymptotic formulae for likelihood-based tests of new physics*, Eur. Phys. J. C71, 1554 (2011) [arXiv:1007.1727].

[15] K. S. Cranmer, *Combining ATLAS and CMS Higgs searches*, proceedings of PHYSTAT 2011 (these proceedings).

Bayesian Analysis Toolkit in Searches

Frederik Beaujean[1]*, Allen Caldwell*[1]*, Daniel Kollár*[2]*, Kevin Kröninger*[3]*, Shabnaz Pashapour*[3*]
[1] Max-Planck-Institut für Physik, [2] CERN,[3] Georg-August-Universität Göttingen
*Corresponding Author

Abstract
The Bayesian Analysis Toolkit, a software package for data analysis based on Bayes' theorem, is introduced. This toolkit takes advantage of Markov Chain Monte Carlo to find the full posterior probability distributions. The tool can easily be used for parameter estimation, limit setting and error propagation. Model comparison and goodness-of-fit estimation are realized in the package through well-established methods. In addition to a brief description of the Bayesian Analysis Toolkit, the use of this tool in searches is described in the example of Banff Challenge 2a problem 1.

1 Introduction

A comprehensive statistical interpretation is an essential part of any data analysis. Typically, one needs to compare model predictions with data, to draw conclusions on the validity of the model as a representation of the data and to extract values of parameters. It is not trivial to implement the required tools for such a task and usually individual researchers develop their own versions of these tools. Therefore, it is beneficial to have a set of common statistical tools and numerical algorithms that can be validated regularly and easily adapted to solve arbitrary problems.

One of the problems to be addressed is to search for the presence of a signal over some background. For example, in the current status of particle physics that the Higgs boson has not yet been observed and there is no clear direction on what nature has in store for us as the new physics, we need to have the ability to spot the smallest signals in a reliable way to continue our path in better understanding the workings of nature.

In this article, we describe the Bayesian Analysis Toolkit (BAT) [1, 2], its philosophy and functionalities as well as the use of this toolkit to address one of the Banff challenge problems [3].

2 The Bayesian Analysis Toolkit

The Bayesian Analysis Toolkit is a C++ software package to address statistical problems based on Bayes' theorem. It comes in form of a library working with ROOT [4], developed to provide the users with easy to implement methods. The tool has an interface with CUBA [5], MINUIT [6] and RooStats [7]. Bayes' theorem for a single model has the form

$$P(\vec{\lambda}|\vec{D}) = \frac{P(\vec{D}|\vec{\lambda})P_0(\vec{\lambda})}{\int P(\vec{D}|\vec{\lambda})P_0(\vec{\lambda})\,d\vec{\lambda}}\ , \tag{1}$$

i.e., the probability of parameter set $\vec{\lambda}$ given data $\vec{D}$, the *posterior* probability, is proportional to the probability of data given parameters, also known as the *likelihood*, times the initial probability for the parameters, the *prior* probability[1]. The denominator ensures the normalization of the posterior probability and is just the integral of the numerator over the allowed region of $\vec{\lambda}$. One can also interpret this formula as a learning rule stating that: The knowledge about the model and its parameters before the experiment, the prior, is updated using the probability of the new data for different values of the parameters, resulting in posterior knowledge.

The approach taken in the BAT technical development is to fulfill two main requirements:

[1] Throughout this article the term probability is used for both probability and probability density.

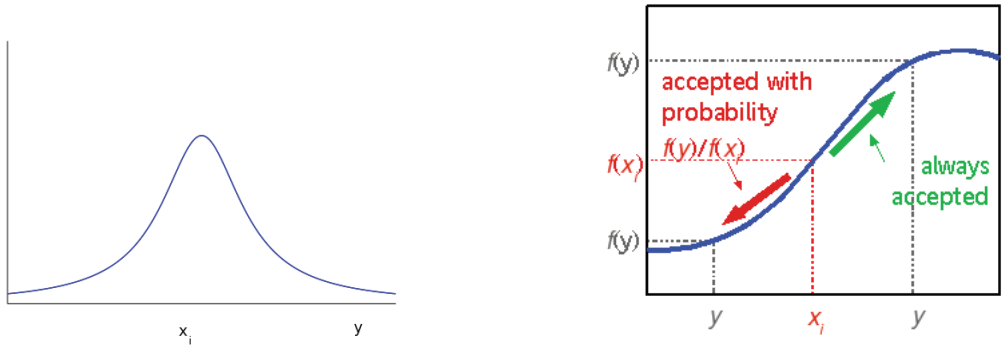


Fig. 1: The proposal function to pick the point to move to for the MCMC random walk (left), and a simple sketch to show the decision-making process in the MCMC process (right).

i) provide a flexible framework which allows formulation of arbitrary models,
ii) provide a reliable mapping of the full posterior probability density.

For each data analysis, there is a case specific part which includes the model and the data and the other part consists of the common tools. The user only needs to deal with the case specific issues and the rest is taken care of by the common tools in BAT. The user should create the model by defining the parameters, $\vec{\lambda}$, the likelihood, $P(\vec{D}|\vec{\lambda})$, and the priors, $P_0(\vec{\lambda})$, and read in the data. Common tasks such as normalization, mode finding, goodness-of-fit test, marginalization and presentation of the output in a nice format are handled by BAT functions. The key tool in BAT, allowing for the mapping of the posterior probability in multidimensional parameter space and the extraction of quantities of interest, is the Markov Chain Monte Carlo.

2.1 Markov Chain Monte Carlo

The feasibility of Bayesian inference has been revolutionized by the use of Markov Chain Monte Carlo (MCMC) (see e.g., [8, 9]). The MCMC can be used to obtain the posterior probability given by Eq. 1 which otherwise is generally a difficult task, specially for models with a large number of parameters. The MCMC can be employed to scan very complicated probability distributions in many dimensions through a random walk to points with higher probabilities in the allowed parameter space. The Metropolis algorithm [10] is the first and the most popular MCMC algorithm and is implemented in BAT. This procedure is followed to map out a function $f(\vec{x})$:

1. start at a random $\vec{x_i}$
2. generate a random point around $\vec{x_i}$, the proposal point, according to a proposal function,
3. calculate the values of the function at the current point, $\vec{x}_i$, and the proposal point, $\vec{y}$, and compare them:
   - if $f(\vec{y}) \geq f(\vec{x}_i)$, set $\vec{x}_{i+1} = \vec{y}$,
   - if $f(\vec{y}) < f(\vec{x}_i)$, set $\vec{x}_{i+1} = \vec{y}$ with probability $r = f(\vec{y})/f(\vec{x}_i)$,
   - if $\vec{y}$ is not accepted, stay where you are, $\vec{x}_{i+1} = \vec{x}_i$.
4. generate a new $\vec{y}$ around the new $\vec{x}$ (go to 2).

For an infinite number of steps, the $f(\vec{x_i})$ is gauranteed to converge to $f(\vec{x})$. However, for a finite number of steps, one has to check for convergence. Figure 1 shows the proposal function, a Cauchy distribution, and a schematic of how the MCMC works.

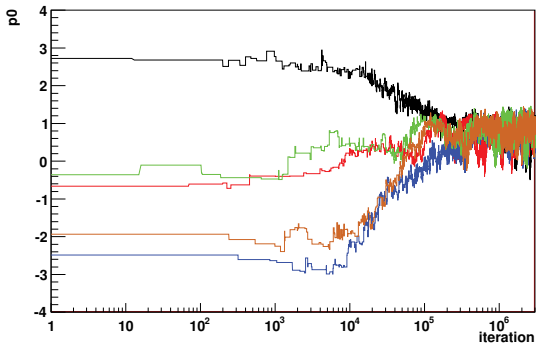


Fig. 2: Parameter value at each iteration for 5 Markov chains that shows the convergence of the 5 chains after aprroximately 10^5 iterations.

2.2 Convergence

To achieve the convergence of MCMC and find reasonable run parameters a *pre-run* is performed in BAT before the MCMC is used for the analysis of the posterior. In the pre-run phase, we use several chains in the parameter space in parallel. The steps in parameter space are done consecutively for each parameter and chain.

The set of steps from an update of the first parameter of the first chain to the last parameter of the last chain makes one iteration. The efficiency for accepting or rejecting new points is evaluated separately for each parameter and chain over a number of iterations. The proposal function is set to a Cauchy function by default and the width of the function is adjusted during the pre-run to match the required efficiency of the sampling. After a finite number of steps are taken, for example 1000 steps, we update the width of the proposal function to optimize performance, until an efficiency between 15% to 50% is reached for each parameter. Users can also define their own proposal function.

The convergence is determined based on the *R-value* [11] which should be about 1 once the convergence is reached. The R-value is the ratio of the current variance estimate to the within-sequence variance with a factor to account for the extra variance of the Student's t distribution. Figure 2 shows the parameter value, $p0$, in five different chains as a function of the iteration. Convergence, defined via the R-value, is reached after aprroximately 10^5 iterations.

2.3 Main Analysis Run and Functionality

The analysis run is performed for a defined number of iterations with run parameters found in the pre-run. At this stage all the scales are fixed and samples are collected for posterior analysis. At each step for each parameter a 1-dimensional (1-D) histogram is filled with the values of the parameter and the so-called marginalized distribution is obtained. This represents the posterior probability function of a single parameter of a model given data when all the other parameters are integrated over.

Similarly one can create 2-dimensional (2-D) histograms for every pair of parameters when all the other parameters are integrated over. This will show the correlation between the two parameters. The full output of the MCMC can be saved during the run for future analysis. Figure 3 shows an example of these distributions for a 6-parameter model. It enables you to compare all the 6 parameters and their correlations in one plot.

In addition to the posterior probability functions, BAT can be used for other statistical calculations. Evaluation of arbitrary functions of parameters is also implemented in BAT and typically is used for error propogation. Given that MCMC covers the full parameter space, the location of the global maximum is updated at each step. Therefore, a bi-product of the MCMC is the location of the global mode of the

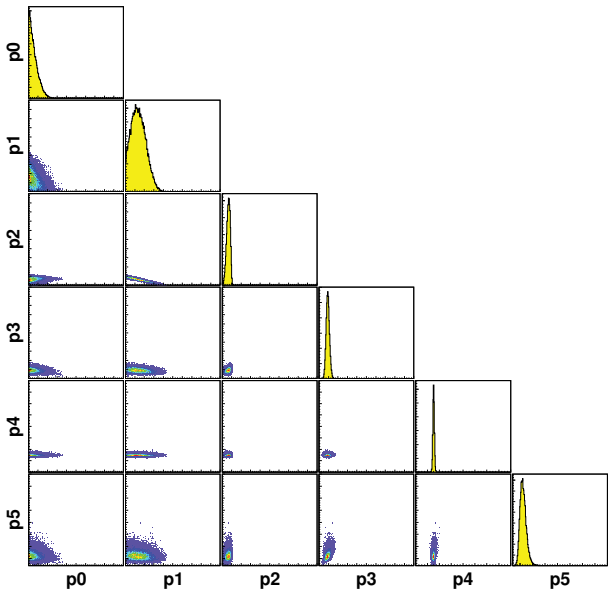


Fig. 3: An example plot : 1-D and 2-D marginalized distributions for a 6-parameter, p0 to p5, model, allowing one to easily examine the parameters and their correlations.

posterior probability. MCMC is not optimized for this task and as such the estimated mode is not accurate but it can be used as a good starting point for other minimization programs, e.g., MINUIT. A numerical integration over the posterior probability is also possible in BAT using the sampled mean algorithm with and without importance sampling. Alternatively, one can compile BAT with the CUBA library which allows the use of well-tuned routines for integration in many dimensions.

3 Signal Discovery

One of the most common tasks in data analysis is to look for the presence of a signal over some background. Usually the general form of background is known and there are predictions for the signal model. The task is to check for the presence of signal over the background given the data at hand. Here a method to search for signal using BAT is suggested. As an example the result of the Banff Challenge 2a Problem 1 is discussed.

A simple strategy to look for the signal is to define the models under investigation, a null hypothesis, $P_0(H_1)$, for background only case, a possible signal including the background, $P_0(H_2)$, such that the total probability of the two cases is one, $P_0(H_1) + P_0(H_2) = 1$. You also need to set the possible values for the parameters given a signal is present, $P(\mu, A, \sigma|H_2)$. Here, μ is the average value of the observed data, A is the rate for the background and σ is the width of the signal model. Variables μ, A, and σ are the parameters for the model. After defining the models, one can calculate the posterior probabilities for the hypothesis and check the validity of the models under investigations.

Once one decides on the choice of priors and probabilities, it is very straightforward to implement this simple strategy in BAT. One just needs to define the parameters and their ranges, the priors and the likelihood method and can read the data, all of which can be done by using the most basic class in BAT, `BCModel` class, or one of its inheritance. Afterwards, the functions available in BAT can be used to calculate the posterior probabilities and the p-value, to provide the marginalized 1D/2D distributions, the full MCMC output and to do many other tasks.

3.1 An Example - Banff Challenge 2a Problem 1

Several interesting statistical issues have been raised in the workshop at the Banff International Research Station on Statistical Issues Relevant to Significance of Discovery Claims [3]. These issues have been illustrated as specific examples to be addressed by the participants. One of these examples is to simulate the task of discovering a signal or new phenomena. The details of the Banff challenge can be found in [3].

In brief, two hypothesis have been considered, a background only case in the form of

$$B(x) = Ae^{-Cx} \,, \tag{2}$$

where x is the mark of the event and is between 0 and 1, $C = 10 \pm 0$, and the background rate is drawn from a truncated Gaussian distribution such that $A = 10000 \pm 1000$ and $A \geq 0$. The second hypothesis considers a signal, in addition to the above background, of the form

$$S(x) = De^{-(x-E)^2/2\sigma^2} \,, \tag{3}$$

where $D \geq 0$, and $\sigma = 0.03$ and, in the signals generated for the simulated data, the peak position E is between 0 and 1 exclusive of both sides. Participants are given 20K simulated datasets with a mixture of the two hypotheses to analyse and provide a yes-no decision whether a signal is to be claimed. The Type-I error should not be more than 0.01 and if a signal is claimed, the peak position and its 68% interval should also be reported. No prior was provided.

Our approach was to follow a Bayesian logic. A two-step decision procedure has been employed. We start with a background only fit to the data and calculate the p-value. If the p-value is greater than 1%, we discard the possibility of the signal presence, if it is less than 1%, we further analyse the data. We bin the data and choose our likelihood as the product of Poisson probabilities for each bin. We use BAT's fast Poisson p-value estimate corrected for the degrees of freedom [12] to get the p-value. For the background rate, we consider a truncated Gaussian distribution prior in a range of 0 to 20K. If we had made a signal discovery claim based on the p-value cut, we would have a Type-I error of 0.0138 ± 0.0009.

However, we do not make a claim based on p-value cut, instead we start with step two and calculate the probability of background only model, $P(B|D)$. Our prior for the background only model is chosen to be 0.95. In other words, we take a case where we have strong prior belief in the null hypothesis. Furthermore, we set the strict requirement that $P(B|D) < 0.001$ to claim evidence for the presence of a signal. Flat priors are assumed for E and D with respective ranges of 0 to 1 and 0 to 2000. The "Look Elsewhere Effect" is already taken care of with the choice of priors [13]. Our Type-I error for our final result is 0.0 which is an important result stating we have no false positive.

Using this strategy, 271 datasets were flagged as having signal. Figure 4 shows the measured peak position vs. the measured signal rate for the datasets claimed to have a signal. For the found signals, 63% of the measured peak positions and 50% of the signal rates fall within 68% of their true values.

The result will change by the choice of different priors, however, in the real experiment, you have to make a judgment based on how far out the new physics is believed to be.

4 Summary

The Bayesian Analysis Toolkit is introduced. The philosophy of its design, its functionalities and general characteristics are briefly discussed. The implementation and performance of Markov Chain Monte Carlo in BAT is described. As an example, the use of BAT in search for a signal is outlined and the result for the Banff Challenge 2a problem 1 is reported.

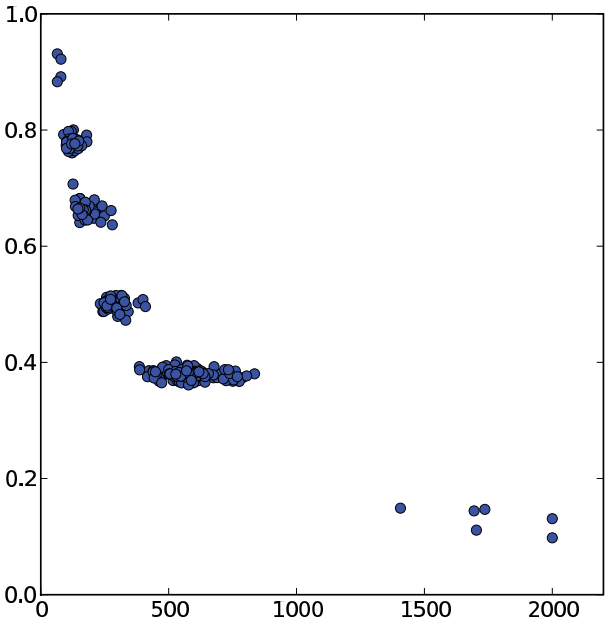


Fig. 4: The measured peak position, E, vs. the measured signal rate, D, for the datasets claimed to have a signal.

References

[1] A. Caldwell, D. Kollár and K. Kröninger, *BAT - The Bayesian Analysis Toolkit*, Comp. Phys. Comm. **180**, 2197 (2009).

[2] `http://www.mppmu.mpg.de/bat/`

[3] T. R. Junk, *Banff Challenge 2,* these Proceedings.

[4] R. Brun and F. Rademakers, *ROOT - An object oriented data analysis framework*, Nuclear Instruments and Methods in Physics Research **A**, 81 (1997).

[5] T. Hahn, *CUBA – a library for multidimensional numerical integration*, Comp. Phys. Comm. **168**, 78 (2005).

[6] F. James, *MINUIT - Function Minimization and Error Analysis*, CERN Program Library Long Writeup **D506** (1994-1998).

[7] W. Verkerke and D. Kirkby, *The RooFit toolkit for data modeling*, [arXiv:physics/0306116] (2003); L. Moneta *et al. The RooStats project*, [arXiv:1009.1003v2] (2011).

[8] S. Karlin and H. Taylor, *A first course in Stochastic processes*, Academic Press (1975).

[9] W.R. Gilks, S. Richardson and D.J. Spiegelhalter, *Markov Chain Monte Carlo in Practice*, Chapman (1996).

[10] N. Metropolis *et al. Equation of State Calculations by Fast Computing Machines*, J. Chem. Phys. **21**, 1087 (1953).

[11] A. Gelman and D.B. Rubin, *Inference from iterative simulation using multiple sequences*, Statistical Science **7**, 457 (1992).

[12] F. Beaujean *et al.*, *p-values for model evaluation*, Phys. Rev. D **83**, 012004 (2011).

[13] A. Caldwell, *Signal discovery in sparse spectra: a Bayesian analysis*, these Proceedings.

Highlights from PHYSTAT 2011

Glen Cowan
Physics Department, Royal Holloway, University of London, Egham TW20 0EX, UK

Abstract
The PHYSTAT 2011 Workshop held at CERN from 17-20 January, 2011, brought together particle physicists working on statistical data analysis together with astrophysicists and statisticians to exchange ideas and report on recent developments. Highlights from the first three days of the meeting are summarized here. The fourth day, devoted to the problem of unfolding (deconvolution) is covered elsewhere in these proceedings.

1 Introduction

The highlights from PHYSTAT 2011 fall into several broad categories: frequentist methods, Bayesian methods, and tools and applications, reviewed below in Sections 2, 3 and 4, respectively. Apologies are extended in advance for any personal bias or imbalance in the emphasis of topics covered.

2 Frequentist methods

The frequentist methods discussed at PHYSTAT 2011 include use of order statistics for discovery [1], issues related to setting limits [2, 4], and treatment of systematic uncertainties using profile or marginal likelihoods [5, 6]. Additional contributions in this area covered the multiple testing problem or "look-elsewhere effect" as well as methods for combining results [7, 8].

2.1 Order statistics for discovery

Statistical tests for the discovery of new physics have often focused on specific signal models, and the test is thus optimal if the new model is correct. One runs the risk, however, of only discovering the types of phenomena that have been thought of in advance. It is therefore important to carry out some tests that probe the data in a more general way and explore departures from the Standard Model expectations that are not motivated by specific alternatives.

An example of this was presented by Cox [1], who proposes using order statistics as a basis for various tests. Suppose one carries out a test at n positions (these could be, e.g., the n bins of a histogram), and obtains as a result a set of n p-values $P_1, \ldots, P_n$. These can be transformed using $Z = -\ln P$ and then ordered, i.e., $Z_{(1)} \leq Z_{(2)} \leq \ldots \leq Z_{(n)} = \max Z_j$.

A plot of the ordered Z can then be used descriptively and forms the basis for formal tests. Under the null hypothesis, the Z values form a straight line of unit slope. A single outlying point thus indicates an easily identifiable alternative. An incorrectly modeled shape for the histogram would lead to a smooth curve, and if the bins are correlated then one obtains a straight line but with a slope different from unity.

Such a method has a clear application in Particle Physics in a search for a bump in a histogram. Here, however, one would need to use a modified version of the test where the departure from the null (i.e., the new signal) is smeared out over several adjacent bins.

2.2 Frequentist limits

Before the existence of a given signal process is well established, it is often of interest to test the signal model using different values of its parameters and to see which values can be excluded. Specifically, one is often interested in testing parameters related to the overall rate of the signal process, and seeing which

values can be excluded on the grounds that the predicted rate is too high relative to what is observed in the data. This corresponds to using a one-sided test to obtain an upper limit on the rate.

A long recognized difficulty with such one-sided tests is that effectively all physically allowed values of the signal rate may be excluded. For any unbiased test the probability to reject a given signal rate under the assumption of the background-only hypothesis is at least equal to the size of the test, e.g., $\alpha = 0.05$. This holds true even for very small signal rates, that is, for signal models to which one effectively has no sensitivity. If the number of events in data fluctuates low relative to the expected background, then one may end up rejecting even a very low signal rate and thus setting a limit that is substantially smaller than the intrinsic resolution of the measurement.

Already in the era of the LEP Higgs searches, the CLs procedure [9] was developed to prevent data fluctuations from leading to unrealistically strong limits. Here, the p-value of the hypothesized signal model is divided by one minus the p-value of the background-only hypothesis, and the signal model is only excluded if this ratio is found below a small threshold α. The threshold thus plays the role of the significance level of the test, but it is not quite the same because the ratio of p-values is necessarily greater than the p-value of the signal model (the numerator of the ratio). Thus one is less likely to exclude a given signal rate and the CLs upper limit is in general higher than the corresponding limit based on a simple one-sided test.

More recently, as mentioned by Demortier [2], it has been suggested to only regard a parameter value as excluded if its p-value is below the test size α and also if one has sufficient sensitivity to that value. As a measure of sensitivity, one can require a certain minimum probability of discovering the signal (i.e., rejecting the background-only model) if it is true. This is essentially the power of a test of the background-only hypothesis with respect to the signal alternative, hence the name "Power-Constrained Limits" or PCL.

Alternatively one may take as the measure of sensitivity the power of a test of the signal model with respect to the background-only alternative. This is the approach recently used by the ATLAS Collaboration [3].

A similar approach is taken in the method described by van Dyk [4, 10] for reporting the results of a search for an astrophysical source. The proposed procedure is to give: (1) whether the source was detected, (2) a confidence interval for the source intensity and (3) the sensitivity of the observation, quantified as the minimum source strength for which one would have a given detection probability. Van Dyk has emphasized the importance of communicating both the usual upper limit (called an upper bound in the astrophysics community) as well as the sensitivity, rather than only the maximum of these two numbers.

2.3 Systematic uncertainties in likelihood-based tests

An important element of any analysis, frequentist or Bayesian, is ensuring that the model adequately describes the data. This means that one must have an accurate representation for the probability of an outcome, say, x, as a function of the model parameters θ. If this model is not sufficiently accurate, then the situation can be improved by including additional nuisance parameters into the model. These provide an added degree of flexibility so that for some point in the enlarged parameter space, the model will be closer to the truth. Of course as more nuisance parameters are included, one's sensitivity to the parameters of interest is diminished.

Conway [5] and Röver [8] brought up an important issue concerning two distinct methods for dealing with nuisance parameters: profiling and marginalization. Suppose originally one measures x, and the experiment is modeled with the likelihood $L(x|\theta)$, where θ is a parameter of interest. Now it may be that the model is found to be inadequate, and so it is enlarged by inserting a nuisance parameter ν, so one now has the likelihood $L(x|\theta, \nu)$.

To constrain the nuisance parameter, one may set up a control measurement y with likelihood

$L(y|\nu)$. This now becomes part of the full likelihood. If x and y are independent, this is simply

$$L(x, y|\theta, \nu) = L(x|\theta, \nu)L(y|\nu) \ . \tag{1}$$

To eliminate the nuisance parameter ν, one can form the profile likelihood

$$L_{\rm p}(x, y|\theta) = L(x, y|\theta, \hat{\hat{\nu}}(\theta)) \ , \tag{2}$$

where the value $\hat{\hat{\nu}}(\theta))$ is the value of ν that maximizes the likelihood for the specified θ.

Alternatively, in the Bayesian approach one can regard the measurement y as supplying the prior information about the nuisance parameter ν. This prior can be written

$$\pi(\nu) \propto L(y|\nu)\pi_0(\nu) \ , \tag{3}$$

where $\pi_0(\nu)$ reflects prior knowledge about ν even before the control measurement y. (It could be called the ur-prior, using the German prefix for original or primordial.) In the Bayesian approach one eliminates the nuisance parameter by integrating to find the marginal likelihood,

$$L_{\rm m}(\theta) = \int L(x|\theta, \nu)\pi(\nu)\, d\nu \ . \tag{4}$$

The point to notice here is that the observed value of y is taken once to determine the prior $\pi(\nu)$, but is then not viewed as a quantity that varies upon repetition of the experiment. Thus if one simulates measurements based on the model (4), it is only x that is generated. In contrast, the profile likelihood (2) models both the measurements x and y. One must therefore simulate both of these values to determine the distribution of a statistic based on $L_{\rm p}(\theta)$.

It is easy to show that in simple cases, e.g., Gaussian measurements and a constant ur-prior, the two approaches (profiling and marginalization) are equivalent. And in the examples shown by Conway [5] and Röver [8], essentially no difference between the two methods can be seen. Trotta and Cranmer [11], however, discussed cases where this is not true.

A further important example where the two methods are not equivalent is when the main measurement x is discrete, and the control measurement y is continuous. For example, x could represent a Poisson-distributed number of events, and y could be a correction factor related to the efficiency, modeled as following a Gaussian distribution. The distribution of a test statistic based on the marginal likelihood (4), will be discrete, since the dependence on the continuous value y has disappeared after integration over the nuisance parameter. Therefore confidence intervals based on the marginal likelihood will suffer from the over-coverage that is well known in discrete problems. A statistic based on the profile likelihood (2), however, follows a continuous distribution, because it retains a dependence on the continuous variable y. Thus the over-coverage due to discreteness is absent, which should be regarded as an advantage of this approach.

A further important aspect of tests based on the profile likelihood ratio is that one can use analytic formulae to approximate the distributions needed to carry out statistical tests. The formulae are exact only in the large sample limit, but for practical examples the approximations were shown to be reasonably accurate for surprisingly small data samples [6].

2.4 The look-elsewhere effect

Important progress was reported by Vitells [7] and Ranucci [12] on the problem of multiple testing, usually called in particle physics the 'look-elsewhere effect'. The problem often relates to finding a peak in a distribution when the peak's position is not predicted in advance. In the frequentist approach using

a p-value, one must determine the probability, under the background-only hypothesis, to find a peak as significant as the one found more more so anywhere in the search region.

The 'brute-force' solution to this problem involves generating data under the background-only hypothesis and for each data set, fitting a peak of unknown position and recording a measure of its significance. To establish a discovery one often requires a p-value less than 2.9×10^{-7}, corresponding to a 5σ effect. Thus determining this with Monte Carlo requires generating and fitting an enormous number of experiments, perhaps several times 10^7.

In contrast, if the position of the peak were known in advance, then the fit to the distribution would be much faster and easier, and furthermore one can in many cases use formulae valid for sufficiently large samples that bypass completely the need for Monte Carlo (see, e.g., [6]). But this 'fixed-position' p-value would not be correct in general, as it assumes the position of the peak was known in advance.

Vitells described a method that allows one to modify the p-value computed under assumption of a fixed position to obtain the correct value by use of a relatively small Monte Carlo calculation. Suppose a test statistic q_0 is observed to have a value u, and the model contains a nuisance parameter θ (such as the peak position) which is only defined under the signal model (there is no peak in the background-only model). Then Vitells and Gross [13] find that the desired p-value can be written

$$P(q_0 > u) \approx N_1 e^{-u/2} + \frac{1}{2} P(q_0(0) > u) \; , \tag{5}$$

where $P(q_0(0) > u)$ is the 'fixed-position' p-value, and N_1 is the mean number of 'upcrossings' of the the statistic q_0 above the level u. The value of N_1 can be estimated by finding the number of upcrossings above some much lower value, u_0, from a relation due to Davis [15],

$$N_1 \approx \langle N_{u_0} \rangle e^{u_0/2} \; . \tag{6}$$

By choosing u_0 sufficiently low, the value of N_1 can be estimated by simulating, say, only 100 experiments, rather than the 10^8 needed for a 5σ discovery.

Gross and Vitells also indicate how to extend the correction to the case of more than one parameter, e.g., where one searches for a peak of both unknown position and width, or for searching for a peak in a two-dimensional space, such as an astrophysical measurement on the sky [14]. Here one may find some number of regions where signal appears to be present, but within those regions there may be islands or holes where the significance is lower. In the generalization to multiple dimensions, the number of upcrossings of the test statistic q_0 is replaced by the expectation of a quantity called the Euler characteristic, which is roughly speaking the number of disconnected regions with significant signal minus the number of 'holes'.

It should be emphasized that an exact accounting of the look-elsewhere effect requires that one specify where else one looked, e.g., the mass range in which a peak was sought. But this may be have been defined in a somewhat arbitrary manner, and one might have included not only the mass range but other variables that were also inspected for peaks but where none was found. Thus it perhaps not worth expending great effort on an exact treatment of the look-elsewhere effect, as one would do in the 'brute-force' method mentioned above. Rather, the more easily obtained fixed-position p-value can be reported along with an approximate correction to account for the range of parameter space in which the effect could have appeared.

Ranucci [12] also reported on the analogous problem in a time-series analysis. That is, if one examines any time series long enough, a feature that appears significant will eventually appear. The methods proposed for dealing with this problem are similar to those reported by Vitells.

3 Bayesian methods

In Bayesian statistics, probabilities are assigned to hypotheses (e.g., parameter values), in contrast to the frequentist approach where one only speaks of the probability of (repeatable) data outcomes. Given, say, a parameter θ and data x, Bayes' theorem is used to find the posterior probability of θ given x,

$$p(\theta|x) \propto L(x|\theta)\pi(\theta) \ , \tag{7}$$

where $L(x|\theta)$ is the likelihood and $\pi(\theta)$ is the prior probability. An important difficulty in the Bayesian approach stems from the requirement to supply priors. Although one may wish in cases to have these reflect a complete lack of prior information, it has long been realized that this is not a uniquely defined concept.

For example, Bayesian methods have a long tradition in particle physics for the problem of a Poisson distributed value n used to make inference about the mean μ. A widely used prior pdf for μ is the (improper) constant prior for $\mu \geq 0$, which is often thought of as reflecting a complete lack of prior knowledge about μ. This is not really true, as it is not invariant under a change of parameter (e.g., it is not flat in $\ln \mu$), and furthermore it cannot possibly represent a meaningful degree of belief. Nevertheless it provides a simple benchmark and has been widely used.

A. Caldwell [16] proposed elicitation of prior probabilities through consensus of the particle physics community. As impossible as this task may sound, it could prove to be an interesting exercise and may well result in useful benchmarks.

In the absence of meaningfully usable prior information one may try to determine priors from formal rules, as described in the review by Kass and Wasserman [19]. A pioneering element of this approach is the reference prior due to J. Bernardo and J. Berger [20]. These were addressed by several speakers at this meeting and some important points are summarized in Sec. 3.1.

A further important element of Bayesian statistics that has not yet found wide application in particle physics is Bayesian model selection, which was addressed by Berger [18] and reviewed below in Sec. 3.2. Implicit Bayesian methods were described by Demortier [2] and the example of Approximate Bayesian Computation (ABC) is summarized in Sec. 3.3.

3.1 Reference priors

The particle physics community has been reluctant to assign subjective prior probabilities to important model parameters, no doubt driven by the desire to remain 'objective' and also because of the difficulties in reaching any sort of consensus on what these probabilities should be. A general prescription for prior probabilities from formal rules is thus very attractive to the community, and so the method of reference priors due to Bernardo and Berger [20] can perhaps provide a way forward.

As described by Bernardo [17], Demortier [2] and Pierini [21], to find the reference prior for a given problem one begins by considering the Kullback-Leibler divergence of the posterior $p(\theta|x)$ relative to a prior $\pi(\theta)$, obtained from the data $\vec{x} = (x_1, \ldots, x_n)$, which are assumed to consist of n independent and identically distributed values of a random variable x:

$$D_n[\pi, p] = \int p(\theta|\vec{x}) \ln \frac{p(\theta|\vec{x})}{\pi(\theta)} \, d\theta \ . \tag{8}$$

This is effectively a measure of the gain in information provided by the data. The reference prior is chosen so that the expectation value of this information gain is maximized for the limiting case of $n \to \infty$, where the expectation is computed with respect to the marginal distribution of the data,

$$p(\vec{x}) = \int L(\vec{x}|\theta)\pi(\theta) \, d\theta \ . \tag{9}$$

The techniques for finding reference priors are relatively straightforward for the case of a single parameter, where it turns out to be the same as the well-known Jeffreys prior. Finding a general algorithm suitable for the multiparameter case, however, proves to be problematic. In particular the multiparameter reference prior can in general depend on the ordering of the parameters. Further discussion and applications to particle physics problems can be found in Ref. [22].

The interpretation of the posterior probabilities derived from reference priors is the subject of some debate. One may, for example, derive the result, then disregard its Bayesian origins and simply exploit its frequentist properties. For example, one can use the posterior pdf to derive an interval for the parameter, which will than have a certain probability to cover to the true parameter value in the same sense as a frequentist confidence interval. One may also use a reference prior as part of a sensitivity analysis, i.e., a study of how the posterior probabilities change under variation of the prior. These questions will no doubt receive further study should reference priors find wider application in particle physics.

3.2 Bayesian model selection

In the particle physics community, the usual frequentist measure of significance for establishing a discovery has been based on the p-value of the background-only hypothesis. That is, one gives the probability, under assumption of no new signal, to see data as signal-like as what was actually observed or more so. This is of course not exactly what one wants, which would more naturally be the probabilities of the background-only model or various signal models. The fact that the p-value is often confused for the probability of the no-signal model only makes matters worse.

The natural substitute for a p-value in the Bayesian paradigm is the posterior probability that signal is present or absent given the data. But this requires the prior probability for the hypotheses, and this is where constructing a result that is of value to the broader scientific community becomes difficult. What, after all, is the prior probability for the existence of the Higgs boson? Or of supersymmetry, or some other perhaps very speculative extension to the Standard Model? These things are highly subjective, and mixing them into the reporting of an experimental result cannot help matters.

As discussed by Berger [18], however, one can summarize an experimental result by use of a Bayes factor, B_{01}, which quantifies the degree to which one of two hypotheses, H_0 or H_1, is preferred by the data. This requires no overall prior probabilities for H_0 or H_1, but priors must be given for all of the internal parameters of the two models.

For a pair of hypotheses H_0 and H_1 the Bayes factor is defined as the posterior odds divided by the prior odds,

$$B_{01} = \frac{P(H_0|x)}{P(H_1|x)} \frac{\pi_1}{\pi_0} = \frac{P(x|H_0)}{P(x|H_1)} \; . \tag{10}$$

Here x refers to the data and π_i ($i = 0, 1$) are the prior probabilities. That is, B_{01} is the same as the posterior odds if one were to assume equal prior probabilities, and it is thus an indicator of which model is preferred by the data. The second equality in (10) follows from Bayes' theorem, and therefore the Bayes factor is also equal to the ratio of likelihoods.

If a model contains any internal parameters, then to obtain the likelihood these must be characterized by a meaningful prior pdf and marginalized, i.e.,

$$P(x|H_i) = \int P(x|H_i, \theta_i) \pi_i(\theta_i) \, d\theta_i \; , \tag{11}$$

where θ_i are the internal parameters for model H_i ($i = 0, 1$) and $\pi_i(\theta_i)$ is the corresponding prior pdf. It is important to note that in this case the prior pdf cannot be improper, as this would only be defined up to an arbitrary constant and the Bayes factor would not be well defined. Furthermore, if an improper

prior is made proper by imposing a cut-off, then the Bayes factor will retain a dependence on this cut-off. Thus all internal parameters of the models must be characterized by meaningful, proper priors.

As an example, Berger examined the problem of a number of events N, assumed to be Poisson distributed with a mean $s + b$, where s and b are the contributions from signal and background processes, respectively. The Bayes factor for an observed value N is

$$B_{01}(N) = \frac{\text{Poisson}(N|0+b)}{\text{Poisson}(N|s+b)\pi(s)\,ds} = \frac{b^N e^{-b}}{\int_0^\infty (s+b)^N e^{-(s+b)}\pi(s)\,ds} \tag{12}$$

In both numerator and denominator, the probabilities must be integrated over all internal parameters of the two models. In this example, this is only relevant in the denominator, where one has the signal parameter s, characterized by a prior pdf $\pi(s)$.

The prior $\pi(s)$ could be chosen subjectively, but for most problems there would be no consensus for what to use. One could show the result for a variety of subjective choices, which would convey some feel for how important prior information is in the given problem. Alternatively one could use what is called the 'intrinsic prior', which for this problem is $\pi^I(s) = b(s+b)^{-2}$.

Finally, one would use the Bayes factor for discovery in a manner analogous to how a frequentist would use the p-value for the background-only hypothesis. That is, for a sufficiently small value of B_{01} or the p-value, one would reject the background-only model. The numerical values of cannot be directly compared, however, as illustrated by Berger in the example below.

Taking $N = 7$ and $b = 1.2$ gives a p-value of 0.00025, and as this number is very small one naturally thinks the probability of $s = 0$ must therefore also be small, and there can be a great temptation to identify the two numbers at least symbolically. But using the intrinsic prior in this case gives a Bayes factor $B_{01} = 0.0075$, i.e., a factor of 30 greater than the p-value.

That B_{01} is substantially greater than the p-value for this problem cannot be pinned on the prior used. One can place a lower bound on the Bayes factor by making the prior $\pi(s)$ a delta function centred on the ML estimator, $\hat{s} = \max(0, N - b)$. In this case one finds $B_{01} = 0.0014$, still quite a bit larger than the p-value of 0.00025. So the lesson of the exercise is that the smallness of the p-value cannot be mentally transferred so easily onto a similarly small probability for the hypothesis.

In fact in many problems it may be more useful to report the Bayes factor as a function of a parameter rather than integrating over it. For example, rather than B_{01}, one could show B_{0s} as a function of s. Of course in problems with a larger number of parameters this may become impractical.

An important impediment to the use of Bayes factors, however, is related to numerical challenges in computing the required marginal likelihoods represented by Eq. (11). One approach mentioned at the meeting [11, 23] involves a tool developed in the astrophysics community called nested sampling [24]. The key is to reparametrize the problem by defining

$$X(\lambda) = \int_{L(\theta)>\lambda} \pi(\theta)\,d\theta\ , \tag{13}$$

so that the desired integral can be written

$$\int L(\theta)\pi(\theta)\,d\theta = \int_0^1 X(\lambda)\,d\lambda\ . \tag{14}$$

An implementation of the nested-sampling algorithm is available in the `MultiNest` package [25].

Finally one should note that the Bayes factor is in many ways more intuitive than the p-value, as it addresses more directly the question of which model one believes to be true. For many years the particle physics community has used $p = 2.9 \times 10^{-7}$ for the p-value of the background-only model as a discovery threshold (a 5σ effect). But one's readiness to announce a discovery should surely depend

on factors such as the degree to which the data are better described by an alternative model, and this is directly found in the Bayes factor.

3.3 Approximate Bayesian Computation (ABC)

The methods for dealing with systematic uncertainties often rely on having a parametric model containing corresponding nuisance parameters. In the frequentist framework, tests based on the profile likelihood can be used to eliminate the nuisance parameters, or in the Bayesian approach one assigns appropriate priors and marginalizes.

In many cases, however, a parameter μ may appear in a Monte Carlo model for a given process, but one does not have a parametric function for the probability of the data x given μ. Demortier described an approach known as Approximate Bayesian Computation (ABC), in which one can approximate the posterior probability $p(\mu|x)$ without requiring direct access to $p(x|\mu)$ [2].

First the prior pdf for μ, $\pi(\mu)$, is sampled to obtain a value μ, and then this is used in the Monte Carlo model to generate a data set x^*. One then computes a distance measure that quantifies the separation between x^* and the data actually observed. If this distance is less than a given threshold the simulated μ is accepted, otherwise it is rejected. The distribution of accepted μ values is then used as an approximation for the posterior probability $p(\mu|x)$.

In principle this method could be used when combining measurements from two different experiments, although in that case the generation of Monte Carlo events would have to use the same parameter values and therefore some coordination would be necessary. ABC methods represent in any case an interesting approach that can address an important need in particle physics.

4 Applications and tools

The many contributions at PHYSTAT 2011 on statistical tools and applications clearly showed the community's increase in sophistication since the first PHYSTAT meeting more than ten years ago.

Cranmer [23] described the preparation for combination of the searches for the Higgs boson by ATLAS and CMS. This will exploit the full likelihood representing the joint outcomes of both experiments, with proper treatment of common nuisance parameters. The resulting model can be used in a variety of ways, such as in tests based on the profile likelihood or in a Bayesian analysis. The software for this task is being developed as part of the `RooStats` package [26, 27]. Studies on the combination of different decay channels shown by Zhukov [28] further illustrated the power of this software.

The Bayesian Analysis Toolkit (BAT) described by Pashapour [29] is a package designed for Bayesian computation, specifically, Markov Chain Monte Carlo integration for marginalization of posterior probabilities, and includes automated handling of tasks such as convergence diagnostics. As the user base for this package grows it will be important to include other aspects of Bayesian analyses, such as computation of marginal likelihoods (needed for Bayes factors) as well as support for various types of priors, particularly the reference priors mentioned in Sec. 3.1.

Prosper summarized lessons from the Tevatron [30]. The developments include multiparameter Bayesian analyses resulting in posterior densities for measured cross sections as well as searches for single top-quark production based on sophisticated multivariate classifiers. It will be interesting to see what role such classifiers will play in searches at the LHC, since their increased sensitivity can come with a loss of transparency. A 5σ signal from a Boosted Decision Tree may initially be met with some skepticism unless it is backed up by 4σ evidence from a cut-based analysis, and so the team that pursues both approaches may win the competition.

Among the most encouraging reports at the meeting were those on statistical practice by the LHC experiments, from which a rapid flow of publications is now emerging. The talks by Harel (CMS) [31], Casadei (ATLAS) [32] and Morata (LHCb) [33] show that different analysis groups are following

different routes, and a movement towards some level of uniformity may take some time to achieve.

S. Forte reported on quantifying uncertainties related to parton densities [34], demonstrating that theorists as well as experimentalists are involved in application of statistical methods. Forte presented an analysis of the NNPDF Collaboration in which neural networks are used to parametrize parametrize parton densities. The increased flexibility of the neural network relative to the parametric functions used by other groups is found to provide a more satisfactory assessment of parton uncertainties.

As in past meetings, the PHYSTAT workshop provided an important opportunity to learn from colleagues in other fields about their statistical practice. Röver [8] and Sardy [35] reported on searches for gravitational waves. The analyses employ regularization methods to suppress noise that involve a bias-variance trade-off that is similar to what particle physicists encounter when unfolding a distribution for effects of limited resolution.

Lahav [36] summarized astrophysical applications, where one finds a far greater use of Bayesian methods than currently seen in particle physics. HEP can learn from the astrophysics community about tools such as nested sampling for computing the marginal likelihoods needed in Bayesian model selection. An exoplanet search provided an outstanding example for particle physicists of how to present clearly a result obtained from a variety of Bayesian priors.

Finally, congratulations to the winners, and thanks to the organizers, of the Banff Challenge 2a, which was reported by Junk [37]. This addressed a number of tricky issues, including the look-elsewhere effect and poorly constrained nuisance parameters. We look forward to the next round.

5 Outlook and conclusions

It is clear that great progress has been made in the methods and software used in HEP since the first 'Confidence Limits' workshop at CERN in 2000. Use of sophisticated multivariate classifiers has become an industry, both frequentist and Bayesian approaches for dealing with systematic uncertainties have made important advances, and new software tools allow experiments to combine results in a way that fully exploits the available information and correctly accounts for common systematics. The 'look-elsewhere' effect has long been a serious problem and the contributions seen at this meeting go a long way towards solving it.

There are also many areas where progress is ongoing, such as in finding Bayesian reference priors for important HEP problems, and developing new, physically motivated ways to improve models so as to account for systematic effects. The issue of how best to report limits and intervals has still not found a fully satisfactory solution, but at least the tools are becoming available that allow a variety of approaches to be pursued easily.

The HEP community continues to cling to a 5σ discovery threshold, that is, a p-value of 2.9×10^{-7}. As pointed out by van Dyk, this can be viewed as sweeping problems such as the look-elsewhere effect and poorly understood systematics under the rug. This is one of many issues that one hopes will be revisited as real discoveries from the LHC begin to arrive.

Acknowledgements

I thank the speakers of PHYSTAT 2011 for their highly interesting contributions. I am grateful to the non-particle physicists and the statisticians for sharing their insights and expertise. And finally I thank the organisers, especially Louis Lyons, Albert de Roeck and Michelangelo Mangano, for arranging such a stimulating and productive meeting.

References

[1] D. Cox, these proceedings.

[2] L. Demortier, these proceedings.

[3] G. Cowan, K. Cranmer, E. Gross and O. Vitells, *Power-Constrained Limits*, arXiv:1105.3166 (2011).

[4] D. van Dyk, these proceedings.

[5] J. Conway, these proceedings.

[6] G. Cowan, these proceedings; see also G. Cowan, K. Cranmer, E. Gross and O. Vitells, Eur. Phys. J. C (2011) 71:1554; arXiv:1007.1727.

[7] O. Vitells, these proceedings.

[8] C. Röver, these proceedings.

[9] T. Junk, Nucl. Instrum. Methods Phys. Res., Sec. A **434**, 435 (1999); A.L. Read, J. Phys. G **28**, 2693 (2002).

[10] Vinay L. Kashyap et al., *On Computing Upper Limits to Source Intensities*, Astrophysical Journal, 719, 900-914 (2010); arXiv:1006.4334.

[11] R. Trotta, these proceedings.

[12] G. Ranucci, these proceedings.

[13] E. Gross and O. Vitells, Eur. Phys. J C70 (2010) 525-530; arXiv:1005.1891.

[14] E. Gross and O. Vitells, *Estimating the significance of a signal in a multi-dimensional search*, arXiv:1105.4355.

[15] R.B. Davis, Biometrika 74, (1987) 33-43.

[16] A. Caldwell, these proceedings.

[17] J. Bernardo, these proceedings.

[18] J. Berger, these proceedings.

[19] R.E. Kass and L. Wasserman, J. Amer. Statist. Assoc. 91 (1996) 1343.

[20] J.M. Bernardo, J. Roy. Statist. B 41 (1979) 113–147; J.M. Bernardo and J.O. Berger, J. Amer. Statist. Assoc. 84 (1989) 200–207. See also J.M. Bernardo, *Reference Analysis*, in *Handbook of Statistics*, 25 (D.K. Dey and C.R. Rao, eds.), 17–90, Elsevier, 2005, and references therein.

[21] M. Pierini, these proceedings.

[22] L. Demortier, S. Jain and H. Prosper, Phys. Rev. D Phys. Rev. D 82, 034002 (2010); arXiv:1002.1111.

[23] K. Cranmer, these proceedings.

[24] J. Skilling, Bayesian Analysis (2006) 1, Number 4, pp. 833–860.

[25] F. Feroz, M.P. Hobson and M. Bridges, Mon. Not. Roy. Astron. Soc., 398, 4, 1601-1614 (2009); arXiv:0809.3437.

[26] L. Moneta, K. Belasco, K. Cranmer *et al.*, "The RooStats Project," proceedings of ACAT, 2010, Jaipur, India [arXiv:1009.1003 [physics.data-an]]. [https://twiki.cern.ch/twiki/bin/view/RooStats/]

[27] G. Schott, these proceedings.

[28] V. Zhukov, these proceedings.

[29] S. Pashapour, these proceedings.

[30] H. Prosper, these proceedings.

[31] A. Harel, these proceedings.

[32] D. Casadei, these proceedings.

[33] J. Morata, these proceedings.

[34] S. Forte, these proceedings.

[35] S. Sardy, these proceedings.

[36] O. Lahav, these proceedings.

[37] T. Junk, these proceedings.

Unfolding

Unfolding: Introduction

Louis Lyons
Blackett Laboratory, Imperial College, London SW7 2BW, UK
and Particle Physics, Oxford OX1 3RH, UK

Abstract
As a non-expert on unfolding, I wanted to make a few 'obvious' non-controversial remarks about unfolding. It turns out, however, that even such innocuous comments can become the subject of heated debate.

1 The Problem

Given a one dimensional histogram for a particular variable x, obtained in a detector with known experimental resolution, can we estimate the histogram that we would have obtained had the detector not introduced any smearing? We assume that the smearing is specified by a matrix M, whose $(i,j)^{th}$ element is the probability that an event actually in bin j of the true histogram for x appears in bin i for the smeared data. It may be that the matrix M is known exactly. More likely is that its elements have statistical errors from its estimation via a simulation of detector effects; and/or it has systematic uncertainties because it was derived from an approximate model.

Of course we will also need to provide a covariance matrix for our de-smeared histogram. There are some obvious extensions of this example: the original distribution in x could be unbinned rather than in a histogram; the problem could involve more than one dimension, etc.

Thus the High Energy Physics unfolding problem is different from the more common statistics situation, where the issue is to remove the effects of smearing from an optical image in order to sharpen it, and to decide, for example, whether the photograph is of a dog, a cat or a person. In that case, estimates of uncertainties in pixel intensities, or in their correlations, are not of interest.

2 Why Unfold?

Because folding an assumed true distribution in x is simpler than unfolding an observed distribution in an attempt to obtain the true one, it is in general preferable to avoid unfolding. Then a comparison between a predicted theory and observed data is performed at the level of the smeared theory with the actual data, rather than between the pure theory and the unfolded data. This can even be performed for future theories, provided that the smearing matrix is published along with the data. The argument that a theorist, say in 2051, will find it difficult to smear his/her theory is very weak, since multiplying a vector by a matrix is computationally and conceptually no more difficult than comparing the original theory with the unsmeared data, which involves a non-trivial error matrix.

This then raises the question of when it might be necessary to unfold. A few cases are listed:

a) Comparing or combining experimental distributions from experiments with different smearing matrices M.

b) Tuning a Monte Carlo simulation, by fitting the parameters involved in the QCD theory to the data. Apparently this proceeds too slowly if the theory has to be smeared at each step of the iterative fitting.

c) Obtaining a plot for posterity that shows the estimate of the true distribution, rather than including the non-fundamental effects of experimental resolution. However, the unfolded distribution can contain strong bin-to-bin correlations, and physicists are accustomed to making eyeball judgements only about histograms whose bin contents are uncorrelated.

3 Matrix Method

Let d_i be the contents of the i^{th} bin of the data histogram and t_j that of the j^{th} bin of the distribution without smearing then,

$$d_i = \Sigma M_{ij} t_j. \tag{1}$$

Note that there can be fewer bins for the unfolded distribution than for the data.

The maximum likelihood solution for t can have large bin-to-bin oscillations, with estimated bin contents actually being negative. These effects become less serious for wide bins when the off-diagonal elements of the smearing matix are smaller. However, they then become more sensitive to the model assumed for deriving the smearing matrix; any such systematic effect should be taken into account in estimating the error matrix for the unsmeared distribution.

4 Bin-by-bin Correction Factors

This is an easy-to-use method, but unfortunately it has some serious drawbacks. Monte Carlo simulation is used for an assumed true distribution with a_i events in the true histogram, which after smearing becomes p_i events in the pseudo-data histogram. (Note that p_i in a given bin depends on all the a.) Then the correction factor C_i for the i^{th} bin is simply defined by

$$a_i = C_i \times p_i, \tag{2}$$

and depends on the assumed distribution. These factors are used to correct the observed data d_i to the estimated truth e_i by

$$e_i = C_i \times d_i. \tag{3}$$

An example of this approach for a simple 2-bin distribution is shown in Table 2, for the smearing matrix of Table 1. The numbers of true unsmeared events in bins 1 and 2 are 800 and 200, respectively. With the smearing matrix of Table 1, this results in 760 and 240 expected events in the 2 bins of the smeared distribution. We set the observed d_1 and d_2 equal to their expected values.

Each row of Table 2 corresponds to a different assumption about the a_i. ($a_1 + a_2$ is always taken as 1000, as in the assumed real data). The second row of numbers has them set at their true values, 800 and 200, respectively; then the estimates e_1 and e_2 are correct. For all other rows, the estimates are biassed, even for the case where both correction factors are unity; and also when the assumed numbers are set equal to the observed ones. Also their sum is not equal to the total number of observed events[1]. Furthermore, the errors on the bin contents of the unfolded histogram are in general taken as uncorrelated.

Calculating the uncertainties on the unfolded bin contents is problematic. For example, with $d_i = 100 \pm 10$ and $C_i = 0.1$, it is tempting to write $e_i = 10 \pm 1$. This uncertainty is wrong (it is even smaller than $\sqrt{d_i}$), as it is merely the uncertainty on the expected number, and does not include the statistical fluctuation on the observed number. Perhaps more importantly, only diagonal errors are obtained in this method.

Because of the sensitivity of the derived answers to the assumed unfolded distribution, which is needed for calculating the C_i, and because of the problems with obtaining the error matrix for the unfolded distribution, this method is *not* recommended[2].

5 Questions to be Resolved

5.1 Bin size

If the bin size of the unfolded distribution is too narrow, the smearing matrix has large off-diagonal elements. On the other hand, large bin width results in a loss of sensivity to high frequency components

[1] This contrasts with the matrix method.

[2] Though in some cases an iterative approach may work.

Bin	Truth 1	Truth 2
Observed 1	0.9	0.2
Observed 2	0.1	0.8

Table 1: The smearing matrix, whose elements M_{ij} are the probabilities that, as a result of smearing, an event in bin j of the true distribution appears in bin i of the distribution for the actual data.

a_1	a_2	p_1	p_2	C_1	C_2	e_1	e_2	Sum
1000	0	900	100	1.11	0	844	0	844
800	200	760	240	1.05	0.83	800	200	1000
760	240	732	268	1.04	0.90	789	215	1004
667	333	667	333	1.00	1 00	760	240	1000
500	500	550	450	0.91	1.11	691	267	958
200	800	340	660	0.59	1.21	447	291	738
0	1000	200	800	0	1.25	0	300	300

Table 2: Problems with correction factors. The correction factors C_i are calculated from assumed true numbers a_i and the corresponding smeared numbers p_i in each bin. The estimated numbers e_i are calculated as C_i times the actual observed numbers 760 and 240, and are to be compared with the true numbers 800 and 200 respectively. 'Sum' is $e_1 + e_2$, and in general is not equal to the observed sum of 1000.

of the true distribution; and to uncertainties in the elements of the smearing matrix, caused by their sensitivity to the distribution of the variable of interest across a bin. Some recommendation about the choice of bin width would be useful.

5.2 Regularisation

Because the maximum likelihood solution to the unfolding problem can result in an unfolded distribution with large bin-to-bin oscillations, some form of regularisation is generally employed to damp down these oscillations and to produce a smoother solution. This is usually achieved by adding a term to the likelihood which penalises large second derivatives; or by removing high frequency modes from the solution. Alternatively, orthogonal decomposition with suppression of small components can work for all distributions except those with sharp features.

A variety of regularisation methods is available, and again advice would be useful on the best form and the optimal regularisation strength to use in a given problem.

5.3 Size of errors

As already mentioned, calculating the uncertainties in the estimated unfolded distribution is often not trivial. (In cases of difficulty, a bootstrap method or some other method of varying the bin contents may be useful.) In particular, the question arises as to whether the uncertainty on an unfolded number of events e_i can be smaller than $\sqrt{e_i}$; that is, can the estimate in a situation with smearing in x have a smaller uncertainty than if x had been determined precisely? The answer may be *yes* because regularisation provides some form of local averaging, which can reduce the uncertainty on e_i.

5.4 Assessing a solution

It is not obvious how to assess which solution is best out of a series of solutions to an unfolding problem. Some of the problems are:

– A criterion of the largest p-value would favour a solution with large errors.
– Taking an unweighted sum of squared deviations ignores the fact that some bins are determined more precisely than others.
– Would we be satisfied if the minimum χ^2 were for a solution with wild anti-correlated oscillations?

Issues such as these need to be resolved before an Unfolding Challenge, along the lines of Banff Challenges 1 (for intervals) and 2 (for discovery), can meaningfully be set.

Bob Cousins has suggested a test that should be satisfied by any method that is used to unsmear two different data distributions, produced from two known 'true' distributions and specified detector smearing. Then he would expect

$$\Delta\chi^2_s = \Delta\chi^2_u \tag{4}$$

where $\Delta\chi^2$ is the difference in χ^2 between models 1 and 2; and the subscript 's' refers to the χ^2 being calculated by comparing smeared theory with the data, while 'u' refers to the comparison between the de-smeared data and the original theory. The reason for using 2 theoretical models rather than just one is that an unsmearing method could be tuned to work for a specific theory. Opinion is divided as to whether all reasonable theories would pass the test; or whether it is obvious that regularisation would invalidate it.

6 Today's Talks

We are very fortunate to have with us Victor Panaretos, a statistician from Lausanne, who will give the Statisticians' view of our problem, and will be present throughout the day, to encourage us to use statistically acceptable methods, and at very least to prevent us from straying too far afield. The other speakers before lunch (Blobel, Zech, Kartvelishvili and Bierwagen) will talk about the more common methods developed by High Energy Physicists. The afternoon speakers will deal with other HEP methods; a software framework for unfolding methods; and about the methods used in practice in 3 of the LHC collaborations.

7 Postscript

It was hoped that as a result of the meeting we could produce a consensus of points on which the major unfolding programme developers were in agreement. Perhaps as might have been predicted from the particularly lively discussions before, during and after the sessions, this turned out not to be possible. However, it is a goal worth striving for in the near future.

It is a pleasure to thank Volker Blobel for illuminating discussions.

A Statistician's View on Deconvolution and Unfolding

Victor M. Panaretos
École Polytechnique Fédérale de Lausanne, Switzerland

Abstract
We briefly review some of the basic features of unfolding problems from the point of view of the statistician. To illustrate these, we mostly concentrate on the particular instance of unfolding called deconvolution. We discuss the issue of ill-posedness, the bias-variance trade-off, and regularisation tuning, placing emphasis on the important class of kernel density estimators. We also briefly consider basic aspects of the more general unfolding problem and mention some of the points that where raised during the discussion session of the unfolding workshop.

1 Introduction

Unfolding and deconvolution can be seen to arise as variants of the statistical problem of *estimation*, when there is the additional complication of measurement error. In a classical setting, we are able to collect data in the form of realisations of random variables $(X_1, ..., X_n)$. These are often assumed independent and identically distributed according to a cumulative probability distribution F_X belonging to some known class of distributions $\mathcal{F}$. The problem of point estimation can then be formulated as follows:

1. Assume that $\mathcal{F}$ is known but F_X is unknown.
2. Observe a realisation $(x_1, ..., x_n)$ of $(X_1, ..., X_n)$, generated by F_X.
3. Determine which model from $\mathcal{F}$ generated the sample on the basis of $(x_1, ..., x_n)$ (i.e. estimate F_X).

Depending on the degree of specification of the collection $\mathcal{F}$, we may distinguish two broad classes of estimation problems. The first one, called *parametric estimation*, considers collections $\mathcal{F}$ that can be parametrised (put in one-to-one correspondence) by some subset of Euclidean space, $\mathcal{F} = \{F_\theta : \theta \in \Theta \subseteq \mathbb{R}^p\}$. The functional form of the elements of $\mathcal{F}$ is thus completely known, except for a finite-dimensional Euclidean parameter. The second broad class of estimation problems considers collections $\mathcal{F}$ that are only weakly specified, in the sense that they are taken to be subsets of a function space defined through some broad qualitative constraints. For example, $\mathcal{F}$ could be taken to be the collection of distributions possessing densities that are twice continuously differentiable.

An essential difference between the parametric and nonparametric frameworks concerns the effective dimensionality of the problem. In the parametric case, the estimation problem reduces to the determination of a finite dimensional parameter whose dimension remains fixed as the sample size grows. In the nonparametric framework, there is no finite-dimensional reduction of the problem: even if sample size might constrain us to approximate the truth by a function of effectively finite dimension, this dimension will typically increase along with sample size. Note here, that if the dimension of the parameter of a parametric model is allowed to grow with the sample size (i.e. the more the data, the richer the model we employ), then, as sample size increases, the distinction from a nonparametric model decreases.

Occasionally, the problem –be it parametric, or nonparametric– is further perturbed by measurement limitations. That is, instead of the collection $\boldsymbol{X} = (X_1, ..., X_n)$, we can only observe a proxy $\boldsymbol{Y} = (Y_1, ..., Y_n)$, that results from some random perturbation of $\boldsymbol{X}$, governed by a *measurement error mechanism*:

$$Y_i = g(X_i, \varepsilon_i), \qquad i = 1, ..., n.$$

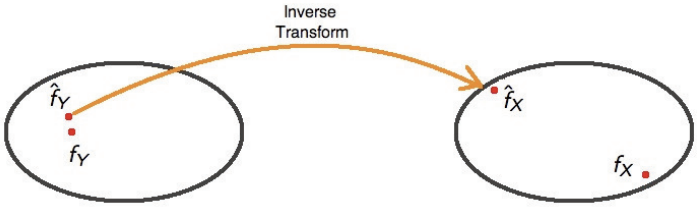


Fig. 1: Schematic illustration of ill-posedness: small errors in the estimation of f_Y may be translated into large errors in estimation of f_X

Here, one typically assumes that the random errors ε_i are mutually independent, identically distributed and independent of the collection $\boldsymbol{X}$, the function g is smooth and that the functions $\{g_t(\cdot) = g(t, \cdot); t \in \mathbb{R}\}$, are invertible with differentiable inverses (i.e. if we knew the input and the response, then we should be able to uniquely and stably determine the error). The error inputs ε_i are unobservable, but their distribution F_ε will be assumed to be completely known. All random variables are assumed real.

It follows that the data we observe are from a distribution F_Y and not the distribution of interest F_X. Assuming that all random variables involved possess density functions denoted by f_X and f_Y, respectively, we may write

$$f_Y(y) = \int f_{Y|X}(y|x) f_X(x) dx = \int h(x, y) f_X(x) dx,$$

where h is connected to the density f_ε of ε and the measurement function g through the change of density formula:

$$h(x, y) = f_{Y|X}(y|x) = \left| J_{g_x^{-1}}(y) \right| \times f_\varepsilon \left(g_x^{-1}(y) \right).$$

Here, J stands for the Jacobian of the transformation. The unfolding problem (or one version of what statisticians call a measurement error model) then consists in estimating the density f_X when one observes realisations from the density f_Y.

In principle, the measurement error problem does not pose real conceptual difficulties in the parametric case because from a qualitative point of view the problem remains the same: the density f_Y will still depend on the original parameter θ, and therefore estimation of θ can be carried out directly in the Y-space without additional complications, at least provided that θ remains identifiable, or the likelihood is not significantly "flattened" (lack, or almost lack, of identifiability can, however, be an issue and would lead to serious complications; one way to address these is by identifiability constraints, which essentially amount to *parametric regularisation*, though we will not pursue this further here).

In the nonparametric case, however, f_X is completely unknown, and hence one cannot escape the measurement error problem and work solely on Y-space. Rather, one will need to first estimate f_Y by some $\hat{f}_Y$, and then attempt to invert the integral transform connecting f_Y with f_X, using $\hat{f}_Y$ as a proxy for f_Y. This naive approach, however, can lead to serious errors. For well-behaved h, the integral transformation involved will have a discontinuous (unbounded) inverse transform. Consequently, an element of ill-posedness enters the picture, as small errors in the estimation of f_Y may be translated into large errors in estimation of f_X through the inversion process.

If the measurement errors are not independent, then the observed variables $(Y_1, ..., Y_n)$ will no longer constitute an independent random sample, but will instead form a stationary process. Consequently, the integral expression $\int h(x, y) f_X(x) dx$ will still hold marginally for the density of each Y_i, but the joint density will no longer be the product density. Nevertheless, even in such cases, one can attempt to proceed using the same estimators as in the independent case, provided that the dependence among the errors is weak (where the notion of 'weak dependence' can be formalised through appropriate

mixing conditions). The stronger the dependence structure of the errors, the less reliable such estimators will become.

2 Nonparametric Deconvolution

An interesting special case of the unfolding problem is obtained when attention is restricted to the case $g(X, \varepsilon) = X + \varepsilon$. The integral equation relating the measured and true density reduces to a convolution equation

$$f_Y(y) = \int_{\mathbb{R}} f_\varepsilon(y - x) f_X(x) dx \Leftrightarrow f_Y = f_\varepsilon * f_X.$$

We will concentrate on this special problem in the nonparametric case as, on the one hand, it contains the germs of generality, while on the other hand, it is very well understood, having been extensively studied in the statistics literature. Our presentation in this section will borrow heavily from Meister [9], who provides an elegant overview of statistical deconvolution, and where precise versions of the statements given here, along with proofs may be found.

2.1 Inversion and Ill-Posedness

The fact that the integral equation relating the two densities is a convolution immediately suggests an estimation technique based on direct inversion of the convolution operator:

1. Let ϕ_X, ϕ_ε, ϕ_Y be the characteristic functions (Fourier transforms) corresponding to the densities f_X, f_ε and f_Y, respectively.
2. Then $f_Y = f_\varepsilon * f_X \implies \phi_Y = \phi_X \phi_\varepsilon$.
3. So if $\hat{\phi}_Y$ is an estimate of the characteristic function of Y, we could estimate ϕ_X by $\hat{\phi}_Y / \phi_\varepsilon$, to obtain:

$$\hat{f}_X(t) = \frac{1}{2\pi} \int_{\mathbb{R}} e^{-itu} \frac{\hat{\phi}_Y(u)}{\phi_\varepsilon(u)} du.$$

This, of course, raises the question of how can one estimate the characteristic function ϕ_Y. Provided that approximation error is measured in square integrated distance, the answer is provided by the Plancherel identity: good estimates of f_Y give good estimates of ϕ_Y (and vice versa) by Fourier transforming,

$$\int [f_1(y) - f_2(y)]^2 dy = \int \left| \widetilde{f_1}(u) - \widetilde{f_2}(u) \right|^2 du.$$

Here $\widetilde{g}$ denotes the Fourier transform of g. So if $\hat{f}_Y$ is a good estimator of f_Y, then $\widetilde{\hat{f}}_Y$ is just as good an estimator of ϕ_Y, so that we may estimate f_Y, and then apply the Fourier transform to get an estimator of ϕ_Y. However, the continuity (boundedness) of the operation $f_X \mapsto f_X * f_\varepsilon$ (and corresponding discontinuity (unboundedness) of the inverse operation) will reveal that it is not enough to try to estimate f_Y (or, equivalently ϕ_Y) accurately.

Observe that, by the Plancherel identity, we have

$$\|f_X - \hat{f}_X\|^2 = \int |f_X(x) - \hat{f}_X(x)|^2 dx = \int \left| \phi_X(u) - \frac{\hat{\phi}_Y(u)}{\phi_\varepsilon(u)} \right|^2 du = \int \left| \frac{\phi_Y(u) - \hat{\phi}_Y(u)}{\phi_\varepsilon(u)} \right|^2 du.$$

where the last equality follows from $f_Y = f_X * f_\varepsilon$. Now, typically, $\phi_\varepsilon(u)$ decays "fast" as $u \to \pm\infty$. Therefore, a small discrepancy between $\hat{\phi}_Y$ and ϕ_Y for a large frequency u, can be blown up to an arbitrarily large discrepancy between f_X and $\hat{f}_X$. Such discrepancies are guaranteed to occur by the inherent statistical variability, but also because of sample limitations: it is intuitively clear that estimating the highest frequency characteristics of f_Y accurately based on a finite sample is essentially not possible.

2.2 Regularisation

We have seen that at the essence of ill-posedness lies the fact that $\frac{\hat{\phi}_Y(u)}{\phi_\varepsilon(u)}$ is potentially a bad estimator of $\phi_X(u)$ for u outside some bounded interval. Nevertheless, if f_X is going to be square integrable, we expect $|\varphi_X(u)|$ to become negligible as $u \to \pm\infty$. That is, we might a priori know based on qualitative properties of f_X that ϕ_X is practically zero outside some domain $\left[-\frac{1}{b}, \frac{1}{b}\right]$. Given this information, we can set our estimator to be $\frac{\hat{\phi}_Y(u)}{\phi_\varepsilon(u)}$ when $u \in \left[-\frac{1}{b}, \frac{1}{b}\right]$, and set it to zero outside that domain to avoid the blow-up. That is, we employ the *regularised* estimator

$$\frac{\mathbf{1}\{-1/b \leq u \leq 1/b\}\hat{\phi}_Y(u)}{\phi_\varepsilon(u)} = \frac{\mathbf{1}\{-1 \leq bu \leq 1\}\hat{\phi}_Y(u)}{\phi_\varepsilon(u)} = \frac{\widetilde{K}(bu)\hat{\phi}_Y(u)}{\phi_\varepsilon(u)},$$

where we used the notation $\widetilde{K}(bu) = \mathbf{1}\{-1 \leq bu \leq 1\}$ to imply that we understand $\mathbf{1}\{-1 \leq bu \leq 1\}$ as the Fourier transform of some *kernel function* K. More generally, we could use a weight function $\widetilde{K}(bu) := \widetilde{K}_b(u)$, where $\widetilde{K}(u)$ is supported on $[-1, 1]$ and bounded. This corresponds to estimating ϕ_Y by $\widetilde{K}(bu)\hat{\phi}_Y(u)$ which leaves the naive estimate unaffected for small frequencies, tames the variation for higher frequencies, and kills the variation completely beyond a certain frequency threshold.

2.3 The Naive and Kernel Estimators

To illustrate these ideas, we consider the so-called 'plug-in' estimator of ϕ_Y, defined as

$$\hat{\phi}_Y^{nve}(u) = \frac{1}{n}\sum_{j=1}^{n} e^{iuY_j}.$$

We labeled this as the naive estimator ('*nve*'), since it merely uses the empirical version of the characteristic function. It is, nevertheless, a reasonable estimator for every fixed u by the strong law of large numbers: it is unbiased, $\mathbb{E}[\hat{\phi}_Y^{nve}(u)] = \phi_Y(u)$, and for any u, $\mathbb{E}|\hat{\phi}_Y^{nve}(u) - \phi_Y(u)|^2 = O(n^{-1})$. However, these are properties that hold locally, i.e. for a fixed u. When employing $\hat{\phi}_Y^{nve}$ as part of a deconvolution estimator to estimate f_X, we see that it is "totally affected" by ill-posedness: the induced estimator $\hat{f}_X^{nve}(x) = \text{“}(2\pi)^{-1}\sum_{j=1}^{n}\int \frac{e^{iu(Y_j - x)}}{n\phi_\varepsilon(u)}du\text{”}$ is not even well-defined (hence the quotation marks – the quantity inside the integral is not integrable). Nevertheless, we may employ the regularisation strategy presented in the previous section, in order to obtain a regularised version of the naive estimator that is not only well-defined, but also has controllable variation:

$$\hat{f}_X^{kernel}(x) = \frac{1}{2\pi}\int e^{-iux}\frac{\widetilde{K}(bu)\sum_{j=1}^{n}\frac{1}{n}e^{iuY_j}}{\phi_\varepsilon(u)}du.$$

We call this the kernel estimator, because we have built it by dampening the high frequency components of the naive estimator using the Fourier transform of some kernel K.

2.4 Error Properties of the Kernel Estimator

Since the object being estimated is a function, it is not immediately clear what criterion one might employ in order to measure the accuracy of the deconvolution estimator. Natural building blocks for error measures are divergences on function space, which can then yield error measures by means of averaging (averaging meaning taking the expectation with respect to the sample observations $Y_1, ..., Y_n$). The choice of divergence reflects which aspects of f_X we wish to emphasize the most. For example, one could define an error measure based on the Cramér-Von Mises divergence as $\mathbb{E}[\int(\hat{F}_X(x) - F_X(x))^2 f_X(x)dx]$, placing greater emphasis on regions of high density. If interest lies primarily on the tails of the distribution, then one could employ an Anderson-Darling divergence, $\mathbb{E}[\int(\hat{F}_X(x) - F_X(x))^2[F_X(x)(1 -$

$F_X(x))]^{-1} f_X(x)dx]$. Perhaps the most widely studied error measure is the mean integrated squared error measure, $\mathbb{E}\int(\hat{f}_X(x) - f_X(x))^2 dx = \mathbb{E}\|\hat{f}_X - f_X\|^2$, which places equal emphasis on different parts of the domain of f_X. In what follows, we will concentrate on this particular error measure. This is also partly a matter of convenience, since the convolution operator is naturally linked with Fourier analysis on L^2. Assuming that f_X is square-integrable and that $\phi_\varepsilon(u) \neq 0$ everywhere, then, if $\widetilde{K}(u)$ is bounded and supported on a bounded interval, it can be shown that

$$\mathbb{E}\|f_x - \hat{f}_X^{kernel}\|^2 = \frac{1}{2\pi n}\int |\widetilde{K}(bu)|^2 \left[\frac{1}{|\phi_\varepsilon(u)|^2} - |\phi_X(u)|^2\right] du + \frac{1}{2\pi}\int |\widetilde{K}(bu) - 1|^2 |\phi_X(u)|^2 du.$$

This error expression provides insight into the nature of the estimation error in deconvolution. In the statistical terminology, the first term represents the variance component of the error, whereas the second term represents the bias component. The variance term represents the component of the error that is due to statistical variation as well as the instability of the inversion process. The bias component describes the systematic error due to regularisation. The expression reveals the existence of a fundamental trade-off between the two terms, governed by the *regularisation parameter* b:

1. In the first term (variance term), $\widetilde{K}$ controls blow-up caused by rapid decay of ϕ_ε. This component is decreasing in b (i.e. decreases as the length of the interval $[-b^{-1}, b^{-1}]$ decreases).
2. In the second term (bias term), $\widetilde{K}$ controls "how far we are on average" from estimating ϕ_X. This component is increasing in b (i.e. increases as the length of the interval $[-b^{-1}, b^{-1}]$ decreases).

Therefore, the choice of b must be made judiciously, in order to balance these two effects, and minimise the overall mean squared error.

One may, however, pose the question of whether it is possible to do better than the kernel estimator in terms of overall error. Said differently, is the kernel estimator "optimal" (or at least fairly reasonable), or should we rather concentrate on something different? As is typically the case in statistics, this question cannot be answered exactly, i.e. for fixed sample size, nor in complete generality (very weak specification of the properties of the densities involved). A partial answer can be given in an asymptotic regime, as the sample size is taken to increase to infinity, $n \to \infty$, and under a stronger specification of the function class of the densities involved and the properties of the error density. Roughly speaking, we could require that the class $\mathcal{F}$ contains relatively smooth functions (e.g. densities possessing a β-th derivative with uniformly bounded L^2 norm) and that the tail decay of the error characteristic function is of polynomial order, say with an exponent $\alpha < 0$ (such error densities are called smooth, to be contrasted with supersmooth error densities, where the decay of the characteristic function is exponential). Notice that the latter assumption is related to the typical magnitude of the errors: the rate of decay of the error characteristic function is connected with the typical magnitude of the error (slow tail decay of the characteristic function means that the error density is concentrated around zero).

Under these assumptions, and if the kernel K satisfies certain regularity conditions, choosing $b \sim n^{-\frac{1}{2\beta+2\alpha+1}}$ as $n \to \infty$, we can obtain the following asymptotic upper bound for the error of the kernel estimator:

$$\sup_{f\in\mathcal{F}} \|\hat{f}_X^{kernel} - f\|^2 = O(n^{-\frac{2\beta}{2\beta+2\alpha+1}}).$$

On the other hand, if we additionally assume that $|\frac{d}{du}\phi_\varepsilon(u)| \leq \text{const} \times |t|^{-\alpha}$, and $\beta > 1/2$, we have

$$\sup_{f\in\mathcal{F}} \|\hat{f}_X - f\|^2 \geq \text{const} \times n^{-\frac{2\beta}{2\beta+2\alpha+1}},$$

for any estimator $\hat{f}_X$ of f_X, as $n \to \infty$ (the precise regularity conditions can be found in Meister [9, Sec. 2.4]). That is, asymptotically, we will never be able to perform order of magnitude better than the kernel estimator over the whole function class, since the upper bound for the kernel estimator coincides up to

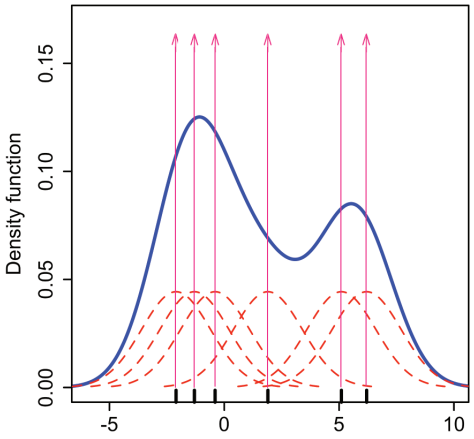


Fig. 2: Schematic representation of the workings of a kernel estimator in the time domain.

a constant with the lower bound for *any* other estimator. Of course these results leave much room for discussion, as they are uninformative when it comes to the exact finite sample behaviour of the kernel estimator relative to other estimators. In addition, these results show rate optimality, but there is an undetermined constant involved. Finally note that they consider the worst case error over the function class. They should therefore be interpreted with care as qualitative statements. Optimality results of a similar flavour are available when the error density is super-smooth (e.g. Gaussian), but with slower convergence rates.

2.5 Time Domain Interpretation of the Kernel Estimator

The kernel estimator of ϕ_Y amounts to replacing the naive estimator, $\frac{1}{n}\sum_{j=1}^{n} e^{iuY_j}$, by the regularised estimator $\widetilde{K}(bu) \cdot \frac{1}{n}\sum_{j=1}^{n} e^{iuY_j}$. Applying the inverse Fourier transform, we see that this amounts to estimating f_Y by

$$\left[\frac{1}{b}K\left(\frac{x}{b}\right)\right] * \left[\frac{1}{n}\sum_{j=1}^{n}\delta(Y_j - x)\right] = \sum_{j=1}^{n}\frac{1}{nb}K\left(\frac{Y_j - x}{b}\right),$$

where $\widetilde{K}(u) = \int K(t)e^{itu}dt$ and δ is Dirac's delta. Intuitively, instead of using the "empirical density" as the estimator of f_Y, the kernel estimator uses a smoothed version. In a sense, the empirical density 'fits best' to the observed data but is far too rough (it is not even a function in a proper sense). This roughness of the density estimate would translate into a slow tail decay of its Fourier transform, leading to the problems observed in Section (2.3). This is the essence of the bias-variance trade-off in the time domain: we need to attempt to fit the data well but at the same time maintain a certain level of smoothness to avoid ill-posedness issues.

2.6 Tuning The Regularisation Parameter

The bias-variance tradeoff phenomenon requires that we tune the regularisation parameter in order to obtain the optimal amount of regularisation. For supersmooth error densities (e.g. Gaussian), the (asymptotically) optimal regularisation parameter is independent of the unknown function f_X, and depends only on the error density – in fact, provided that the error density is known, we can determine the exact value of the regularisation parameter (there is no unknown constant involved).

However, in the case of simply smooth error density [assumption (A3)] we saw that the bandwidth depends on the smoothness properties of the unknown f_X. In fact, even if we knew these exactly, we

would still only know how to choose a bandwidth up to a constant. An alternative is to choose b using the data as a guide. For example, we could attempt to choose b to minimise $\|\hat{f}_{b,X}^{kernel} - f_X\|^2$. This, of course, depends on the unknown density f_X, but can be estimated using the data. The squared norm admits the decomposition

$$\|\hat{f}_{b,X}^{kernel} - f_X\|^2 = \|\hat{f}_{b,X}^{kernel}\|^2 - 2\Re\underbrace{\left[\langle \hat{f}_{b,X}^{kernel}, f_X\rangle\right]}_{\mathbb{E}\left[\overline{\hat{f}_{b,X}^{kernel}}(X)\right]} + \|f_X\|^2,$$

where $\Re$ denotes the real part of a complex number. We notice that the first term depends only on the data, the second term is estimable from the data and the third term is independent of b. One could employ leave-one-out cross validation in the Fourier domain to estimate $\mathbb{E}\left[\overline{\hat{f}_{b,X}^{kernel}}(X)\right]$ and select the value of b that minimises the overall expression (see Meister [9, Sec. 2.5.1] for details).

The problem with cross-validation is that b becomes a random variable, dependent on $\hat{f}_{b,X}^{kernel}$. In essence, we use the data twice and so our error and optimality results are not guaranteed to hold true. Nevertheless, at least asymptotically, cross validation can be seen to be adaptive – meaning that with increasing sample size, it will eventually provide the optimal error rates, thus *adapting* to the potentially unknown smoothness class of the unknown density. Under some additional smoothness assumptions on f_X (which we omit for brevity), as well as the α-polynomial tail decay assumption on ϕ_ε, if we use the sinc kernel to construct $\hat{f}_{b,X}^{kernel}$ and optimise the empirical integrated squared error on a fine enough grid $G(n,\alpha)$ (depending on n and α), we obtain:

$$\limsup_{n\to\infty}\left\{\frac{\inf\limits_{b\in G(n,\alpha)}\widehat{\mathrm{ISE}}(b)}{\inf\limits_{b>0}\mathbb{E}\|\hat{f}_{b,X}^{kernel} - f_X\|^2}\right\} \leq 1, \qquad \text{with probability 1.}$$

Here, $\widehat{\mathrm{ISE}}(b)$ denotes the cross-validated mean integrated squared error corresponding to a regularisation parameter b. See Meister [9, Thm. 2.17] for the precise statement.

2.7 A Pointwise Central Limit Theorem

One may be interested to obtain a confidence interval for the value of the unknown density estimate at a point x. For this reason, one would require an approximate distribution for $\hat{f}_X$ at the point x. If $f_Y = f_X * f_\varepsilon$ is uniformly bounded, then, under regularity conditions on the kernel and polynomial decay of the error characteristic function (smooth error density),

$$\frac{\hat{f}_X^{kernel}(x) - \mathbb{E}[\hat{f}_X^{kernel}(x)]}{\sqrt{\mathrm{Var}[\hat{f}_X^{kernel}(x)]}} \xrightarrow{d} N(0,1),$$

for all $x \in \mathbb{R}$, provided that b is selected so that $bn \overset{n\to\infty}{\longrightarrow} \infty$. This central limit theorem can be used in conjunction with the available bounds on the squared bias in order to construct asymptotic confidence intervals.

It should be noted that there is no corresponding central limit theorem when the error density is supersmooth (exponential decay of the characteristic function), e.g. in the case of Gaussian errors.

2.8 Induced Estimators of the Distribution Function

It might be the case that we are not interested in estimating the density f_X per se, but rather that we are interested in estimating the probability of a certain interval $[c,d]$,

$$F_{c,d} = F_X(d) - F_X(c) = \mathbb{P}[c \leq X \leq d] = \int_c^d f_X(x)dx.$$

Here F_X denotes the distribution function of the random variable X. From the mathematical point of view, this is an 'easier' problem, as we are required to estimate a smooth functional of the density function. Since we have already done the hard work of estimating f_X, it is natural to use a plug-in estimator for this purpose, i.e. use the estimator

$$\hat{F}_{c,d} = \int_a^b \hat{f}_X^{kernel}(x)dx = \int \mathbf{1}\{c \le x \le d\}\hat{f}_X^{kernel}(x)dx.$$

Plancherel's identity now shows that

$$\hat{F}_{c,d} = \frac{1}{n\pi}\sum_{j=1}^{n}\int e^{iu(c+d)/2}\sin[(u(d-c)/2]\frac{\widetilde{K}(bu)e^{iuY_j}}{u\phi_\varepsilon(u)}du.$$

Under some smoothness assumptions on the unknown density function, we can also obtain error bounds for this estimator of the distribution function, both for a fixed unknown density, as well as uniformly in a prescribed function class. In particular, we suppose that the class $\mathcal{F}$ contains uniformly bounded and smooth densities (possessing a globally Hölder-continuous $\lfloor\beta\rfloor$-th derivative, $\lfloor\beta\rfloor$ denoting the integer part of β). Then, provided the kernel satisfies certain regularity conditions, the plug-in estimator satisfies

$$\mathbb{E}|\hat{F}_{c,d} - F_{c,d}|^2 \le \|f_Y\|_\infty \frac{2}{\pi n}\int \frac{\left|\sin[u(d-c)/2]\widetilde{K}(bu)\right|^2}{|u|^2|\phi_\varepsilon(u)|^2}du + 4C^2b^2\left(\int |K(t)||t|dt\right)^2.$$

Furthermore, if we additionally assume an α-polynomial decay of ϕ_ε, we may also obtain the following uniform error bounds for $\sup_{f\in\mathcal{F}}\mathbb{E}|\hat{F}_{c,d} - F_{c,d}|^2$:

1. For $0 < \alpha < 1/2$, if we simply put $\widetilde{K} = 1$, the bound is $O(1/n)$.
2. For $\alpha = 1/2$ and $b \sim n^{-1}$, the bound is $O\left(\frac{\log n}{n}\right)$.
3. For $\alpha > 1/2$ and $b \sim n^{-1/(2\beta+2\alpha+1)}$, the bound is $O\left(n^{-(2\beta+2)/(2\beta+2\alpha+1)}\right)$.

Notice that for $0 < \alpha < 1/2$ (slow tail decay of the error characteristic function), the rate is essentially the parametric estimation rate. The precise regularity conditions may be found in Meister [9, Sec. 2.7].

3 Nonparametric Unfolding

We now turn to making some broad remarks on a more general version of unfolding. This arises when considering a more general binary operation $g(X, \varepsilon)$. The integral equation now becomes

$$f_Y(y) = \int h(x,y)f_X(x)dx \implies f_Y = \mathcal{L}f_X,$$

where $\mathcal{L} : f \mapsto \int h(x,y)f(y)dy$ is a more general integral operator. For simplicity, we assume that the densities involved are square integrable and supported on $[a,b]$, and that $h(x,y) = \sum_{k=1}^{\infty}\lambda_k\varphi_k(x)\varphi_k(y)$ for an orthonormal basis $\{\varphi_k\}$ of $L^2[a,b]$ and a square summable sequence of non-zero real coefficients $\{\lambda_k\}$ (so that h is square-integrable and symmetric). Then, the eigenfunctions of the operator $\mathcal{L}$ are precisely the φ_k, with corresponding eigenvalues λ_k, $\mathcal{L}\varphi_k = \lambda_k\varphi_k$. We have $f(x) = \sum_{k=1}^{\infty}a_k\phi_k$ for all square integrable f, with $a_k = \langle f, \varphi_k\rangle = \int_a^b f(t)\varphi_k(t)dt$. In this setting, we may write

$$f_Y = \mathcal{L}f_X = \mathcal{L}\left[\sum_{k=1}^{\infty}\langle f_X, \varphi_k\rangle\varphi_k\right] = \sum_{k=1}^{\infty}\langle f_X, \varphi_k\rangle\mathcal{L}[\varphi_k] = \sum_{k=1}^{\infty}\lambda_k\langle f_X, \varphi_k\rangle\varphi_k.$$

On the other hand, we must also have $f_Y = \sum_{k=1}^{\infty}\langle f_Y, \varphi_k\rangle\varphi_k$ so that $\langle f_Y, \varphi_k\rangle = \lambda\langle f_X, \varphi_k\rangle\varphi_k$, since the Fourier representation is unique. Going backwards, we can invert the transform by taking $f_X =$

$\sum_{k=1}^{\infty} \frac{\langle f_Y, \varphi_k \rangle}{\lambda_k} \varphi_k$. This short analysis suggests an estimation strategy similar to that employed in the case of deconvolution: (1) First estimate f_Y by some reasonable estimator $\hat{f}_Y$, (2) Then apply the inverse transform to $\hat{f}_Y$,

$$\hat{f}_X = \sum_{k=1}^{\infty} \frac{\langle \hat{f}_Y, \varphi_k \rangle}{\lambda_k} \varphi_k.$$

We now observe how ill-posedness manifests itself in this context: since the sequence of eigenvalues is assumed square integrable, it must be that $\lambda_k \downarrow 0$. Note that since $\{\varphi_k\}$ is a basis,

$$\|f_X - \hat{f}_X\|^2 = \sum_{k=1}^{\infty} \left(\langle f_X, \varphi_k \rangle - \frac{\langle \hat{f}_Y, \varphi_k \rangle}{\lambda_k} \right)^2 = \sum_{k=1}^{\infty} \frac{\left(\langle f_Y, \varphi_k \rangle - \langle \hat{f}_Y, \varphi_k \rangle \right)^2}{\lambda_k^2}.$$

The situation is thus similar to the deconvolution setting: small estimation errors in the estimation of f_Y (e.g. in the estimation of $\langle f_Y, \varphi_k \rangle$ for large k) magnify to large estimation errors of f_X due to the blow-up of the inverse eigenvalues. As in deconvolution, we will need to 'tame' this blow-up by means of regularisation. To this aim, we can argue that since $\|f_X\|^2 = \sum_{k=1}^{\infty} \langle f_X, \varphi_k \rangle^2 < \infty$, it must be that $\langle f_X, \varphi_k \rangle \to 0$. Therefore, we may choose a truncation level B (regularisation parameter), and enforce:

$$\langle \hat{f}_X, \varphi_k \rangle = \begin{cases} \frac{\langle \hat{f}_Y, \varphi_k \rangle}{\lambda_k} & \text{if } k \leq B, \\ 0 & \text{otherwise.} \end{cases}$$

Equivalently, we estimate f_Y by $\sum_{k=1}^{B} \langle \hat{f}_Y, \varphi_k \rangle \varphi_k$, obtaining the estimator of the original density:

$$\hat{f}_X = \sum_{k=1}^{B} \frac{\langle \hat{f}_Y, \varphi_k \rangle}{\lambda_k} \varphi_k.$$

To illustrate the approach, we apply this *spectral truncation regularisation* to a naive 'plug-in' estimator of f_Y. This is obtained by noting that $\langle f_Y, \varphi_k \rangle = \int_a^b \varphi_k(u) f_Y(u) du = \mathbb{E}[\varphi_k(Y)]$. Consequently $f_Y(u) = \sum_{k=1}^{\infty} \mathbb{E}[\varphi_k(Y)] \varphi_k(u)$. Again, the law of large numbers seems to suggest to estimate the expectations $\mathbb{E}[\varphi_k(Y)]$ by their empirical versions (their optimal unbiased estimators) to obtain the naive estimator

$$\hat{f}_Y^{nve}(u) = \text{“}\sum_{k=1}^{\infty} \left[\frac{1}{n} \sum_{j=1}^{n} \varphi_k(Y_j) \right] \varphi_k(u)\text{”}.$$

Just as with deconvolution, though, this is not even a well-defined estimator. To see why, notice that the series would appear to "converge" to the "Fourier series expansion" of the empirical density $\frac{1}{n} \sum_{j=1}^{n} \delta(u - Y_j)$ with respect to the $\{\varphi_k\}$ basis, since $\frac{1}{n} \sum_{j=1}^{n} \varphi_k(Y_j) = \int \left[\frac{1}{n} \sum_{j=1}^{n} \delta(u - Y_j) \right] \varphi_k(u) du$. But this empirical density is far from being square integrable, and such an expansion is undefined. Nevertheless, any finite sum formed by the first B summands of the series is well defined, that is, the spectrally truncated naive estimator $\hat{f}_Y^{trunc}$, is well-defined, yielding the truncated series estimator of f_X,

$$\hat{f}_X^{trunc}(x) = \sum_{k=1}^{B} \left[\frac{1}{n\lambda_k} \sum_{j=1}^{n} \varphi_k(Y_j) \right] \varphi_k(x).$$

In terms of error properties, one can see that the truncated estimator satisfies a similar bias-variance tradeoff as before,

$$\mathbb{E}\|f_X - \hat{f}_X^{trunc}\|^2 = \sum_{k=1}^{B} \frac{\text{Var}[\varphi_k(Y)]}{n\lambda_k^2} + \sum_{k=B}^{\infty} \langle f_X, \varphi_k \rangle^2.$$

In the first term (variance term), B controls the blow-up caused by the decay of eigenvalues as compared to the behaviour of $\text{Var}[\varphi_k(Y)]/n$. In the second term (bias term), B controls our systematic deviation from f_X, expressed in terms of what part of f_X we are missing completely due to the truncation.

The spectrally truncated naive estimator also allows one to appreciate the potential effects that dependence among the measurement error inputs ε_i may bring about: if the Y_i are stationary but not independent, then we can still rely on $n^{-1}\sum_{i=1}^n \varphi_k(Y_i)$ as an estimator of $\mathbb{E}[\varphi_k(Y)]$ by the ergodic theorem, but the quality of the estimator for finite n may suffer, depending on the strength of the dependence (the bias term will remain the same, but the terms in the series of the variance component will now be $(n\lambda_k)^{-2}\sum_{i=1}^n\sum_{j=1}^n \text{Cov}(\varphi_k(Y_i),\varphi_k(Y_j))$).

We now turn to show that, in fact, the spectrally truncated estimator admits a kernel estimator interpretation. Let $K_B(x,y)=\sum_{k=1}^B \varphi_k(x)\varphi_k(y)$, and observe that

$$\hat{f}_Y^{trunc}(x)=\sum_{k=1}^B\langle \hat{f}_Y^{nve},\varphi_k\rangle\varphi_k(x)=\sum_{k=1}^B\varphi_k(x)\int_a^b\varphi_k(y)\hat{f}_Y^{nve}(y)dy=\int_a^b K_B(x,y)\hat{f}_Y^{nve}(y)dy.$$

We conclude this brief discussion of the more general unfolding framework, by pointing out a slightly different observation scenario that one may consider in practice. Instead of observing an iid sample from f_Y, we might assume that we observe f_Y itself, subject to some error, $f_Y=\mathcal{L}f_X+\epsilon$. For example, ϵ can be thought of as white noise, meaning that we are able to observe

$$\langle f_Y,\varphi_k\rangle=\lambda_k\langle f_X,\varphi_k\rangle+\langle\epsilon,\varphi_k\rangle,\qquad k=1,2,...$$

with $\langle\epsilon,\varphi_k\rangle$ an iid white Gaussian noise sequence. This point of view would lead to the more classical statistical inverse problem framework, which is very well understood (see Cavalier [3]). Similar considerations apply, with spectral truncation and Tikhonov regularisation being the main approaches to relax the ill-posed problem. Variants of cross validation, or other methods for the tuning of the amount of regularisation such as Stein risk and the risk hull method have been studied in this context (see, Cavalier & Golubev [4]).

4 Discussion and Further Details

A question that took up a significant part of the discussion at the end of the unfolding session was, plainly stated, “to fold or to unfold"? In particular, should one attempt to estimate the folded function in Y-space and then unfold (invert the integral transform) in order to obtain their estimate in X-space, or rather should one look for the density in X-space such that, when folded (pushed forward through the integral transform), it would yield a density in Y-space that is most consistent with the data. From the mathematical point of view, the two views are essentially equivalent. If one chooses to fold instead of unfold, then one still needs to apply the same sort of regularisation: functions in X-space that are significantly far apart, may yield almost the same folded density in Y-space. To see this, consider two densities in X-space whose Fourier coefficients with respect to the first B eigenfunctions of the folding operator are identical, but the remaining coefficients are significantly different, though not different enough to counterbalance the decay of the eigenvalues of the operator. In this case, regularisation would amount to restricting one’s search on a subspace of X-space spanned by the first B eigenfunctions of the folding operator (where B would be a regularisation parameter to be tuned judiciously). Either approach could therefore be adopted, depending on what is most convenient from a practical point of view – but folding *does not circumvent* the problem of ill-posedness if regularisation is not applied.

Among the points raised was the use of cross-validation to select a regularisation parameter. The issue was connected to the feasibility of conducting the leave-one-out cross validation given that the sample size may be of the order of hundreds of thousands of observations. This, however, is not necessarily a problem: leave-one-out cross validation is employed in situations where the sample size is relatively

small, in order to avoid splitting the sample into two parts (a validation and an estimation part). For very large samples, one could employ an approach that is less computationally intensive, by splitting the data.

Another issue was that the folding operator is often only approximately known. This will indeed perturb things, but the main principles remain the same. For example, a simple approach to inject randomness into the operator in the setting of Section 3 would be by assuming that the eigenvalues are in fact a random element in the space of square summable sequences. This would not change the approach fundamentally. More complicated scenarios could introduce random eigenfunctions.

Below we provide a very short reference list on some of the topics covered in this overview. Silverman [10] provides an accessible introduction to nonparametric density estimation, while Meister [9] contains an elegant overview of the statistical deconvolution problem. One of the earliest papers in statistical deconvolution is Stefanski & Carroll [11], while a treatment of convergence rates can be found in Carroll & Hall [2] and Fan [7]. Details on bandwidth selection can be found in Delaigle & Hall [6], while Bissantz et al. [1] study the problem of constucting confidence bands for deconvolved density estimates. Hall & Lahiri [8] consider the problem of distribution estimation in the deconvolution setting. Finally, Cavalier [3] provides a review of some of the basic aspects of more general statistical inverse problems –such as the unfolding problem– in the context provided in the end of Section 3; in the same context, Cavalier & Hengartner [5] consider the case where the eigenvalues of the folding operator are noisy.

Acknowledgement

I wish to thank David Cox for kindly reading through a draft version of this report and providing useful comments. My thanks also go to Louis Lyons, for providing suggestions that helped improve the presentation.

References

[1] Bissantz, N., Dümbgen, L., Holzmann, H. & Munk, A. (2007). Nonparametric Confidence bands in deconvolution density estimation. *J. Roy. Stat. Soc. Ser. B,* ***69****: 483–506.*

[2] Carroll, R.J. & Hall, P. (1988). Optimal rates of convergence for deconvolving a density. *J. Am. Statist. Assoc.,* ***83****: 1184–1186.*

[3] Cavalier, L. (2008). Nonparametric statistical inverse problems. *Inverse Problems,* ***24****: 034004.*

[4] Cavalier, L. & Golubev, Yu. (2006). Risk hull method and regularization by projections of ill-posed inverse problems. *Annals of Statistics,* ***34****: 1653–1677.*

[5] Cavalier, L. & Hengartner, N.W. (2005). Adaptive estimation for inverse problems with noisy operators. *Inverse Problems,* ***21****: 1345.*

[6] Delaigle, A. & Hall, P. (2006). On the optimal kernel choice for deconvolution. *Stat. Prob. Lett.,* ***76****: 1594–1602.*

[7] Fan, J. (1991). On the optimal rates of convergence for non-parametric deconvolution problems. *Ann. Stat.* ***19****: 1257–1272.*

[8] Hall, P. & Lahiri, S.N.(2008). Estimation of distributions, moments and quantiles in deconvolution problems. *Ann. Stat.* ***36****: 2110–2134.*

[9] Meister, A. (2009). *Deconvolution Problems in Nonparametric Statistics*. Springer.

[10] Silverman, B.W. (1998). *Density Estimation*. Chapman & Hall.

[11] Stefanski, L.A. & Carroll, R.J. (1990). Deconvoluting kernel density estimators. *Statistics,* ***20****: 169–184.*

Unfolding Methods in Particle Physics*

Volker Blobel
University of Hamburg, Hamburg, Germany

Abstract
Measured distributions in particle physics are distorted by the finite resolution and limited acceptance of the detectors. The transformation to the underlying true distribution is called unfolding in particle physics and belongs to the class of linear inverse problems.

1 Inverse problems

1.1 Direct and inverse processes

The distributions $f(t)$ of a physics variable t to be measured in particle physics experiments are often not directly accessible. Because of limited acceptance and finite resolution the distribution $g(s)$ of the measured variable s is related to the distribution $f(t)$ by migration, distortions and transformations. Using Monte Carlo (MC) methods the direct process from an assumption $f(t)^{\text{model}}$ on the true distribution $f(t)$ to the expected measured distribution $g(s)$ can be simulated. The inverse process from the actually measured distribution $g(s)$ to the related *true* distribution $f(t)$ is difficult and ill-posed: small changes in the measured distribution can cause large changes in the reconstructed *true* distribution, if naive methods are used. In particle physics the inverse process is usually called *unfolding*. The direct and the inverse process

direct process (MC)	true/MC dist.	$f(t)$	$\Longrightarrow$	$g(s)$ measured dist.
inverse process (unfolding)	measured dist.	$g(s)$	$\Longrightarrow$	$f(t)$ true dist.

are described by the Fredholm integral equation of the first kind

$$\int_{\Omega} K(s,t)\ f(t)\,\mathrm{d}t = g(s) \tag{1}$$

with a Kernel function $K(s,t)$ describing the physical measurement process (Refs. [1]– [4] and references therein). In particle physics the Kernel function $K(s,t)$ is usually implicitly known from a Monte Carlo sample based on an assumption $f(t)^{\text{model}}$.

1.2 Discretization and linear solution

The inverse problem given by the Fredholm integral equation has to be discretized in order to allow a numerical solution, with the result of the linear equation

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}\ . \tag{2}$$

The relations between the functions/distributions and the matrix and vectors are:

true distribution $f(t) \Rightarrow \boldsymbol{x}$	n-vector of unknowns
measured distibution $g(s) \Rightarrow \boldsymbol{y}$	m-vector of measured data
Kernel $K(s,t) \Rightarrow \boldsymbol{A}$	rectangular m-by-n response matrix .

*This paper is the abridged version of a contribution to the forthcoming book O. Behnke et al. (eds.) with the working title: "Contemporary Data Analysis Methods", Wiley-VCH, ISBN 978-3-527-41058-3. The book will also contain more references.

The variables s, t and vectors $\boldsymbol{x}$, $\boldsymbol{y}$ are assumed to be one-dimensional in the following[1]. Several different discretization methods are possible. Real data are collected usually by integrating a signal over a short interval (bin), given by a grid $\{s_0,\, s_1, \ldots s_m\}$, often with equidistant bin limits in a histogram. The elements y_i correspond to integrals of $g(s)$ from s_{i-1} to s_i for $i = 1,\, 2, \ldots,\, m$ and are calculated according to equation (2) by the product $y_i = \boldsymbol{a}_i^{\mathrm{T}} \boldsymbol{x}$, where the vector $\boldsymbol{a}_i^{\mathrm{T}}$ is a row vector of matrix $\boldsymbol{A}$ and $y_i = A_{i1}x_1 + A_{i2}x_2 + \ldots + A_{in}x_n$. If the response is determined by a Monte Carlo sample, the same method can be used for the discretization $K(s,t) \Rightarrow \boldsymbol{A}$ and $f(t) \Rightarrow \boldsymbol{x}$; in this case element x_j is the average of $f(t)$ in bin j. Elements of the response matrix $\boldsymbol{A}$ are (positive) probabilities, and include the description of inefficiencies of the measurement detector. Other methods are possible, for example $f(t)$ can be discretized by a superposition of B-splines [3] which avoids discontinuities in the unfolded distribution, or the discretization can be based on numerical quadrature.

Assuming an accurate response matrix $\boldsymbol{A}$ and the relation $\boldsymbol{A}\,\boldsymbol{x}_{\text{exact}} = \boldsymbol{y}_{\text{exact}}$, the measured distribution deviates from the exact one only by statistical data errors. The data errors are represented by an m-vector $\boldsymbol{e}$, and the actually measured distribution $\boldsymbol{y}$ is given by

$$\boldsymbol{y} = \boldsymbol{y}_{\text{exact}} + \boldsymbol{e} = \boldsymbol{A}\,\boldsymbol{x}_{\text{exact}} + \boldsymbol{e}\ .$$

In particle physics the statistical properties of the measurements are usually well known. Often the elements of the vector $\boldsymbol{y}$ are counts, following Poisson statistics. In general the expectation value and variance are

$$\mathrm{E}\left[\boldsymbol{y}\right] = \boldsymbol{y}_{\text{exact}} \qquad\qquad \mathrm{V}\left[\boldsymbol{y}\right] = \mathrm{V}\left[\boldsymbol{e}\right] = \mathrm{E}\left[\boldsymbol{e}\boldsymbol{e}^{\mathrm{T}}\right] = \boldsymbol{V}_y\ , \tag{3}$$

i.e., an unbiased measurement $\boldsymbol{y}$ with $\mathrm{E}\left[\boldsymbol{e}\right] = 0$ is assumed, and the covariance matrix $\boldsymbol{V}_y$ of the measurement[2] is known.

In particle physics, unlike other fields, not only the result vector $\boldsymbol{x}$ has to be determined, but also the covariance matrix $\boldsymbol{V}_x$ of the result vector . If the linear Fredholm equation is solved for the estimate $\widehat{\boldsymbol{x}}$ by a linear transformation of the data vector according to $\widehat{\boldsymbol{x}} = \boldsymbol{A}^{\dagger}\boldsymbol{y}$, the propagation of the data uncertainties to the unfolding uncertainties is straightforward: $\boldsymbol{V}_x = \boldsymbol{A}^{\dagger}\boldsymbol{V}_y{\boldsymbol{A}^{\dagger}}^{\mathrm{T}}$. The case $m = n$ with a quadratic matrix $\boldsymbol{A}$ could be solved by the inverse matrix $\boldsymbol{A}^{\dagger} = \boldsymbol{A}^{-1}$, but often the matrix $\boldsymbol{A}$ has a bad condition or is even singular and $m = n$ should be avoided. In the recommended case $m > n$ the n-by-m matrix $\boldsymbol{A}^{\dagger}$ can be constructed from the m-by-n matrix $\boldsymbol{A}$ and used to determine the estimate $\widehat{\boldsymbol{x}}$:

$$\widehat{\boldsymbol{x}} = \boldsymbol{A}^{\dagger}\boldsymbol{y} = \boldsymbol{A}^{\dagger}\,\boldsymbol{y}_{\text{exact}} + \boldsymbol{A}^{\dagger}\boldsymbol{e} = \boldsymbol{A}^{\dagger}\boldsymbol{A}\,\boldsymbol{x}_{\text{exact}} + \boldsymbol{A}^{\dagger}\boldsymbol{e}\ . \tag{4}$$

The pseudo-inverse $\boldsymbol{A}^{\dagger}$, also called Moore-Penrose generalized inverse, satisfying the relation $\boldsymbol{A}^{\dagger}\boldsymbol{A} = \boldsymbol{I}$, is a generalization of the inverse matrix, and allows the solution in the naive least squares sense, derived from the requirement

$$\min_{\boldsymbol{x}} F(\boldsymbol{x}) \quad \text{with} \quad F(\boldsymbol{x}) = \left(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\right)^{\mathrm{T}} \boldsymbol{V}_y^{-1} \left(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\right)\ , \tag{5}$$

where the inverse of the data covariance matrix $\boldsymbol{V}_y$ is included to take into account the different accuracy of the elements of the data vector. The least squares solution from the normal-equations formalism can be expressed by the pseudo-inverse

$$\boldsymbol{A}^{\dagger} = \left(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{V}_y^{-1}\boldsymbol{A}\right)^{-1} \boldsymbol{A}^{\mathrm{T}}\boldsymbol{V}_y^{-1}\ ,$$

(with $\boldsymbol{A}^{\dagger}\boldsymbol{A} = \boldsymbol{I}$) with the matrix $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{V}_y^{-1}\boldsymbol{A} = \boldsymbol{C}$. The covariance matrix $\boldsymbol{V}_x$ is given by $\boldsymbol{V}_x = \boldsymbol{A}^{\dagger}\boldsymbol{V}_y{\boldsymbol{A}^{\dagger}}^{\mathrm{T}} = \left(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{V}_y^{-1}\boldsymbol{A}\right)^{-1} = \boldsymbol{C}^{-1}$. Although the estimate $\widehat{\boldsymbol{x}}$ has the expectation $\boldsymbol{x}_{\text{exact}}$ (see equation

[1] The variables and the vectors can be multi-dimensional in practice, even with different dimensions for the true and the measured distribution.

[2] Covariance matrices are written with a subscript like $\boldsymbol{V}_y$; matrices $\boldsymbol{V}$ without subscripts are orthogonal matrices from a decomposition (Section 2).

(4)) because of $\boldsymbol{A}^\dagger \boldsymbol{A} \equiv \boldsymbol{I}$, this naive solution is often not satisfactory. It can be strongly oscillating with large negative correlation coefficients between neighbouring points and large positive correlation coefficients between next-to-immediate neighbours.

1.3 Parametrized unfolding

Unfolding was considered above to determine a discretized version $\boldsymbol{x}$ of a distribution $f(t)$ without a specific parametrization. A predicted probability density function (pdf) $f(t)$ without unknown parameters can be checked for compatibility with the data by *folding*; however folding does not provide information on the *sensitivity*. If a certain parametrization $f(t) \equiv f(t;\, \boldsymbol{a})$ depending on a vector of parameters $\boldsymbol{a}$ (to be fitted) is assumed, motivated e.g., by the theoretical analysis of the problem, this parametrization can be directly used in unfolding, using the response matrix $\boldsymbol{A}$, without the need to introduce a regularization. The bin content y_i is approximated using the elements of an auxiliary vector $\boldsymbol{x}$:

$$y_i = a_i^{\mathrm{T}} \boldsymbol{x} \quad \text{with} \quad x_j(\boldsymbol{a}) = \int_{t_{j-1}}^{t_j} \mathrm{d}t\; f(t;\, \boldsymbol{a}) \qquad\qquad j = 1,\, 2, \ldots,\, n \tag{6}$$

assuming a grid $\{t_0,\, t_1, \ldots t_n\}$ for the variable t. Unfolding is then the solution of the minimization problem

$$\min_{\boldsymbol{a}} F(\boldsymbol{a}) \quad \text{with} \quad F(\boldsymbol{a}) = (\boldsymbol{A}\boldsymbol{x}(\boldsymbol{a}) - \boldsymbol{y})^{\mathrm{T}}\, \boldsymbol{V}_y^{-1}\, (\boldsymbol{A}\boldsymbol{x}(\boldsymbol{a}) - \boldsymbol{y}) \;. \tag{7}$$

The function value $F(\widehat{\boldsymbol{a}}) = \chi^2_y$ should follow the χ^2-distribution with $m - n_{\mathrm{par}}$ degrees of freedom, if the parametrization has n_{par} parameters. A standard fit program like MINUIT (CERN) with numerical derivatives can determine the parameter vector $\widehat{\boldsymbol{a}}$ and its covariance matrix.

An example of a parametrized unfolding, taken from Ref. [5], is shown in Figure 1. A pdf $f(t) = (1 + a\,t) \,/\, (1 + a/2)$ with t in the interval $[0, 1]$ is measured with a Gaussian resolution with standard deviation of 0.3. Figure 1(a) shows a simulated example for $a = 1$ with 5 000 entries of the measured distribution in the interval $[-0.3, 1.3]$. A 20-by-20 response matrix is determined by a simulation of 50 000 cases, using a uniform distribution (parameter $a = 0$) in $[0, 1]$. The result of the parameter fit according to equation (7) with the result $\widehat{a} = 1.09 \pm 0.18$ is shown in Figure 1(b) together with the simulated true histogram. Figure 1(c) shows the histogram of the fitted slope a from 10^5 simulations, together with a Gaussian curve of standard deviation 0.18; the fitted parameter has on average the correct value $a = 1$ with a slightly asymmetric distribution. These results agree with the results of Ref. [5].

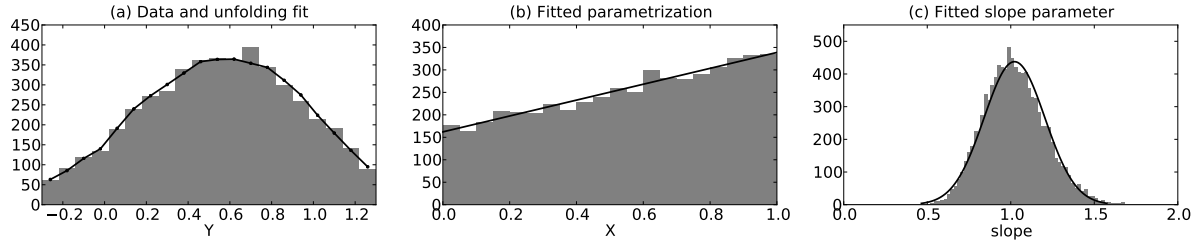


Fig. 1: Example for parametrized unfolding

1.4 Convolution and deconvolution

A function $f(t)$ with period 1 can be approximated by a sum of cosine functions, which is a complete system, periodic in $[0, 1]$ and orthogonal in the interval $0 \le t \le 1$. The approximation is given by $f(t) = a_0 + a_1 \cos(\pi t) + a_2 \cos(2\pi t) + \cdots$. The terms are the basis functions of the discrete cosine transformation, shown in Figure 2. The special case of a Kernel $K(s,\, t) \equiv K(s - t)$ is called a *convolution* and the inverse process is called *deconvolution*. A convolution of the function $f(t)$ by a Gaussian

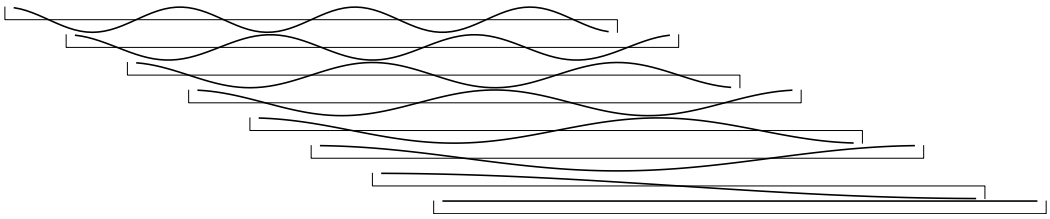


Fig. 2: The first eight basis functions of the discrete cosine transformation over the range $0 \ldots 1$

resolution function with standard deviation σ is considered. For a single term $\cos(k\pi t)$ the form of the term is not changed by the convolution with the Gaussian:

$$\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(s-t)^2}{2\sigma^2}\right) \times \cos(k\pi t)\ \mathrm{d}t = \exp\left(-\frac{(k\pi\sigma)^2}{2}\right) \times \cos(k\pi s)\ ,$$

but the amplitude is attenuated by an exponential factor, which will become $\ll 1$ for larger indices k. The convoluted function $g(s)$ (see equation (1)) is *smoother* than $f(t)$ and can be approximated again by a cosine sum with coefficients α_k instead of a_k. The coefficients α_k of the convoluted function $g(s)$ will become small and negligible asymptotically much faster than of the original function $f(t)$. Deconvolution is simple in this case: the coefficients α_k, determined from $g(s)$, have to be multiplied by the inverse exponential factor, to reconstruct the coefficients a_k. With increasing index k, the exponential correction factors of the coefficients α_k soon become extremely large, increasing the relative uncertainty of the coefficients by a factor $\gg 1$. Thus the number of terms of the original function $f(t)$ which can be reconstructed is *limited* because of the finite resolution.

2 Solution with orthogonalization

2.1 Singular value decomposition (SVD)

The standard numerical method for the analysis of ill-posed problems $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}$ is the singular value decomposition (SVD) of the m-by-n matrix matrix $\boldsymbol{A}$, defined for any m and n. Assuming $m \geq n$ (called *thin* SVD) the SVD is of the form[3] with elements

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T = \sum_{i=1}^{n} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^{\mathrm{T}}\ ,$$

where $\boldsymbol{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n) \in \mathbb{R}^{m\times n}$ and $\boldsymbol{V}(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n) \in \mathbb{R}^{n\times n}$ are matrices with orthonormal columns and the diagonal matrix $\boldsymbol{\Sigma} = \mathbf{diag}\{\sigma_1, \ldots, \sigma_n\} = \boldsymbol{U}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{V}$ has non-negative diagonal elements σ_i, called singular values, in non-increasing order. The *condition* of matrix $\boldsymbol{A}$ is defined as the ratio of the largest to the smallest singular vector: $\mathrm{cond}\,(\boldsymbol{A}) = \sigma_1/\sigma_n$. The condition is an upper bound on the magnification factor of the ratio of relative errors of the estimate $\widehat{\boldsymbol{x}}$ to the data $\boldsymbol{y}$. The m-vectors $\boldsymbol{u}_i$ and the n-vectors $\boldsymbol{v}_i$ are called left and right singular vectors of $\boldsymbol{A}$. The SVD matrices with the property $\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} = \boldsymbol{V}^{\mathrm{T}}\boldsymbol{V} = \boldsymbol{V}\boldsymbol{V}^{\mathrm{T}} = \boldsymbol{I}$ will be used for the least squares solution $\widehat{\boldsymbol{x}}$ of the problem. The singular vectors $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$ have an increasing number of sign-changes with increasing index and decreasing singular value, similar to the cosine functions in Figure 2 .

In order to take the uncertainty of the data $\boldsymbol{y}$, given by the covariance matrix $\boldsymbol{V}_y$, into account, a pre-scaling (also called pre-whitening) of the problem is required. For uncorrelated data this is achieved by dividing the rows of the linear system by the standard deviation $\sqrt{(\boldsymbol{V}_y)_{ii}}$ of the data. The fastest

[3]This is a decomposition into *outer* products of two vectors. The outer product $\boldsymbol{a}\boldsymbol{b}^{\mathrm{T}}$ of two vectors $\boldsymbol{a}$ and $\boldsymbol{b}$, also called dyadic product, is a rank-1 matrix $\boldsymbol{B}$ with elements $B_{jk} = a_j b_k$.

method for correlated data is based on the Cholesky decomposition of the matrix $\boldsymbol{V}_y = \boldsymbol{R}^{\mathrm{T}}\boldsymbol{R}$ with an upper triangular matrix $\boldsymbol{R}$. In the following it is assumed that a pre-scaling of $\boldsymbol{A}$ and $\boldsymbol{y}$ has already been done (i.e., $\boldsymbol{A} := (\boldsymbol{R}^{-1})^{\mathrm{T}}\boldsymbol{A}$ and $\boldsymbol{y} := (\boldsymbol{R}^{-1})^{\mathrm{T}}\boldsymbol{y}$) with the result $\boldsymbol{V}_y = \boldsymbol{I}$, before the singular value decomposition. If the elements y_i are counts with standard deviation $\sqrt{y_i}$, the magnitude of the singular values is proportional to the number of measured events. For the case of a Gaussian response matrix with standard deviation σ the decrease of the singular values is approximately described by the exponential factor $\exp\left(-ak^2\sigma^2\right)$ (with some constant a) in the convolution example from Section 1.4.

2.2 Symmetric eigenvalue decomposition

The eigenvalue decomposition of a symmetric n-by-n matrix $\boldsymbol{C}$ is the orthogonalization method to be used in maximum likelihood methods based of the Poisson statistics [3] ($\boldsymbol{C}$ = Hessian) and in normal-equations least squares ($\boldsymbol{C} = \boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$, assuming pre-scaling of matrix $\boldsymbol{A}$), and can be achieved by a SVD. In the SVD of this matrix the left and right singular vectors are identical:

$$\boldsymbol{C} = \boldsymbol{A}^{\mathrm{T}}\boldsymbol{A} = \left(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}}\right)^{\mathrm{T}}\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}} = \boldsymbol{V}\boldsymbol{\Sigma}^2\boldsymbol{V}^{\mathrm{T}} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^{\mathrm{T}} \; .$$

The diagonal matrix $\boldsymbol{\Lambda} = \mathbf{diag}\left\{\lambda_1,\, \lambda_2, \ldots\right\}$ has non-negative diagonal elements λ_i, called eigenvalues, equal to the square of the singular values σ_i of the matrix $\boldsymbol{A}$. The symmetric eigenvalue decomposition of the matrix $\boldsymbol{C}$ is mathematically equivalent to the singular value decomposition of the matrix $\boldsymbol{A}$.

2.3 Least squares using the SVD

The use of the SVD in least square problems allows some insight into the structure of the matrix $\boldsymbol{A}$ of ill-posed problems $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}$. The matrix product $\boldsymbol{A}\boldsymbol{x}$ expressed using the SVD matrices

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T\boldsymbol{x} = \sum_{j=1}^{n} \sigma_j \left(\boldsymbol{v}_j^{\mathrm{T}}\boldsymbol{x}\right)\boldsymbol{u}_j = \boldsymbol{y}$$

shows that contributions to $\boldsymbol{y}$ with small singular values σ_j, corresponding to higher-frequency contributions, are suppressed. If all singular values are non-zero, the least squares estimate $\widehat{\boldsymbol{x}}$ is given by

$$\widehat{\boldsymbol{x}} = \boldsymbol{A}^{\dagger}\boldsymbol{y} = \boldsymbol{V}\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{U}^{\mathrm{T}}\boldsymbol{y}\right) = \sum_{j=1}^{n} \frac{1}{\sigma_j}\left(\boldsymbol{u}_j^{\mathrm{T}}\boldsymbol{y}\right)\boldsymbol{v}_j = \sum_{j=1}^{n} \frac{1}{\sigma_j} c_j \boldsymbol{v}_j \; . \tag{8}$$

The data $\boldsymbol{y}$ with unit covariance matrix are transformed by $\boldsymbol{U}^{\mathrm{T}}$ to an n-vector $\boldsymbol{c} = \boldsymbol{U}^{\mathrm{T}}\boldsymbol{y}$ with unit covariance matrix $\boldsymbol{V}_c = \boldsymbol{I}$, representing the transformed measurement. The elements $c_j = \boldsymbol{u}_j^{\mathrm{T}}\boldsymbol{y}$ of $\boldsymbol{c}$, called *Fourier coefficients*, tend to decrease rather fast towards small values for larger indices j, if the distribution described by $\boldsymbol{y}$ is smooth. The coefficients c_j are *independent* and, having a *variance of* 1, show the significance of the corresponding contribution to the estimate $\widehat{\boldsymbol{x}}$. The value of a Fourier coefficient c_j will follow a Gaussian $N(0,1)$, if the exact value is small compared to the standard deviation 1. The expression (8) for the estimate $\widehat{\boldsymbol{x}}$ shows that the contribution to the estimate $\widehat{\boldsymbol{x}}$ related to a single Fourier coefficient c_j is multiplied by the *inverse* of the singular value σ_j. Small singular values σ_j will generate large fluctuations in the unfolding result $\widehat{\boldsymbol{x}}$, and can make the result unacceptable. The calculation of the uncertainty of the estimate $\widehat{\boldsymbol{x}}$ is straightforward, because of the linear transformation of the data $\boldsymbol{y}$ in the expression (8); the covariance matrix is

$$\boldsymbol{V}_x = \boldsymbol{A}^{\dagger}\boldsymbol{V}y\boldsymbol{A}^{\dagger^{\mathrm{T}}} = \boldsymbol{V}\boldsymbol{\Sigma}^{-2}\boldsymbol{V}^{\mathrm{T}} = \sum_{j=1}^{n}\left(\frac{1}{\sigma_j^2}\right)\boldsymbol{v}_j\boldsymbol{v}_j^{\mathrm{T}} \; . \tag{9}$$

In other methods e.g., iterative methods (Section 4) an estimate $\widehat{\boldsymbol{x}}$ is determined without the construction of a transformation matrix like $\boldsymbol{A}^{\dagger}$, which makes the above uncertainty calculation impossible.

2.4 Null space and truncated SVD

The SVD defines by matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ a new basis for the measured data and the unfolding result in a frequency space. The measured data $\boldsymbol{y}$ are transformed to independent Fourier coefficients $c_j = \boldsymbol{u}_j^{\mathrm{T}}\boldsymbol{y}$ with fixed standard deviation 1 (*white* noise). The least-square estimate $\widehat{\boldsymbol{x}}$ can be written in the form $\widehat{\boldsymbol{x}} = \sum_j d_j \boldsymbol{v}_j$ with coefficients $d_j = c_j/\sigma_j$, that are still independent, but have standard deviations $1/\sigma_j$ increasing with index j; this property could be called *blue* noise because the uncertainty is increasing with the frequency. Typically the singular values σ_j of a response matrix $\boldsymbol{A}$ decrease to small values without a clear gap between large and small singular values. Due to rounding and other errors there will be no exactly zero singular values, but taking into account potential uncertainties of the elements of matrix $\boldsymbol{A}$ at least a few singular values may be effectively zero, reducing the *effective rank* of the matrix $\boldsymbol{A}$ to a number p (less than n), which is an upper limit on the number of contributions. Especially if the response matrix is determined by a Monte Carlo simulation there are unavoidable uncertainties in the elements. A tolerance δ can be defined to determine the effective rank p by $\sigma_p > \delta \geq \sigma_{p+1}$ with

$$\delta = \epsilon \times \max_{1\leq i\leq m} \sum_{j=1}^{n} |A_{ij}| \ , \tag{10}$$

where e.g., $\epsilon = 0.01$, if the elements A_{ij} are correct to about two digits, as is the case of typical Monte Carlo calculations. Small singular values $\sigma_j < \delta$ would give meaningless contributions to the solution. Assuming an effective rank of p (less than n), the estimate $\widehat{\boldsymbol{x}}$ of equation (8) can be written in the form

$$\widehat{\boldsymbol{x}} = \underbrace{\sum_{j=1}^{p} d_j \boldsymbol{v}_j}_{\boldsymbol{x}_{\text{range}} \in \mathbb{R}^p} + \underbrace{\sum_{j=p+1}^{n} \widetilde{d}_j \boldsymbol{v}_j}_{\boldsymbol{x}_{\text{null}} \in \mathbb{R}^{n-p}} \ . \tag{11}$$

The first term $\boldsymbol{x}_{\text{range}}$, with contributions $d_j = c_j/\sigma_j$, is a rather well-defined element of a p-dimensional subspace of the $\mathbb{R}^n$, but the second term $\boldsymbol{x}_{\text{null}}$ has arbitrary contributions $\widetilde{d}_j$, which in the product $\boldsymbol{A}\widehat{\boldsymbol{x}}$, multiplied by the singular value σ_j, have essential no effect on the expected data $\widehat{\boldsymbol{y}}$. Because the two terms in $\widehat{\boldsymbol{x}} = \boldsymbol{x}_{\text{range}} + \boldsymbol{x}_{\text{null}}$ (equation (11)) are orthogonal, the squared norm of $\widehat{\boldsymbol{x}}$ is the sum of the two squared norms $\left\|\boldsymbol{x}_{\text{range}} + \boldsymbol{x}_{\text{null}}\right\|^2 = \left\|\boldsymbol{x}_{\text{range}}\right\|^2 + \left\|\boldsymbol{x}_{\text{null}}\right\|^2$. The solution recommended in textbooks is the minimum-norm solution with $\widehat{\boldsymbol{x}}_{\text{null}} = 0$ and $\|\widehat{\boldsymbol{x}}\| = \left\|\boldsymbol{x}_{\text{range}}\right\|$. In this case the n-by-n covariance matrix $\boldsymbol{V}_x$ has a rank defect of $n - p$ and cannot be inverted. Alternatively the dimension of estimate $\widehat{\boldsymbol{x}}$ can be reduced to p, with a full-rank p-by-p covariance matrix $\boldsymbol{V}_x$ (see Section 3.6). In a simulation a data sample of 5000 events is generated with $n = 20$ and $m = 40$, assuming a Gaussian response. The effective rank of $\boldsymbol{A}$ as estimated from equation (10) with $\epsilon = 0.01$ is 18. Figure 3 shows the Fourier coefficients $|c_j|$ and the contributions $|d_j|$. The coefficients c_j for $j \geq 8$ are insignificant, giving a lower limit of the number of contributions. The (insignificant) contributions increase after $j = 10$ and the last contributions ($j \geq 16$) would dominate the result. Truncation after $j = 10$ gives an acceptable result without bias.

3 Regularization methods

3.1 Regularization

The standard method for the solution of ill-posed problems is the *regularization* method [1, 2]. The expression to be minimized w.r.t. the unfolding result includes the least squares (equation (5)) or (negative) log-likelihood expression, which ensure a good description of the measured distribution. A second term $\Omega(\boldsymbol{x})$, often of the form $\Omega(\boldsymbol{x}) = \|\boldsymbol{L}\boldsymbol{x}\|^2$ with a certain matrix $\boldsymbol{L}$, requires certain properties like smoothness of the unfolding result and contributes with a weight, given by a regularization parameter $\tau > 0$:

$$\min_{\boldsymbol{x}} F(\boldsymbol{x}) + \tau \|\boldsymbol{L}\boldsymbol{x}\|^2 \ .$$

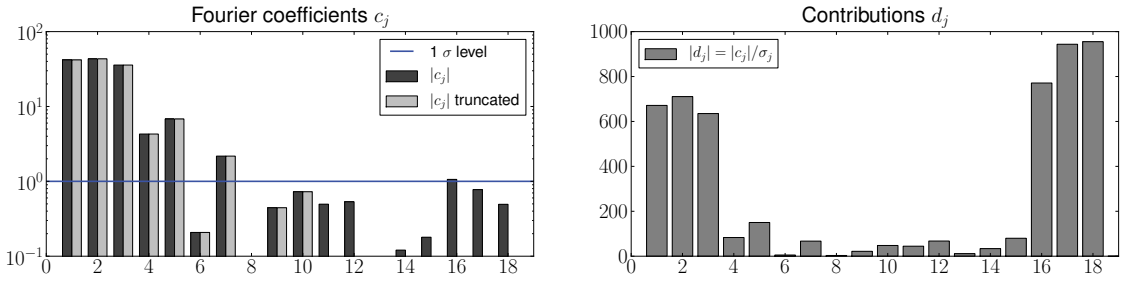


Fig. 3: Truncation

In the regularized solution of the least squares case (equation (5)) the matrix $\boldsymbol{A}^{\dagger}$ is replaced by the regularized matrix $\boldsymbol{A}^{\#}$:

$$\widehat{\boldsymbol{x}} = \boldsymbol{A}^{\#}\boldsymbol{y} = \left[\left(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A} + \tau \boldsymbol{L}^{\mathrm{T}}\boldsymbol{L}\right)^{-1}\boldsymbol{A}^{\mathrm{T}}\right]\boldsymbol{y}\ . \tag{12}$$

The regularization term $\tau \boldsymbol{L}^{\mathrm{T}}\boldsymbol{L}$ is added to the matrix $\boldsymbol{C} = \boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$ of the normal equations, and inserting $\boldsymbol{y} = \boldsymbol{A}\,\boldsymbol{x}_{\mathrm{exact}} + \boldsymbol{e}$ one obtains

$$\widehat{\boldsymbol{x}} = \boldsymbol{A}^{\#}\boldsymbol{A}\,\boldsymbol{x}_{\mathrm{exact}} + \boldsymbol{A}^{\#}\boldsymbol{e} = \boldsymbol{x}_{\mathrm{exact}} + \underbrace{\left(\boldsymbol{A}^{\#}\boldsymbol{A} - \boldsymbol{I}\right)\boldsymbol{x}_{\mathrm{exact}}}_{\text{systematic error}} + \underbrace{\left(\boldsymbol{A}^{\#}\boldsymbol{e}\right)}_{\text{statistical error}}\ . \tag{13}$$

The product $\boldsymbol{\Xi} \equiv \boldsymbol{A}^{\#}\boldsymbol{A}$ is called the *resolution matrix*. For the regularization scheme the resolution matrix is not equal to the unit matrix, and thus the method has a systematic bias $(\boldsymbol{\Xi} - \boldsymbol{I})\,\boldsymbol{x}_{\mathrm{exact}}$. The fact that the regularized solution has a potential bias of the estimate, which depends on the details of the exact distribution $\boldsymbol{x}_{\mathrm{exact}}$, is connected with the attempt to reduce the unnatural oscillations, which are unmeasurable. The *smoothing* effect of the resolution matrix gives no or small systematic errors for smooth exact distributions, and large systematic deviation for unphysical oscillating distributions. The measured distribution $\boldsymbol{y}$ has to be compared with the distribution $\widehat{\boldsymbol{y}}$ corresponding to the estimated unfolded distribution $\widehat{\boldsymbol{x}}$ and given by $\widehat{\boldsymbol{y}} = \boldsymbol{A}\widehat{\boldsymbol{x}} = \boldsymbol{A}\boldsymbol{A}^{\#}\boldsymbol{y}$, where the m-by-m product matrix $\boldsymbol{A}\boldsymbol{A}^{\#}$ is often called the *influence matrix*. The agreement between the measured data $\boldsymbol{y}$ and the vector $\widehat{\boldsymbol{y}}$ predicted by the influence matrix and checked with $\chi^2_y = (\widehat{\boldsymbol{y}} - \boldsymbol{y})^{\mathrm{T}}(\widehat{\boldsymbol{y}} - \boldsymbol{y})$ has to be acceptable.

The deviation of the resolution matrix $\boldsymbol{\Xi} = \boldsymbol{A}^{\#}\boldsymbol{A}$ from the unit matrix $\boldsymbol{I}$, which corresponds to a potential bias, should avoid the unnatural properties of naive unregularized solutions. Sometimes this term is called a penalty function, which seems to express a certain impact on the solution with a bias of the regularized result. For applications in particle physics a non-negligible bias is not acceptable. Below it is shown that the regularization ansatz can be used to separate the significant from the insignificant contributions of the result *without the introduction of a disturbing bias*.

3.2 Norm regularization

The simplest case is the *norm regularization* with $\boldsymbol{L} = \boldsymbol{I}$. For a given value of τ the estimate $\widehat{\boldsymbol{x}}$ can be determined by standard methods of linear algebra (matrix inversion), because of the good condition of the combined matrix. However the solution by the SVD is simple in this case and has several advantages, especially as it allows a clear understanding of the effects of regularization. Using the SVD the solution can be written in the form

$$\widehat{\boldsymbol{x}} = \boldsymbol{V}\ \underbrace{\left[\left(\boldsymbol{\Sigma}^2 + \tau\boldsymbol{I}\right)^{-1}\boldsymbol{\Sigma}^2\right]}_{\text{filter factor matrix } \boldsymbol{F}}\ \boldsymbol{\Sigma}^{-1}\ \underbrace{\left(\boldsymbol{U}^{\mathrm{T}}\boldsymbol{y}\right)}_{\text{coeff.}\boldsymbol{c}} = \left(\boldsymbol{V}\ \boldsymbol{F}\ \boldsymbol{\Sigma}^{-1}\boldsymbol{U}^{\mathrm{T}}\right)\boldsymbol{y}\ ,$$

where the matrix $\boldsymbol{F}$ is diagonal with elements φ_j. Comparison with the unregularized solution (8) shows the additional filter factors φ_j for each term with a strength which depends on the regularization

parameter τ, while the Fourier coefficients $\boldsymbol{c}_j$ are defined as before. The estimate $\widehat{\boldsymbol{x}}$ and its covariance matrix (compare (9)) can be expressed by sums:

$$\widehat{\boldsymbol{x}} = \sum_{j=1}^{n} \frac{1}{\sigma_j}\, \varphi_j\, c_j \boldsymbol{v}_j \qquad \boldsymbol{V}_x = \sum_{j=1}^{n} \left(\frac{1}{\sigma_j^2}\right)\, \varphi_j^2\, \boldsymbol{v_j} \boldsymbol{v_j}^{\mathrm{T}} \qquad \text{with} \quad \varphi_j = \frac{\sigma_j^2}{\sigma_j^2 + \tau} \tag{14}$$

(the squared singular values σ_j^2 are replaced by eigenvalues in case of diagonalization). The filter factors φ_j represent a smooth cut-off (with $\varphi_k = 0.5$ if $\tau = \sigma_k^2$), which can avoid a certain oscillating behaviour (Gibbs phenomenon) in the truncation case. No bias will be introduced if the selected regularization parameter τ is small enough to reduce only the insignificant Fourier coefficients.

The norm regularization corresponds to the original regularization proposal by Tikhonov and by Philipps. The regularization parameter τ can be interpreted as the introduction of the a-priori measurement error $s_{\text{reg}} = 1/\sqrt{\tau}$ for each component of the vector $\boldsymbol{x}$. Individual values of $s_{j,\text{reg}}$ for the components could be introduced, corresponding to a regularization term $\Omega(\boldsymbol{x}) = \sum_j x_j^2/s_j^2$. Norm regularization can be used for unfolding problems with rather smooth solutions $\overline{\boldsymbol{x}}$, requiring only a small number of Fourier coefficients; in other cases some modifications are advisable. One possibility is to change the regularization term $\Omega(\boldsymbol{x}) = \|\boldsymbol{L}\boldsymbol{x}\|^2$ to a term $\Omega(\boldsymbol{x}) = \|\boldsymbol{L}\,(\boldsymbol{x} - \boldsymbol{x}_0)\|^2$ with some a-priori assumption $\boldsymbol{x}_0$ on the resulting vector $\boldsymbol{x}$; this will reduce the number of significant terms. Another possibility is to make the Monte Carlo simulation with an a-priori assumption $f(t)^{\text{model}}$ about the function $f(t)$, and to include the function $f(t)^{\text{model}}$ already in the definition of the response matrix,

$$\int_{\Omega} \left[K(s,t)\; f(t)^{\text{model}}\right] f^{\ddagger}(t)\, \mathrm{d}t = g(s)\;;$$

only an almost constant correction function $f^{\ddagger}(t)$ has to be determined with $f(t) = f(t)^{\text{model}} f^{\ddagger}(t)$. This option is available in unfolding methods of particle physics [3, 4]. The elements A_{ij} of the matrix $\boldsymbol{A}$, which includes $f(t)^{\text{model}}$, are now integers, the number of Monte Carlo events from bin j of $\boldsymbol{x}$, measured in bin i of $\boldsymbol{y}$.

3.3 Regularization based on derivatives

Another regularization scheme is based on *derivatives*; most popular are the second derivates, and this scheme often has advantages over the norm regularization. The matrix $\boldsymbol{L}$ is rather simple if equidistant bins are used. For example the second derivative in bin j is proportional to $(-x_{j-1} + 2x_j - x_{j+1})$ and corresponds to a row $\ldots\ -1\quad 2\ -1\ \ldots$ of the matrix $\boldsymbol{L} \in \mathbb{R}^{(n-2)\times n}$. The solution (12) can again be obtained, for a given regularization factor τ, by matrix inversion. However, a solution using orthogonalization provides an understanding of the details and the separation of significant from insignificant contributions.

Orthogonalization is more complicated than for the norm regularization, because the term $\tau \boldsymbol{L}^{\mathrm{T}}\boldsymbol{L}$ is not diagonal. The *generalized singular value decomposition* can to be used for the corresponding orthogonalization. The orthogonal solution is formally equivalent to the solution (12), with a different definition of the singular or eigenvalues. Compared to the norm regularization the Fourier coefficients refer to a rotated system according to $\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L}$. If, alternatively, the eigenvalue decomposition is used, two rotations (and a scaling) are required to diagonalize simultaneously [3] the two symmetrical matrices $\boldsymbol{C} = \boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$ and $\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L}$ for the solution of the normal equation $\left(\boldsymbol{C} + \tau \boldsymbol{L}^{\mathrm{T}}\boldsymbol{L}\right)\boldsymbol{x} = \boldsymbol{b}$ with $\boldsymbol{b} = \boldsymbol{A}^{\mathrm{T}}\boldsymbol{y}$. The first diagonalization $\boldsymbol{C} = \boldsymbol{U}_1 \boldsymbol{\Lambda} \boldsymbol{U}_1^{\mathrm{T}}$ is used to rewrite the equation in the form

$$\boldsymbol{U}_1 \boldsymbol{\Lambda}^{1/2}\, (\boldsymbol{I} + \tau \boldsymbol{M})\, \boldsymbol{\Lambda}^{1/2} \boldsymbol{U}_1^{\mathrm{T}} \boldsymbol{x} = \boldsymbol{b}$$

with the transformed regularization matrix $\boldsymbol{M} = \boldsymbol{\Lambda}^{-1/2} \boldsymbol{U}_1^{\mathrm{T}} \left(\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L}\right) \boldsymbol{U}_1 \boldsymbol{\Lambda}^{-1/2}$. The second diagonalization $\boldsymbol{M} = \boldsymbol{U}_2 \boldsymbol{S} \boldsymbol{U}_2^{\mathrm{T}}$ is used to rewrite the equation in the form

$$\boldsymbol{R}\, (\boldsymbol{I} + \tau \boldsymbol{S})\, \boldsymbol{R}^{\mathrm{T}} \boldsymbol{x} = \boldsymbol{b}$$

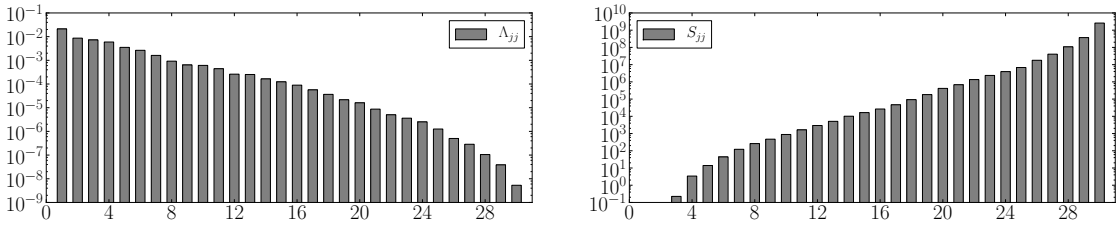


Fig. 4: Eigenvalues Λ_{jj} and S_{jj}

$$\widehat{\boldsymbol{x}} = \left(\boldsymbol{R}^{\mathrm{T}}\right)^{-1} \underbrace{(\boldsymbol{I} + \tau \boldsymbol{S})^{-1}}_{\text{filter factor matrix } \boldsymbol{F}} \underbrace{\left(\boldsymbol{R}^{-1}\boldsymbol{b}\right)}_{\text{coeff.}\boldsymbol{c}}$$

using matrix $\boldsymbol{R} = \boldsymbol{U}_1 \boldsymbol{\Lambda}^{1/2} \boldsymbol{U}_2$ and the inverse $\boldsymbol{R}^{-1} = \boldsymbol{U}_2^{\mathrm{T}} \boldsymbol{\Lambda}^{-1/2} \boldsymbol{U}_1^{\mathrm{T}}$. The filter factor is now given by $\varphi_j = 1/(1 + \tau S_{jj})$ with the element S_{jj} of the diagonal matrix $\boldsymbol{S}$. Figure 4 shows the eigenvalues Λ_{jj} and S_{jj} for the example of Section 3.5, both with increasing frequency from the left to the right; the stronger separation of low- and high frequency contributions by the curvature is visible. Note that the definition of the elements S_{jj} is inverse to the definiton of the elements Λ_{jj}, and the first two eigenvalues S_{11} and S_{22}, corresponding to a constant and to a linear contribution (without a curvature), are zero.

3.4 Determination/Selection of the regularization parameter

There is no generally accepted and unique method to determine the regularization parameter τ, applicable for all cases. An often used method is the L-curve method [1,2]. A *lower limit* of $\tau \approx \sigma_p^2$ is given by the size of the singular value σ_p for an effective rank of p (Section 2.4). An *upper limit* of the regularization parameter τ is determined by the overall χ_y^2 of the agreement of the observed distribution. Each Fourier coefficient, removed in the truncation method, will increase the χ_y^2 value by c_j^2 and n_{df} by one. As long as the coefficients c_j of variance one are compatible with mean zero, the p-value will not change significantly. The p-value will decrease towards zero, if significant Fourier coefficients are removed; this defines the upper limit of τ. It is recommended to study the dependence of several statistical quantities on the value of the parameter τ in repeated solutions over the acceptable range of τ-values.

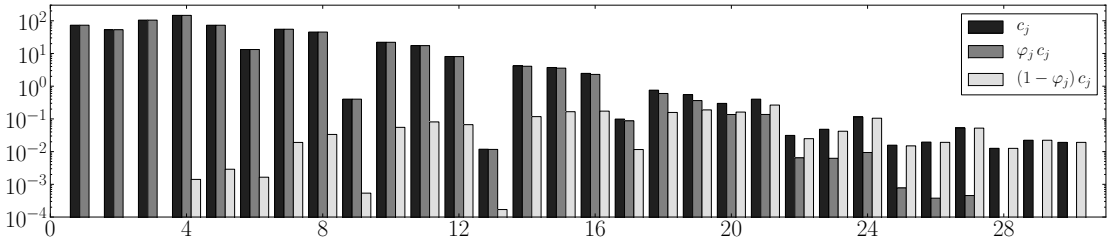


Fig. 5: Fourier coefficients c_j, with filter factors φ_j

3.5 Example: a steeply falling distributions

An example for a difficult unfolding problem is the measurement of the inclusive jet production cross section as a function of the transverse momentum p_{T} in collisions at very high energy, e.g., Reference [6]. The distribution is steeply falling. The transverse momentum p_{T}, as measured in the calorimeter, is systematically underestimated; the bias and the accuracy of the measurement can be determined in a MC simulation. In the publication [6] bin-by-bin correction (see Section 4) is applied, which is essentially

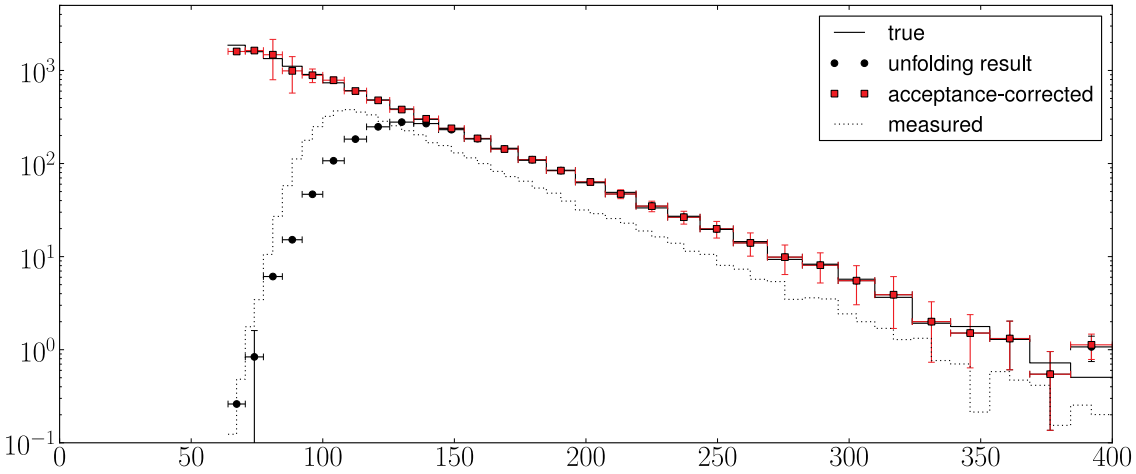


Fig. 6: Unfolding of a steeply falling distribution

only an acceptance correction, unable to correct for a bias; a bias correction is done in a separate step before.

In a simple MC simulation a problem with similar properties is solved by true unfolding according to the method of Section 3.3. Experimental conditions are assumed in analogy to the publication [6]. A pure exponential distribution is assumed with a systematic bias of the measured p_{T}-value to smaller values up to 10 %, and a Gaussian smearing with a relative standard deviation of $\sigma(p_{\mathrm{T}})/p_{\mathrm{T}} = 100\%/\sqrt{p_{\mathrm{T}}\text{ in GeV}/c}$. In addition a trigger acceptance with a rapid decrease below 100 GeV/c is assumed. Because the p_{T}-distribution at low values of p_{T} is unmeasurable, the measured and unfolded p_{T}-range is restricted to 64 to 400 GeV/c, assuming a realistic model function, with a separate acceptance correction after unfolding. The unfolding is performed in the transformed variable $q_{\mathrm{T}} = \sqrt{p_{\mathrm{T}}}$, which has a constant standard deviation $\sigma(q_{\mathrm{T}}) = 0.5$, with a back-transformation to p_{T} after unfolding, resulting in a bin-width increasing with p_{T}. The Fourier coefficients without and with filter factor are shown in Figure 5. The change of the coefficients is always less than the statistical error 1 – thus essentially no bias is introduced. The true, measured and unfolded distribution is shown in Figure 6; below 75 GeV/c the errors are larger than the cross section value.

3.6 Presentation of the regularization result

The result of regularized unfolding is an n-vector $\boldsymbol{x}$, representing the "true" function $f(t)$, together with a covariance matrix $\boldsymbol{V}_x$. In general the covariance matrix is singular with rank $k < n$ (k = number of non-zero eigenvalues after diagonalization of $\boldsymbol{V}_x$). The n bin contents originate from a small number k of effective parameters, and there will be large *positive bin-to-bin correlations*, which give the plot of the unfolded data a very *smooth* appearance, as illustrated in Figure 7 on the left (taken from Reference [4]). The plot with error bars given by the diagonal elements of $\boldsymbol{V}_x$ may be difficult to interprete and to compare with predictions; in principle the (inverse) covariance matrix has to be used for a χ^2 calculation, but this is not possible because of the rank defect. A fit of a parametrization is of course possible with the original data, as described in Section 1.3. The effective number k of degrees of freedom can be estimated by the sum $n_{\mathrm{df}} \approx \sum_j \varphi_j$ of the filter factors [3]. A method to avoid the singular-matrix problem is to present the unfolding result with only n_{df} data points. Combining four positively correlated data points to one will reduce the error by less than a factor $1/2$. This is illustrated in Figure 7 on the right, where the 40 data points are reduced to *almost uncorrelated* 10 bins, because of $n_{\mathrm{df}} \approx 10$, showing the *true information content* of the unfolded data.

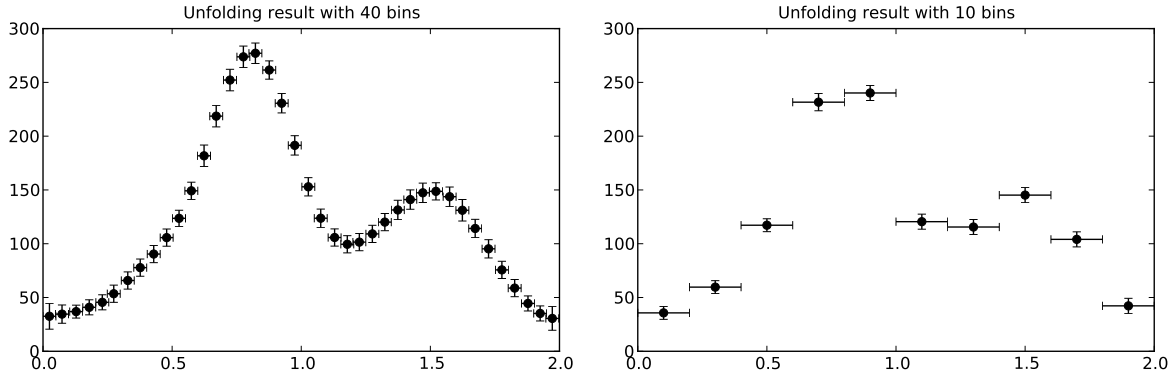


Fig. 7: Unfolding result (data from Ref. [4]) with 40 and with 10 data points

4 Iterative unfolding

Direct matrix methods (i.e. non-iterative methods) like the SVD cannot be used for problems with very large dimension parameters m and n. In those cases iterative methods for unfolding are used where the often sparse response matrix $\boldsymbol{A}$ appears only in products, for example Landweber iteration. These iterative methods are characterized by an implicit regularization, where contributions corresponding to small singular values will have very slow convergence or no convergence at all. Starting from some initial assumption $\boldsymbol{x}^{[0]}$ the first iterations show substantial improvement, but the convergence becomes then rather slow and after a very large number of iterations often a solution with large noise components similar to the naive least squares solution is obtained. This behaviour is called *semi-convergence*. In practice the iteration is stopped early; the number of iterations is the regularization parameter [1, 2]. An objective criterion for stopping the iteration is not known.

Iterative methods are rather popular in particle physics although the number of parameters is rather small and there will be neither cpu-time nor memory-space problems for direct matrix methods. If iterative methods are used, usually an attempt is made, by iterative tuning with reweighting, to perform the MC simulation already with the *correct* input distribution $f(t)^{\text{model}}$, i.e., that distribution that *on average* gives a reasonable description of the observed distributions $\boldsymbol{y}$. In these methods an x-dependent unfolding matrix $\boldsymbol{M}_x$ is iteratively improved and applied to the data $\boldsymbol{y}$ to give an improved estimate $\boldsymbol{x}^{[k+1]} = \boldsymbol{M}_x^{[k]}\ \boldsymbol{y}$. Usually the *unfolding* matrix $\boldsymbol{M}_x$ in iterative methods has only positive elements (and $\boldsymbol{\Xi} = \boldsymbol{M}_x\boldsymbol{A} \neq \boldsymbol{I}$). In the *bin-by-bin correction factor method* the matrix is diagonal with elements $(\boldsymbol{M}_x)_{ii} = x_i^{\text{mc}}/y_i^{\text{mc}}$, determined from a tuned MC simulation. The methods provide a reasonable solution, often however with large but unknown positive correlations between the data points, which is equivalent to a strong smoothing. Because no matrix like the effective regularized inverse $\boldsymbol{A}^{\#}$ is available, no prescription for a direct covariance matrix calculation exists. Estimates of the covariance matrix require e.g., Monte Carlo methods.

Acknowledgement

I would like to thank the organizers of the unfolding workshop within the PHYSTAT2011 meeting, especially Louis Lyons. For a careful reading of the manuscript, valuable comments and detailed suggestions I thank Louis Lyons and Gero Flucke.

References

[1] Per Christian Hansen, *Rank-deficient and Discrete Ill-Posed Problems: Numerical Aspects of Linear Inversion*, SIAM monographs on mathematical modeling and computation, Philadelphia, 1997.

[2] Per Christian Hansen, *Discrete Inverse Problems – Insight and Algorithms*, SIAM Fundamentals of algorithm series, Philadelphia, 2010.

[3] Volker Blobel, *Unfolding methods in high energy physics experiments*, Report DESY 84-118, 1984 (also in Proceedings of the 1984 CERN School of Computing, CERN 85-09, pp. 88-127; see also `http://www.desy.de/~blobel/`).

[4] Andreas Höcker and Vakhtang Kartvelishvili, *SVD approach to data unfolding*, Nucl. Instrum. Methods Phys. Res. A 372, 469 – 481, 1996.

[5] N.D. Gagunashvili, *Parametric fitting of data obtained from detectors with finite resolution and limited acceptance*, Nucl. Instrum. Methods Phys. Res. A 635, 86 – 91, 2011.

[6] A. Abulencia et al., *Measurement of the inclusive jet cross section using the k_T algorithm in $p\bar{p}$ collisions at $\sqrt{s} = 1.96$ TeV with the CDF II detector*, Phys. Rev. D 75, 2007.

Regularization and error assignment to unfolded distributions

Günter Zech
Universität Siegen, Germany

Abstract
The commonly used approach to present unfolded data only in graphical form with the diagonal error depending on the regularization strength is unsatisfactory. It does not permit the adjustment of parameters of theories, the exclusion of theories that are admitted by the observed data and does not allow the combination of data from different experiments. We propose fixing the regularization strength by a p-value criterion, indicating the experimental uncertainties independent of the regularization and publishing the unfolded data in addition without regularization. These considerations are illustrated with three different unfolding and smoothing approaches applied to a toy example.

1 Introduction and general considerations

Unfolding is a difficult mathematical problem, because independent of the amount of data the solution of the Fredholm equation $f'(x') = \int_{-\infty}^{\infty} t(x,x')f(x)dx$ does not lead to a stable solution; here $f(x)$ is the function of interest, $t(x,x')$ is the response- or smearing function and $f'(x')$ is the known distorted function. Therefore regularization methods have been developed.

In particle physics, distributions are usually presented in the form of histograms. Histogramming is a first regularization step. With a wide enough binning, unfolding is reduced to a simple inference problem where the parameters, i.e. the content of the bins, have to be estimated. These parameters and their errors can be determined by a standard fitting procedure. As is common practice in parameter fitting, the compatibility with the data can be expressed by a p-value. So far, the situation is rather clear and not controversial. The regularization by binning has the nice property that the interpretation of the unfolded distribution is straight forward, the point estimates and the error matrix fully document the result, there are no hidden parameters and all true distributions that are compatible with the observed data are admitted by the result.

Problems arise as soon as we choose a binning that is narrow compared to the smearing and try to represent unfolding results graphically. There are usually strong fluctuations between adjacent bins, these are then negatively correlated and the non-diagonal error matrix elements are huge which makes a graphical representation unreadable. But while the point estimates are badly known, the error limits are rather precisely derived from the data. To avoid the unpleasant oscillations, usually a second regularization step is introduced. Contrary to the implicit regularization by binning, here the kind of smoothing and its strength are hidden and their effect is difficult to assess from the published data. Furthermore, the explicit regularization introduces constraints that eliminate high frequency contributions, reduce the errors assigned to the histogram bins and thus exclude solutions that are admitted by the observed data. (We call the errors obtained in a regularized fit nominal errors to distinguish them from the errors defined by the measurement alone.) Theories that are compatible with the data but where their distributions contains narrow structures may be rejected.

As has been discussed by Blobel at this conference, for a given unfolding problem, an effective number of degrees of freedom can be estimated. This is the number of independent significant parameters of the unfolded distribution, i.e. the minimum number of parameters that is required to describe the observed data within their errors. This number depends on the width of the smearing function and on the available statistics. Thus a possible and also sensible solution would be to eliminate the second smoothing step and to publish data with a number of rather wide bins, a number not much larger than the

effective degrees of freedom. The correlations would be relatively mild and could be documented with a correlation matrix.

One common objection to wide bins is that choosing wide bins for the estimated distribution introduces a strong dependence of the smearing matrix on the input distribution used in the Monte Carlo simulation. However, this dependence can be avoided, by combining bins of the unfolded histogram. Another reason for choosing histograms with many bins is that they look much more impressive than a crudely binned histogram (see Fig. 7 in Blobel's contribution) and that they indicate better the anticipated shape of the true distribution. More scientific arguments for a not too wide binning are the following: i) When we increase the statistics by combining the data of different experiments, the spectral resolution should increase but can only be taken advantage of if the bin size is not too large. ii) Normally we choose bins of equal size. Then the minimum number of bins may not be adequate to describe the functional dependence.

All problems can be avoided if a theoretical prediction of the true distribution is available. Then we can fold the theoretical distribution and compare it to the observed data. We do not have to construct a response matrix; binning is only required in x' and not in x, the distorted distribution is simulated and compared to the data. Unknown parameters can be estimated by re-weighting the simulated events [1]. It does not make sense to pass through a histogram with many parameters to finally determine a few parameters of interest. More important, the direct fit, for instance of the amplitude of a narrow peak with given width superposed to a uniform background, would produce a correct result, while the true distribution might be incompatible with an unfolded distribution where high frequencies are filtered out.

In the situation where no generally accepted theoretical description is available, a sensible solution to the stated unfolding problems is to separate the graphical representation of the result from its documentation. We will first sketch the way the data can be documented and then turn to the explicit regularization. We discuss how to fix the regularization strength, illustrate the problems with a simple example and three different unfolding approaches and end with a summary and recommendations.

2 Documentation of the result

To avoid the exclusion of distributions admitted by the data and to permit the combination and comparison of measurements of different experiments the data have to be published without explicit regularization.

There are at least three possibilities to document the full information: i) Unfold without regularization and provide an error matrix. The number of bins has to be low, not much higher than the effective number of degrees of freedom. ii) Publish the raw data together with the response matrix. iii) Publish the eigenvectors of the matrix C (see Sect. 4.1), their weights and the uncorrelated errors of the weights.

There might be statistical or technical difficulties applying any of these approaches which cannot be discussed in a short paper. It should just be mentioned that proposal i) requires enough data to approximate the error distributions of the unfolded histogram bins by Gaussians. Method ii) leaves all the work to the user of the data. To follow point iii), the correlations between the contents of the histogram bins have to be eliminated by diagonalization.

3 Fixing the regularization strength

In addition to the documentation of the data in form of tables we want to illustrate the result of our experiments graphically and in most cases have to add an explicit smoothing step. There is no optimal regularization algorithm, and smoothing is partially subjective. The only obvious requirement is that the smoothing step does not destroy the compatibility of the result with the observed data. A very attractive method favored by statisticians is to suppress small eigenvalue contributions of the least square matrix or equivalently in the SVD decomposition (see Blobel's and Kartvelishvili's contributions to this conference) but there are other methods that work equally well. The performance of a regularization

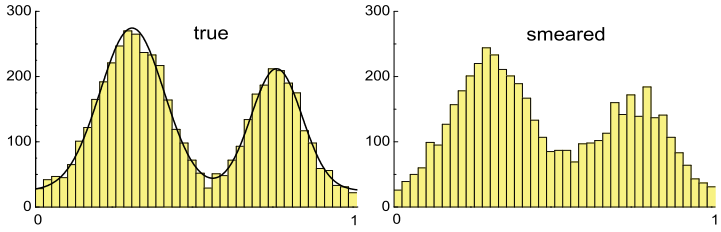


Fig. 1: Distributions used in the unfolding example. Here and in all following graphs the axes correspond to measured variable $0 < x < 1$ (horizontal) and to the number of entries (vertical).

method depends to a certain extent on the specific problem to be solved.

There are several methods to fix the regularization strength, i.e. from the kink of the L-curve [2], from minimum norm or vanishing global correlation. These are mathematical concepts. From a physicist's point of view, the essential criterion is the compatibility of the unfolded histogram with the observed data which can be measured with a p-value derived from the χ^2 statistic. What is important is not the absolute value of χ^2 of the fit of the bin contents of the unfolded histogram but its change $\Delta\chi^2$ caused by the regularization. The p-value is $p = \int_{\Delta\chi^2}^{\infty} u_N(x)dx$ where u_N is the χ^2 distribution for N degrees of freedom, i.e. N bins of the unfolded histogram. Assuming ideal Gaussian errors, this means that the regularized solution is located at the border of a $1-p$ confidence ellipsoid. When we change the regularization strength, rather independent on how we define it, the p-value initially remains close to one as long as we include high frequency contributions that have little effect on the unfolded data. As soon as we start to affect the unfolded histogram, by definition the p-value drops dramatically.

We propose to fix p to 90% or even to a higher value. This means that the parameter set of the regularized histogram is within a 10% confidence interval of the minimum χ^2 point which corresponds to the undistorted measurement. There is no necessity to suppress all fluctuations, we accept them also in standard measurements that are not affected by resolution effects. The proposed p-value is arbitrary, but it should be large, as we want to cut only insignificant features of the data. Its choice is motivated by experience.

It has been argued [3] that $\Delta\chi^2$ should not be referred to N but to the difference between N and the effective degrees of freedom N_{eff}, i.e. $\Delta\chi^2 \approx N - N_{eff}$. ($N_{eff}$ is the number of independent parameters that are necessary to describe the data or the effective rank of the response matrix, see Blobel's contribution.) The additional $N - N_{eff}$ parameters are expected to contribute 1 to χ^2 each. Indeed, for the specific case $N = N_{eff}$ in the framework of SVD one would not regularize at all. On the other hand smoothing N bins of an arbitrary distribution one would usually allow for a change of χ^2 proportional to N. Because of the sharp kink in the distribution of χ^2 as a function of the regularization strength, the two different methods usually will not lead to very different results. Further studies should clarify this issue.

It is debatable what the best value for the number of degrees of freedom (NDF) should be for converting $\Delta\chi^2$ to a p-value. However, since the cut on the p-value is arbitrary, the choice for NDF is not crucial.

It would be very helpful if the community could agree on a common scheme. This would make the comparison of different methods and of unfolded results much easier than it is now.

4 Toy example and three unfolding approaches

These abstract considerations are now illustrated with a simple toy example and 3 different regularization procedures. The true distribution is a superposition of two Gaussians and a uniform distribution: 2500

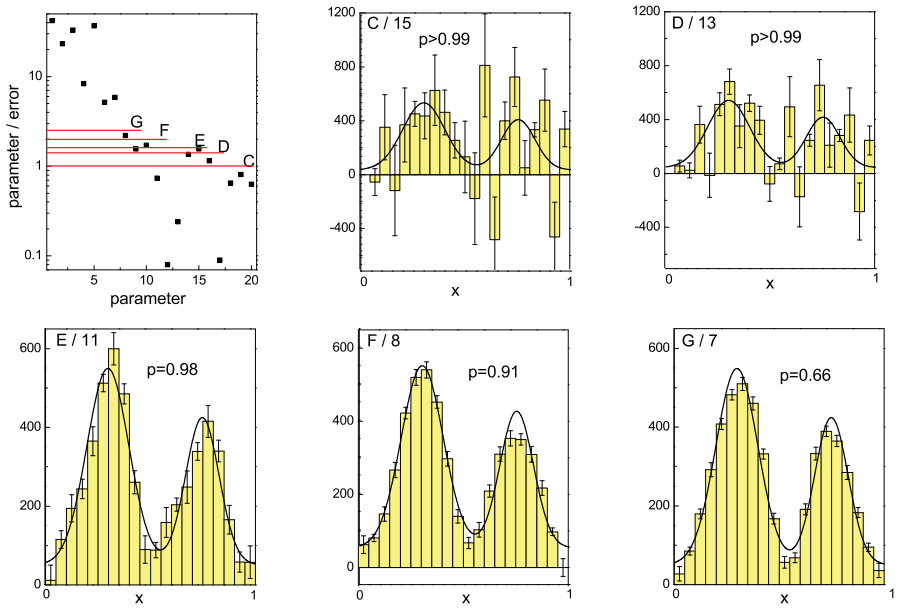


Fig. 2: Significance s of parameters (parameter / error) ordered according to decreasing eigenvalues (top left) and unfolded distribution for different cuts in the significance. The number of contributing eigenvectors is indicated in the top left corner of the plots.

events $N(0.3; 0.1)$, 1500 events $N(0.75; 0.08)$, 1000 events uniform in the range 0 to 1. The resolution function is again a Gaussian, $N(0; 0.07)$. To demonstrate the problems, a relatively large smearing was chosen. The effective number of degrees of freedom is around 8. In Fig. 1 the corresponding true distribution (curve), the original and the smeared histograms from the Monte Carlo simulation are shown. For the unfolding 20 or 10 bins for the unfolded histogram and 40 bins for the observed data were chosen. Our notation is the following: The bin content of the histogram corresponding to the true distribution is represented by the vector θ, the histogram of the observed data by $\hat{d}$, and the response matrix is A. The expectation value d is related to θ by $d = A\theta$.

4.1 Method 1: Truncation of the eigenvalue sequence

The square matrix $C = A^T V^{-1} A$, with V the error matrix of the observed data vector $\hat{d}$, is decomposed into eigenvectors u_i. The true distribution vector θ can be expressed as a sum $\theta = \sum_i c_i u_i$, where the coefficients c_i of the eigenvectors are to be determined. The coefficients c_i are uncorrelated. With decreasing eigenvalues the components of the eigenvectorss oscillate more and more but the coefficients become less significant and their contributions can be eliminated. In theory this is very attractive but in practice the situation is not always simple, as is shown in Fig. 2. The upper left hand plot shows the significance $s = |c_i/\delta c_i|$ for the parameters ordered by eigenvalue. We realize that the parameter with the lowest eigenvalue is not necessarily the one with the smallest significance. The significance of each component is proportional to the square root of the number of events. With increasing statistics more components would become significant and correspondingly the spectral resolution would increase. However, since the sequence of ordered eigenvalues decreases rapidly, statistics in most cases has only a small effect on the spectral resolution. The dominant effect comes from the width of the response function. In the literature cutting or damping low eigenvalue is proposed instead of removing low significant

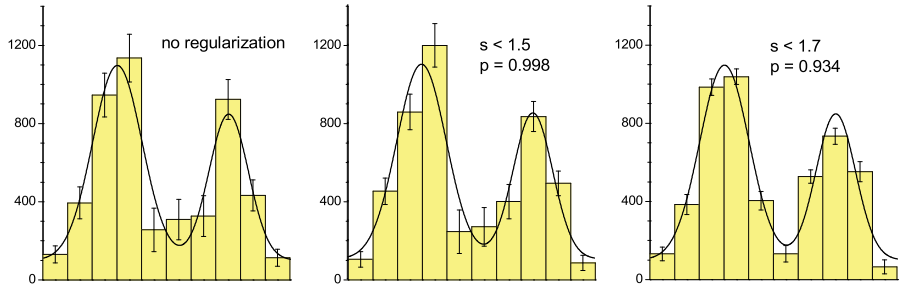


Fig. 3: Unfolded distributions for different cuts in significance.

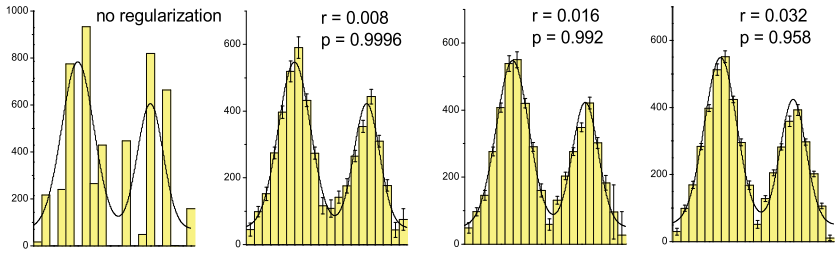


Fig. 4: Distributions unfolded with a Poisson likelihood fit and a curvature penalty.

components, but then the shape of the observed distribution which may favor certain eigenvectors is not taken into account.

In Fig. 2 the unfolded histograms for different cuts in significance are displayed and compared to the true distribution (curves). How should we cut? In principle we should choose the cut without looking at the unfolding results. Cutting contributions that are compatible with zero within their error, $s < 1$, produces unacceptable results. We have to find a compromise between loss in information and smoothing. Retaining between 8 and 11 components seems to be reasonable. The plot labeled E would be the preferred one, as it does not show excessive fluctuations and at the same time corresponds to a large p-value.

The regularization leads to a smooth unfolded distribution but, as expected, the nominal diagonal errors decrease strongly with increasing regularization strength. Because of this dependency, the errors, which are more or less arbitrary, are unreliable indicators of the precision of the result. We show them, to highlight the problem.

The situation becomes more reasonable when we turn to a smaller number of bins as illustrated in the Fig. 3. With 10 bins, which is about the effective number of degrees of freedom, negative bins disappear. The calculated errors become more reasonable because the correlations are smaller. Here we could even renounce explicit smoothing and document the errors with a simple error matrix.

The results presented have been obtained with a least square fit (LSF) and the simple matrix formalism. It would be better to apply a Poisson maximum likelihood fit; negative entries are then suppressed, but the qualitative features and the conclusions would remain the same. Also a smooth cutoff of the low eigenvalue contributions would not alter the result by much.

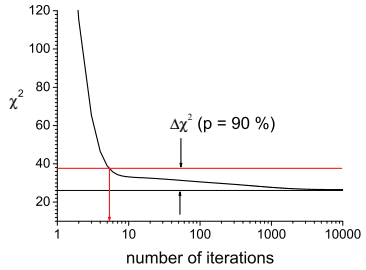


Fig. 5: χ^2 as a function of the number of iterations. Stopping at 5 iterations corresponds to a p-value of about 0.9.

4.2 Method 2: Poisson maximum likelihood fit with penalty regularization

Here we fit the contents of the true histogram by maximizing the log-likelihood

$$\ln L = \ln L_{stat} - R$$

$$\ln L_{stat} = \sum_{i=1}^{M} \hat{d}_i \ln \left(\sum_{j=1}^{N} A_{ij}\theta_j \right) - \sum_{j=1}^{N} A_{ij}\theta_j$$

$$R = r \sum (2\theta_i - \theta_{i-1} - \theta_{i+1})^2$$

which consists of the statistical term and a regularization term which is related to the local curvature. The χ^2 statistic is evaluated for different regularization constants r using Gaussian errors.

Fig. 4 shows some results of this method. Negative entries are excluded but the dependence of the nominal errors on the regularization strength of course remains. The p-values are indicated. A value around 90% corresponds to a reasonable smoothing. As in the general case, we do not know the true distribution, it is difficult to fix the constant r if it cannot be done with a p-value.

Curvature regularization which favors linear distributions is popular, but other penalty functions can be chosen. For a nearly exponential distribution, deviations from linearity of $\log \theta$ could be penalized. The application of penalty functions smooths the distribution even when the smearing matrix is diagonal.

4.3 Method 3: Iterative unfolding, stopping the iteration

Iterative unfolding has become popular because there exists a simple to use program by D'Agostini. Iterative unfolding is an old concept [4]. D'Agostini has re-invented it and interpreted it in the context of Bayesian statistics [5]. The iteration procedure can however also be viewed as a simple mathematical algorithm to invert the matrix equation $d = A\theta$. The method is explained in the Appendix. It is so simple that it can be coded in a few lines. The convergence is very fast at the beginning and suddenly slows down (see Fig. 5). The computation is fast; 10000 iterations were obtained in 1 minute with a standard laptop computer. They produced results almost indistinguishable from an least square or maximum likelihood fit (see Fig. 6). Quite good agreement of the result with the data is usually obtained after a few steps. As with other smoothing methods it is reasonable to fix the regularization strength, e.g. to stop the iteration with the p-value criterion. To compute the p-value we have to determine the minimal χ^2 either by a separate fit or by estimating its value from a large number of iterations.

The nominal errors of the unfolded distribution could in principle be obtained by error propagation, but this is extremely tedious as the analytic relation due to the iteration sequence is extremely complicated. Anyway, the nominal errors are rather useless.

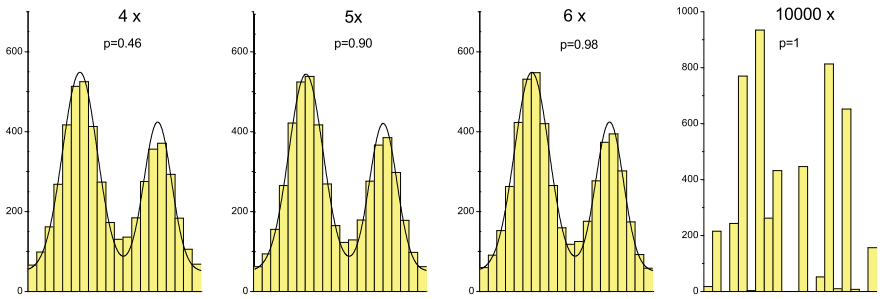


Fig. 6: Iteratively unfolded distribution for different numbers of iterations.

5 Error assignment to the graphical representation

The fitting methods produce error estimates automatically. For other methods the uncertainties can be obtained by the usual error propagation, but these nominal errors depend on the strength of the regularization which on the other hand is unrelated to the statistical accuracy of the data. A strongly regularized distribution may exhibit even smaller diagonal errors than the distribution before the convolution, i.e. smaller than the square root of the number of entries. This is unacceptable, and misleading (see the fake bump in the Alice experiment shown in the contribution to this conference by Gross-Oetringhaus) as we loose information by smearing. We should present errors which are useful in that they indicate whether functions are compatible with the data or not and do not depend on data manipulations.

A sensible graphical presentation of the unfolding result where the values but not the errors depend on the regularization is the following [1]: For each unfolded bin j we attribute a relative statistical error $\delta\theta_j/\theta_j$ equal to one over the square root of the number of observed events associated to that bin. This is equal to one over square root of θ_j if there are no acceptance corrections. A horizontal bar indicates the experimental resolution [1].

6 Summary and recommendations

Experimental results should be published such that 1. the data can be compared to and combined with those of other experiments in such a way that the combined result exhibits smaller statistical uncertainties and superior resolution in the smeared variable, 2. theoretical predictions can be tested, 3. all predictions that are compatible with the data are admitted. These requirements can only be satisfied when the experimental data are published without explicit regularization. When we compare a theory to the measurement, we should fold the theoretical prediction and compare it to the raw data.

A graphical representation with wide bins, such that oscillations are damped, is recommended. In any case the number of bins should be less than twice the effective number of degrees of freedom. If an explicit regularization is applied, we have to be aware that smoothing introduces constraints and modifies the experimental information. We propose to fix the regularization strength using a p-value criterion, which guarantees that the regularized distribution is compatible with the observed data. The errors that are assigned to the unfolded histogram have to be independent of the regularization.

All smoothing approaches that fulfill these requirements are acceptable, however, efficient methods reduce essentially the fluctuations between adjacent bins. All three approaches studied above in a simple example produce sensible and very similar results. Regularization with a penalty term is especially transparent. Iterative unfolding is simple and the smoothing prejudice is included in a flexible way. The truncation of singular value components is mathematically attractive and singular value decomposition (SVD) provides insight into the problem of unfolding.

Appendix: Iterative least square fit

The relation $\hat{d} = A\hat{\theta}$ can be solved iteratively, provided the response matrix A is positive definite. The idea behind the iteration algorithm is the following [1]. Starting with a preliminary guess of $\hat{\theta}^{(0)}$, the corresponding prediction for the observed distribution $d^{(0)}$ is computed. It is compared to $\hat{d}$ and for a bin i the ratio $\hat{d}_i/d_i^{(0)}$ is formed which ideally should be equal to one. To improve the agreement, all true components are changed in proportion to their contribution $A_{ij}\theta_j^{(0)}$ to $d_i^{(0)}$ This procedure when iterated corresponds to the following equations: The prediction $d^{(k)}$ of the iteration k is obtained in a *folding step* from the true vector $\theta^{(k)}$:

$$d_i^{(k)} = \sum_j A_{ij}\theta_j^{(k)} \ .$$

In an *unfolding step*, the components $A_{ij}\theta_j^{(k)}$ are scaled with $\hat{d}_i/d_i^{(k)}$ and added up into the bin j of the true distribution from which it originated:

$$\theta_j^{(k+1)} = \sum_i A_{ij}\theta_j^{(k)} \frac{\hat{d}_i}{d_i^{(k)}} /\varepsilon_j \ .$$

Dividing by the efficiency ε_j corrects for acceptance losses.Empirically, it has been shown that with increasing number of iterations, the result converges to the maximum likelihood fit result for Poisson distributed errors [4].

In D'Agostini's Bayesian approach (see, e.g., Bierwagen, these Proceedings), the same iteration sequence is applied, however, between each iteration the unfolded distribution is smoothed by a polynomial fit. The details of the smoothing step are left to the user. Convergence is implied by the method. The degree of smoothing depends on the intermediate smoothing algorithm. Furthermore, prior densities are introduced for the parameters of the multinomial and the Poisson distributions that are used in the evaluation of the uncertainties. It is not obvious that the priors have a noticeable effect. The reliability of the error estimates is unclear.

Acknowledgment

I would like to thank the organizers and especially Louis Lyons for the invitation to this conference with an interesting program in a pleasant atmosphere. I would like to thank Gerhard Bohm, my referee Gero Flucke and Louis Lyons for a careful reading of the manuscript and for valuable comments.

References

[1] G. Bohm and G. Zech, *Introduction to Statistics and Data Analysis for Physicists*, Verlag Deutsches Elektronen-Synchrotron (2010), http://www-library.desy.de/elbook.html.

[2] P.C. Hansen, *Discrete Inverse Problems – Insight and Algorithms*, SIAM Fundamentals of Algorithms series, Philadelphia (2010).

[3] L. Lyons, private communication.

[4] L.B. Lucy, *An iterative technique for the rectification of observed distributions* Astronomical Journal 79 (6) (1974) 745; Y. Vardi, L. A. Shepp and L. Kaufmann, *A statistical model for positron emission tomography*, J. Am. Stat. Assoc. (1985) 8; 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 (1987) 371.

[5] G. D'Agostini, *A multidimensional unfolding method based on Bayes' theorem*, Nucl. Instr. and Meth. A 362 (1995) 48, G. D'Agostini, *Improved iterative Bayesian unfolding*, arXiv:1010.632v1 (2010).

Bayesian Unfolding

Katharina Bierwagen, Ulla Blumenschein, Arnulf Quadt
2nd Institute of Physics, Georg-August-University Göttingen

Abstract
I give a short introduction to Bayesian Unfolding and describe my work on the improvement of the uncertainty calculation for this method.

1 Introduction

Bayesian Unfolding has been used since 1994 and was introduced by G. D'Agostini (see Ref. [1]). For our application area the method shows problems in the evaluation of the uncertainties. Therefore, performance studies using ensemble testing are shown for the Bayesian Unfolding Method. An improved uncertainty calculation has been developed to providing a better error calculation.

2 Bayes Method

The procedure of Bayesian Unfolding can be explained using a picture of causes C and effects E, in which causes correspond to the true values before smearing and effects to the values after smearing. Each cause can produce different effects, but for a given effect, as is the case in a measurement, the exact cause is not known. However, the probability for an effect produced from a defined cause $P(E_j|C_i)$ can be estimated assuming some knowledge about the migration, efficiency and resolution. This is usually achieved by using Monte Carlo. Now the goal is to estimate the probability $P(C_i|E_j)$ that different causes C_i were responsible for the observed effect E_j. A simple inversion cannot be used to solve this problem, but Bayes theorem yields a solution.

$$P(C_i|E_j) = \frac{P(E_j|C_i) \cdot P_0(C_i)}{\sum_{l=1}^{n_C} P(E_j|C_l) \cdot P_0(C_l)}, \tag{1}$$

$$\hat{n}(C_i) = \frac{1}{\epsilon_i} \sum_{j=1}^{n_E} n(E_j) \cdot P(C_i|E_j) \qquad \epsilon_i \neq 0, \tag{2}$$

with $\hat{n}(C_i)$ the expected number of events in the cause bin i, $n(E_j)$ the number of events in the effect bin j, $P_0(C_i)$ the initial probabilities and ϵ_i the efficiency that the cause i has an effect. This formula can be rewritten in terms of the unfolding matrix M

$$\hat{n}(C_i) = \sum_{j=1}^{n_E} M_{ij} \cdot n(E_j), \tag{3}$$

$$M_{ij} = \frac{P(E_j|C_i) \cdot P_0(C_i)}{[\sum_{l=1}^{n_E} P(E_l|C_i)] \cdot [\sum_{l=1}^{n_C} P(E_j|C_l) \cdot P_0(C_l)]}, \tag{4}$$

which is clearly not equal to the inverse of the migration matrix. For the calculation of the covariance matrix of $\hat{n}(C_i)$ two different sources are taken into account, an uncertainty on the distribution of the effects $n(E_j)$

$$V_{kl}(\underline{n}(E)) = \sum_{j=1}^{n_E} M_{kj} \cdot M_{lj} \cdot n(E_j) \cdot \left(1 - \frac{n(E_j)}{\hat{N}_{true}}\right) - \sum_{\substack{i,j=1 \\ i \neq j}}^{n_E} M_{ki} \cdot M_{lj} \cdot \frac{n(E_i) \cdot n(E_j)}{\hat{N}_{true}}, \tag{5}$$

with the true number of events $\hat{N}_{true}$ and an uncertainty on the migration probabilities $P(E_j|C_i)$

$$V_{kl}(M) = \sum_{i,j=1}^{n_E} n(E_i) \cdot n(E_j) \cdot Cov(M_{ki}M_{lj}). \tag{6}$$

The total uncertainty is calculated from the sum of both covariance matrices

$$V_{kl} = V_{kl}(\underline{n}(E)) + V_{kl}(M). \tag{7}$$

In the following this method with its uncertainties is checked using a toy Monte Carlo.

2.1 The Toy Monte Carlo

This section describes a simple toy Monte Carlo, which is used in the following to check the method.

Three different migration matrices with different migration (large migration (S_1), medium migration (S_2) and low migration (S_3)) for 3 bins without efficiency losses are defined

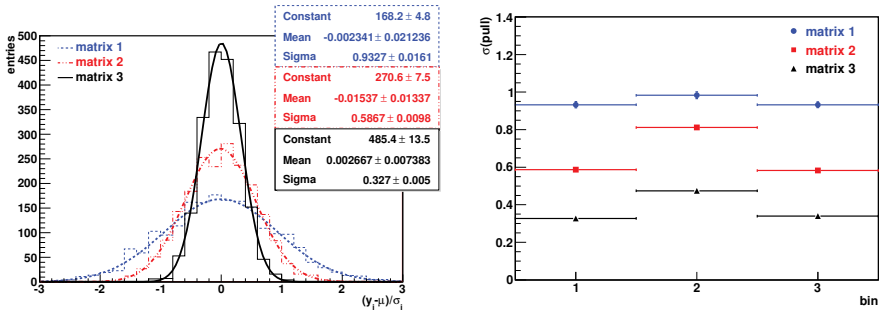
$$S_1 = \begin{pmatrix} 0 & 0.1 & 0.1 \\ 0.2 & 0.3 & 0.5 \\ 0.8 & 0.6 & 0.4 \end{pmatrix}, S_2 = \begin{pmatrix} 0 & 0.1 & 0.8 \\ 0.2 & 0.8 & 0.2 \\ 0.8 & 0.1 & 0 \end{pmatrix}, S_3 = \begin{pmatrix} 0 & 0.025 & 0.95 \\ 0.05 & 0.95 & 0.05 \\ 0.95 & 0.025 & 0 \end{pmatrix}.$$

For the unfolding, two different sources of uncertainties are present: a finite amount of data and a finite number of events for the creation of the migration matrix. Therefore, three different cases were considered: create randomly 2000 test distributions to simulate the finite amount of data events, create randomly 2000 migration matrices from fixed probabilities on top of the test distributions to simulate statistical fluctuations in the migration matrix due to a finite statistics in Monte Carlo and finally create randomly 2000 uniformly distributed true distributions in addition.

2.2 Performance checks

For the performance checks of the method, a C++ implementation of Ref. [1] is used.

In order to quantify if the absolute values and the uncertainties are correct ensemble testing is used. The width of the Gaussian distribution from ensemble testing is compared to the uncertainties calculated by the program using pull distributions: pull $= (y_i - \mu)/\sigma_i$ with the unfolded values y_i, the truth value μ and the calculated uncertainties σ_i. If the uncertainties are correctly estimated, the width of the pull distributions is expected to be compatible with 1. These tests show that the mean values are well described as expected, but the uncertainties are too large.



(a) (b)

Fig. 1: Comparison between the deviations from ensemble tests and the uncertainties calculated by the program for 3 different migration matrices.

Fig. 1 shows the pull distributions for the three different migration matrices. With decreasing migration in the migration matrix, the pull distributions become broader. If the uncertainty calculation is correct and for a correct treatment of the migration effect, the pull distributions are expected to be standard Gaussians (mean zero and unit variance). As shown, less migration leads to an overestimate of the uncertainties, so the migration effect is not treated correctly in the uncertainty calculation. This problem seems to come from the fact that the program assumes a multinomial[1] distribution for the data, but each bin is multinomially distributed and the sum of multinomial distributions is only a multinomial distribution if all distributions are equal. In order to fulfil this requirement, the columns of the migration matrix have to be equal to get the correct estimate for the uncertainty from the program, which is not the typical case in data analysis.

Due to the fact, that the data sample is a sum of multinomial distributions, the formula for the calculation of the covariance matrix V_{kl} for the data sample $\underline{n}(E)$ is changed from Eqn. 5 to

$$V_{kl}(\underline{n}(E) = \sum_{j=1}^{n_E} M_{kj} \cdot M_{lj} \cdot \sum_{r=1}^{n_C} \hat{n}(C_r) \cdot P(E_j|C_r) \cdot (1 - P(E_j|C_r)) - \sum_{\substack{i,j=1 \\ i \neq j}}^{n_E} M_{ki} \cdot M_{lj} \cdot \sum_{r=1}^{n_C} \hat{n}(C_r) P(E_i|C_r) \cdot P(E_j|C_r). \quad (8)$$

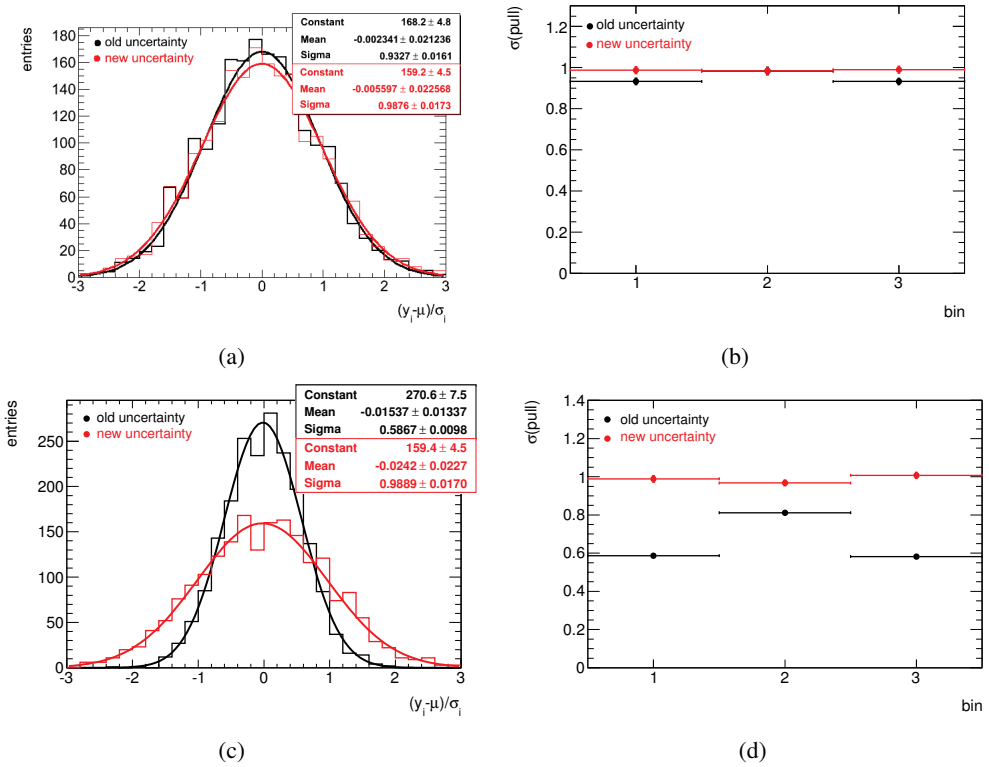
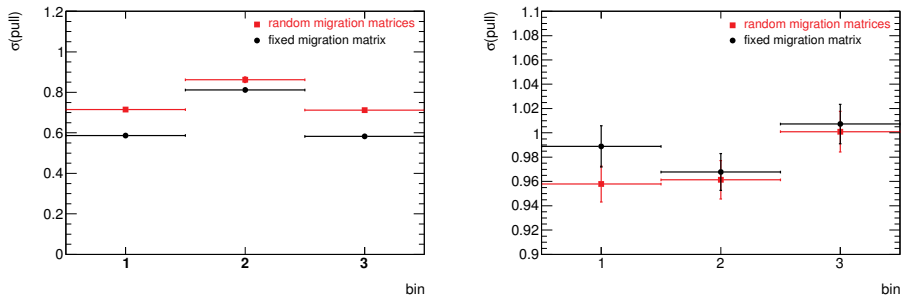


Fig. 2: Comparison of the pull distributions (a,c) and the width of the pull distributions (b,d) for the old and the new uncertainty calculation for large and medium migration.

Fig. 2 shows the comparison of the pull distributions between the old and the new uncertainty calculation for the migration matrix with large migration and with medium migration for the first bin and the width of the pull distributions for each of the three bins. As expected, for large migration,

[1] A multinomial distribution is a generalization of the binomial distribution with more than two possible outcomes.

only a small improvement due to the new uncertainty calculation is visible, whereas for the migration matrix with medium migration a clear improvement is visible. For both cases with the new uncertainty calculation the width of the pull distribution is now compatible with unity. Furthermore, the differences between the pull distributions for a randomly generated migration matrices and a fixed migration matrix vanishes using the new uncertainty calculation (Fig 3).



(a) (b)

Fig. 3: Comparison of the width of the pull distributions for the old uncertainty calculation (a) and the new uncertainty calculation (b).

The new uncertainty calculation shows a clear improvement and solves all previously mentioned problems. Meanwhile, there is an improved Bayesian Unfolding method available, which was also described by G. D'Agostini in Ref. [2] and which is based on the previous method but the uncertainties are treated differently. In this method the quantities are described by probability density functions and the error propagation is done by sampling. The next step is to compare this method with my improvements for the old method. This will be pursued in the near future.

3 Summary

This article describes performance studies using ensemble testing for Bayesian Unfolding. This study is motivated by the fact that the uncertainties for this method seem to be too large compared to the fluctuations. In order to solve this problem, an improved uncertainty calculation is presented which shows a good performance.

References

[1] G. D'Agostini. A multidimensional unfolding method based on Bayes' theorem. *Nuclear Instruments and Methods in Physics Research A*, 362:487–498, February 1995.

[2] G. D'Agostini. Improved iterative Bayesian unfolding. *ArXiv e-prints*, October 2010.

Unfolding with Singular Value Decomposition

V. Kartvelishvili
Lancaster University, United Kingdom

Abstract
An overview is given of the SVD approach to unfolding, including basic principles, error propagation and fitting applications.

1 Introduction

We will assume that an *initial* high statistics Monte Carlo sample was used to create $\hat{A}_{ij}$, the matrix simulating detector response — i.e. the probability for an event generated in the *true* bin j to be found in the *measured* bin i:

$$\sum_j \hat{A}_{ij}\, {x_j}^{\rm ini} = {b_i}^{\rm ini}, \tag{1}$$

which, by construction, is satisfied exactly. Here $b_i^{\rm ini}$ is the vector (histogram) of MC "measured" values, while $x_j^{\rm ini}$ is the vector (histogram) of MC true values.

Our aim is to determine the underlying real distribution x from a real measured distribution b, given $\hat{A}_{ij}$ and some additional information on expected properties of x, i.e. find a meaningful way of solving the simultaneous system of equations

$$\hat{A}x = b. \tag{2}$$

The problem is twofold. Firstly, the right-hand-side b is known with some precision: e.g., for purely statistical errors, the covariance matrix $B = \mathrm{diag}\{b\}$. Secondly, our knowledge of $\hat{A}$ is not perfect either, due to finite MC statistics, as well as imperfections in detector simulation. Even worse, $\hat{A}$ is almost always very close to being degenerate, so an attempt to solve the problem directly and "exactly" is unlikely to be useful.

2 Singular Value Decomposition

For a detailed description of the unfolding algorithm based on the Singular Value Decomposition of the response matrix, see Ref. [1] and references therein. The original development of the algorithm was based on the method presented in Ref. [2].

A Singular Value Decomposition (SVD) of a real $m \times n$ matrix A is its factorization of the form (see Ref. [1] and references therein)

$$A = U\, S\, V^T, \tag{3}$$

where U is an $m \times m$ orthogonal matrix ($U\,U^T = U^T U = I$), V is an $n \times n$ orthogonal matrix ($V\,V^T = V^T V = I$), while S is an $m \times n$ diagonal matrix with non-negative diagonal elements:

$$S_{ij} = 0 \ \text{ for } \ i \neq j, \quad S_{ii} \equiv s_i \geq 0. \tag{4}$$

The numbers s_i are called *singular values* of the matrix A. For example, if A itself is orthogonal, all $s_i = 1$, while if A is degenerate, at least one s_i is equal to zero. By swapping rows of U (and similarly for V), s_i can be ordered from largest to smallest.

Columns of U and V are called the left and right *singular vectors*. They define convenient orthonormal bases in their respective spaces.

With SVD, the linear system $Ax = b$ can be easily diagonalised by introducing rotated vectors z and d, and finding the exact solution looks deceptively simple:

$$\begin{aligned} U\,S\,V^T x &= b \quad \Rightarrow \quad z \equiv V^T x, \quad d \equiv U^T b, \\ s_i z_i &= d_i \quad \Rightarrow \quad z_i = \frac{d_i}{s_i} \quad \Rightarrow \quad x = Vz. \end{aligned} \tag{5}$$

However, there are at least two reasons why this determination of z_i can go horribly wrong: firstly, due to errors in b, some d_i may be poorly known, or not significant at all; secondly, some singular values s_i may be small (or even zero), thus exaggerating the contributions of poorly known coefficients.

Indeed, Fig. 1 shows an example taken from Ref. [2], and also used in Ref. [1]. It is clear that if the r.h.s. b contains random fluctuations, the exact solution x (the right plot) is essentially useless, as these fluctuations get greatly magnified by the small singular values.

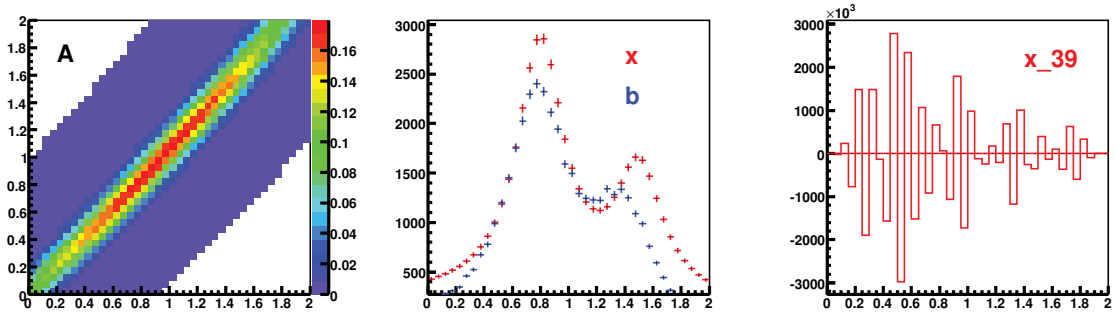


Fig. 1: An example of the response matrix A (left), the "measured" and "true" distribution (b and x, respectively, centre) and the exact solution x of the system $Ax = b$ (right), once the r.h.s. b contains random statistical fluctuations.

So, since the orthogonal matrices U and V are totally harmless, SVD allows the problem to be narrowed down to individual s_i and/or d_i.

3 Rescaling variables and equations

In general, different equations in the simultaneous system $Ax = b$ should contribute to any meaningful solution with different weights, in accordance with the precision of the respective coefficients and the r.h.s. Hence, in order to make the equations and the unknowns more uniform, some rescaling is required.

Rescaling of unknowns. If the Monte Carlo and the data are fast-varying functions ranging over many orders of magnitude, it makes sense to try and find x relative to the true distribution $x^{\rm ini}$ (i.e. relative to the MC used for creating the matrix A). This is achieved by multiplying each column of A_{ij} by $x_j^{\rm ini}$ and simultaneously defining new unknowns $w_j = x_j/x_j^{\rm ini}$. This transformation has a "side-effect" of changing A from "probability to "number-of-events" matrix. The latter is arguably more useful, as bins with higher statistics, and hence more significance, are now enhanced.

Rescaling of equations. In general, if one equation is multiplied by a factor, SVD of the matrix changes. One of the ideas of the GURU algorithm, described in Ref. [1], was to make sure that all the equations are "made equal" by checking that the error in the r.h.s. is always ± 1. In the simplest case of uncorrelated errors in the r.h.s., this is achieved by dividing each equation (i.e. each row of A_{ij} as well as b_i) by the error Δb_i.

After these manipulations, the original system $Ax = b$ has transformed into

$$\sum_j \tilde{A}_{ij} w_j = \tilde{b}_i, \tag{6}$$

where, by construction, the covariance matrix of the r.h.s. is equal to the unit matrix, and the new unknowns w_i are defined relative to the input MC distribution. I.e., $w_i = 1$ would mean that the "unfolded" x is the same as the "truth" MC used to generate the response matrix.

4 Solving equations

Now let's use SVD to solve the rescaled system:

$$\begin{aligned} \tilde{A}w = \tilde{b} \quad &\Rightarrow \quad \tilde{A} = U\,S\,V^T \quad \Rightarrow \quad U\,S\,V^T\,w = \tilde{b} \quad \Rightarrow \quad z \equiv V^T\,w, \quad d \equiv U^T\,\tilde{b}, \\ s_i z_i = d_i \quad &\Rightarrow \quad z_i = \frac{d_i}{s_i} \quad \Rightarrow \quad w = V\,z. \end{aligned} \tag{7}$$

By construction, all d_i are independent and have errors ± 1. Assuming all $s_i \neq 0$, z_i form the *exact* solution of the initial system, which is still highly unlikely to be useful.

Looking at the plots of $|d|$ and s (Fig. 2), it's easy to understand what's happening: many d_i are

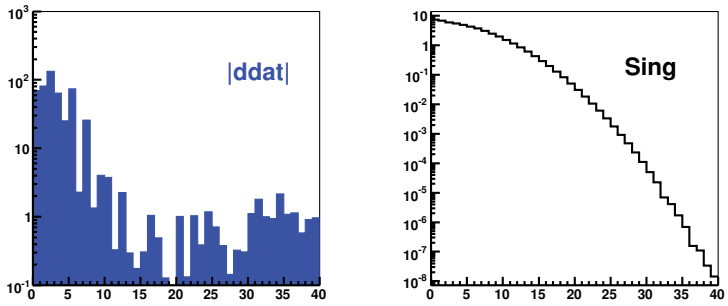


Fig. 2: The distribution of $|d_i|$ (left) and the singular values s_i (right), for the example described above. All components d_i for $i \gtrsim 10$ are essentially insignificant, as by construction all d_i have (independent) errors of ± 1.

insignificant (compatible with zero), while corresponding s_i are small. Respective z_i will be huge, but nonetheless essentially random.

Note that the exact solution of the system, Eq.. (7), is equivalent to the minimisation of the residual χ^2:

$$\chi^2 \equiv (\tilde{A}w - \tilde{b})^T (\tilde{A}w - \tilde{b}) = \min. \tag{8}$$

where (ignoring machine precision and assuming all $s_i \neq 0$) $\chi^2_{\min} = 0$. A simple way of regularisation would be the truncation of the (diagonalised) system at some $i = k$, which would make $\chi^2 = \sum_{i>k} d_i^2$.

While simple truncation is better than keeping all i, this is not a good solution, as biases are hard to control. One of the usual choices, used in high energy physics, is regularisation by adding an *a priori* requirement that the regularised solution is *smooth*. Technically, this requirement is introduced into the χ^2 minimisation condition by adding an extra term [1,2]:

$$\chi^2 \equiv (\tilde{A}w - \tilde{b})^T (\tilde{A}w - \tilde{b}) + \tau (C\,w)^T C\,w = \min. \tag{9}$$

By choosing Cw to be the second finite difference of w, so that

$$(C\,w)^T C\,w = \sum [(w_{i+1} - w_i) - (w_i - w_{i-1})]^2\,, \tag{10}$$

we find a minimum of χ^2 that keeps the "integrated square of the second derivative" of w constrained. The parameter τ essentially plays the role of the Lagrange multiplier in the new conditional minimisation problem, given by Eq. (9).

This new minimisation problem gives rise to a new, overconstrained linear system [1] :

$$\left[\begin{array}{c} \tilde{A}\,C^{-1} \\ \sqrt{\tau}\cdot I \end{array} \right] C\,w = \left[\begin{array}{c} \tilde{b} \\ 0 \end{array} \right]. \tag{11}$$

A fairly straightforward method allows the reduction of the new problem to the old one.

For $\tau = 0$ the new system, Eq. (11), is identical to the old system (6), and can be solved using SVD. However, it's convenient to replace $\tilde{A}$ with $\tilde{A}C^{-1}$, and w with Cw. Then, once the exact solution for $\tau = 0$ is found, the solution for $\tau \neq 0$ is obtained immediately [1]:

$$z_i^{(\tau)} = \frac{d_i}{s_i}\cdot\frac{s_i^2}{s_i^2+\tau}. \tag{12}$$

For large $s_i \gg \tau$, the suppression factor $s_i^2/(s_i^2+\tau)$ is close to 1, but for smaller s_i it works as a low-pass filter. So, for a smooth way of eliminating the wildly oscillating contributions (which correspond to the values of i with non-significant d_i and small s_i) one should choose $\tau \simeq s_k^2$ where k is the index of the last significant d.

The solution for the regularised version of the above example is presented in Fig. 3, taken from Ref. [1].

5 Error propagation

The components of the exact solution, $z_i = z_i^{(\tau=0)} = \frac{d_i}{s_i}$, were uncorrelated, and hence had a diagonal covariance matrix. The covariance matrix of the regularized $z_i^{(\tau)}$,

$$Z_{ik}^{(\tau)} \;=\; \frac{s_i^4}{(s_i^2+\tau)^2}\cdot\delta_{ik},$$

is still diagonal, but the errors corresponding to insignificant d_i are suppressed. Propagating these errors back to covariance matrices of w and x, we have:

$$\begin{aligned} W^{(\tau)} &= C^{-1}V\,Z^{(\tau)}V^T\,C^{T-1}, \\ X_{ik}^{(\tau)} &= x_i^{\rm ini}\,W_{ik}^{(\tau)}x_k^{\rm ini}. \end{aligned} \tag{13}$$

Inevitably, for any non-zero τ, the covariance matrices $W^{(\tau)}$ and hence $X^{(\tau)}$ are not diagonal, and the larger the value of τ, the larger are the bin-to-bin correlations in errors.

This behaviour is not surprising: this type of unfolding is nothing else but a fit, and hence behaves as such. In an extreme case of only one significant d_i, all one can say is that the data is *proportional* to the Monte Carlo, with the error essentially reflecting the overall data statistics — but this error is fully correlated between all bins. Unfolding cannot help here — you need more statistics (and/or a better detector, with a "more diagonal" response matrix) to make more d_i values significant!

We, however, have to admit that a non-diagonal covariance matrix creates a presentational problem: showing just $x_i \pm \sqrt{X_{ii}}$ is, in most cases, misleading.

[1] Note that the matrix C, defined according to Eq. (10), is itself singular and needs to be regularised before the inverse, C^{-1}, can be calculated. One can regularise C by including the first finite differences at the end-points, and by adding a small term proportional to the unit matrix (see Ref. [1] for details).

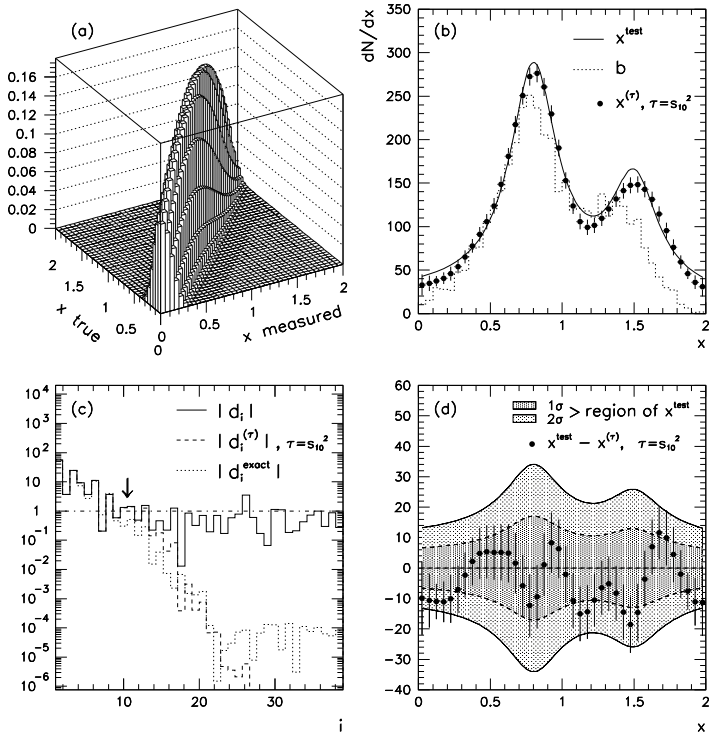


Fig. 3: (a) The response matrix matrix $\hat{A}$. (b) The true distribution $x^{\rm test}$ compared to the measured histogram b and the unfolded distribution $x^{(\tau)}$ for $\tau = s_{10}^2$. (c) The absolute values of d_i (solid line) compared to the regularized r.h.s. (dashed line) and the one unaffected by the statistical fluctuations (dotted line). The horizontal line shows statistical errors in d_i, while the arrow indicates the boundary between the significant and non-significant equations. (d) The deviation of the unfolded distribution from the true one, compared to 1σ and 2σ error bands [1].

6 Calculating the matrix X^{-1}

In order to compare the measurement b with a theoretical prediction $f = f(\alpha)$, one could try *folding* the theory with the detector response matrix, and calculate

$$\chi^2 \;=\; (b - Af)^T B^{-1} (b - Af).$$

Alternatively, because $b = Ax$, one can replace b with Ax and have

$$\begin{aligned} \chi^2 &= (Ax - Af)^T B^{-1} (Ax - Af), \\ &= (x - f)^T A^T B^{-1} A\,(x - f), \\ &= (x - f)^T X^{-1} (x - f), \end{aligned} \tag{14}$$

where $X^{-1} = A^T B^{-1} A$, or, in our notation,

$$X^{-1}_{jk} = \frac{1}{x^{\rm ini}_j x^{\rm ini}_k} \sum_i \tilde{A}_{ij} \tilde{A}_{ik}. \tag{15}$$

Hence, for a problem with fully significant matrix, comparing folded theory with measured data or "raw" theory with unfolded data will give equivalent results. But it is also clear, that in order to compare theory with the unfolded experiment, one needs the *inverse* of the error matrix of the unfolded distribution, rather than the error matrix itself. And from the above it is obvious that the calculation of this inverse does *not* require any regularisation or unfolding.

By substituting $x^{(\tau)}$ for x, one only changes the χ^2 by a "small" amount of order of τ, which may or may not disturb the fit, depending on the model. Hence, if the aim is to fit the unfolded distribution to theory, one is generally better off by folding theory, rather than by unfolding the measurement. However if the effects of regularisation are "small", the results will be equivalent.

If a theoretical parameter happens to be sensitive to an insignificant element of the data, this will show up in the fit one way or another. In any case, in our understanding, the regularised covariance matrix $X^{(\tau)}$, defined in Eq. (13), should only be used for calculating errors on the unfolded distribution. Calculating the *inverse* of $X_{ik}^{(\tau)}$ is neither helpful, nor useful. Instead, if necessary, use the exact X^{-1} from Eq. (15) above.

7 Advice to unfolders

- Three things contribute to the solution: the response matrix, data statistics, and binning.
- Choose binning wisely: large variations in data may be avoided by using variable bin widths. Size of bin-to-bin error correlations may also be affected.
- Rescaling is important: only unfold the equations once they have equal "weights".
- The Monte Carlo sample(s) used for building the response matrix should have as high statistics, and should be as close to the real experiment, as possible.
- Monte Carlo samples used for testing should have the same statistics as the real data; use of larger or smaller samples can be misleading.
- To assess unfolding systematics, vary the matrix within its tolerances, and study the effects on the singular values.
- Even if you are using a different algorithm for unfolding your data, try applying SVD: it will help identify the bottlenecks, and assess any benefits of performing unfolding in the first place.
- It's wise not to expect any miraculous solutions to unfolding problems.

8 Implementations

A number of implementations of the algorithm described in Ref. [1] exist. Some of these are listed below:

- The original Fortran package called `GURU` is still accessible (see Ref. [3]).
- A C++ wrapper for the above Fortran code was developed by G. Hesketh, and can be found at Ref. [4].
- A new C++ implementation of the algorithm, `TSVDUnfold`, now exists in `ROOT` (see Ref. [5], and the talk by K. Tackmann in these proceedings).

9 Acknowledgements

The author is grateful to the organisers for the opportunity to give this presentation, to V. Blobel for many helpful discussions, and to G. Hesketh, A. Hoecker and K. Tackmann for their efforts in re-implementing the algorithm.

References

[1] A. Hoecker and V. Kartvelishvili, "SVD Approach to Data Unfolding," Nucl. Instrum. Meth. A **372** (1996) 469 [arXiv:hep-ph/9509307].

[2] V. Blobel, Unfolding methods in high-energy physics experiments, DESY 84-118 (1984).

[3] `http://www.hep.lancs.ac.uk/internal/guru.tar.gz`.

[4] http://www-d0.fnal.gov/~ghesketh/unfolding.
[5] http://www.root.cern.ch/root/html/TSVDUnfold.html.

An Iterative, Dynamically Stabilized (IDS) Method of Data Unfolding

Bogdan Malaescu
CERN, CH–1211, Geneva 23, Switzerland

Abstract
We describe an iterative unfolding method for experimental data, making use of a regularization function. The use of this function allows one to build an improved normalization procedure for Monte Carlo spectra, unbiased by the presence of possible new structures in data. We unfold, in a dynamically stable way, data spectra which can be strongly affected by fluctuations in the background subtraction and simultaneously reconstruct structures which were not initially simulated.

1 Introduction

Experimental distributions of specific variables in high-energy physics are altered by detector effects. This can be due to limited acceptance, finite resolution, or other systematic effects producing a transfer of events between different regions of the spectra. Provided that they are well controlled experimentally, all these effects can be included in the Monte Carlo simulation (MC) of the detector response, which can be used to correct the data.

We will not focus on the correction of acceptance effects. It is straightforward to perform it on the distribution corrected for the effects arising from a transfer of events between different bins of the spectrum. The detector response is encoded in a transfer matrix connecting the measured and true variables under study. However, as the transfer matrix used in the unfolding is obtained from the simulation of a given physical process, one must perform background subtraction and data/MC corrections for acceptance effects before the unfolding.

Several deconvolution methods for data affected by detector effects have been described in the literature (see for example [1–6]). We present an unfolding method (described in detail in Ref. [7]) allowing one to obtain a data distribution as close as possible to the "real" one for rather difficult, yet realistic, examples. This method is based on the idea that if two conditions are satisfied, namely the MC simulation provides a relatively good description of the data and of the detector response, one can use the transfer matrix to compute a matrix of unfolding probabilities. If the first condition is not fulfilled one can iteratively improve the transfer matrix. The first step of our method provides a good result if the difference between data and normalized reconstructed MC is relatively small over the entire spectrum. If this is not the case, one should proceed with a series of iterations. The regularization of the method is dynamical, coming from the way the data-MC differences are treated in each bin, at each step of the unfolding method.

This method is to be applied to binned, one dimensional data. But, it can be directly generalized to multidimensional problems.

2 Important ingredients of the unfolding procedure

2.1 The regularization function

We use a regularization function $f(\Delta x, \sigma, \lambda)$ to dynamically reduce fluctuations and prevent the transfer of events which could be due to fluctuations, particularly in the subtracted background. This function provides information on the significance of the absolute deviation Δx between data and the simulation in a given bin, with respect to the associated error σ. It is a smooth monotone function going from 0, when $\Delta x = 0$, to 1, when $\Delta x >> \sigma$. λ is a scaling factor, used as a regularization parameter. As we

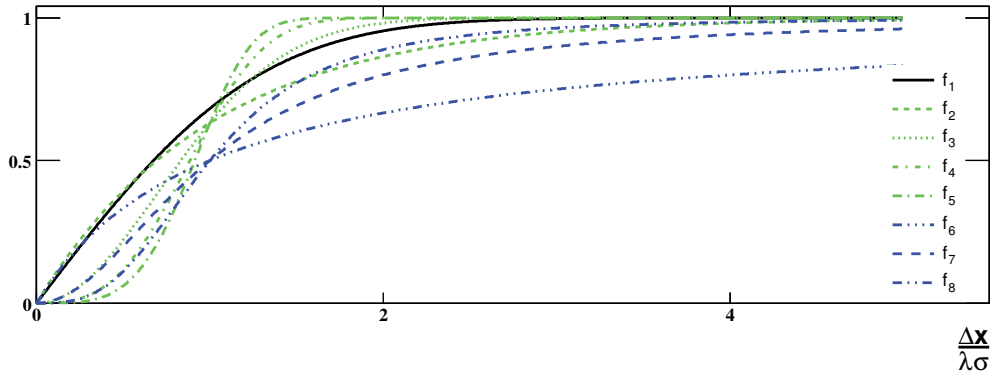


Fig. 1: Behaviour of the functions $f_{1..8}$ with respect to $\Delta x/(\lambda\sigma)$.

shall see in the following, changing the regularization function used in our method will change the way we discriminate between significant deviations and statistical fluctuations.

For the unfolding procedure, we can consider several functions of the relevant variable $\Delta x/(\lambda\sigma)$ (see Fig. 1). In general, we will use different λ parameters for the regularization function for each componenent of the unfolding procedure described in the following. We will see however that some of these parameters can be unified (i.e. assigned identical values) or even dropped (when a trivial value is assigned to them).

2.2 The MC normalization in presence of new structures in data

A rather tricky point is the way the unfolding deals with new structures not considered in the MC simulation but that are present in the data. These structures are affected by the detector effects, and hence they need to be corrected. It seems that the Singular Value Decomposition (SVD) [1] and the iterative [2,3,5,6] methods provide a natural way of performing this correction. However, if the new structures in the data contain a relatively important number of events, they could also affect the normalization of MC spectra with respect to the data (which is needed in the unfolding procedure, as we shall see in the following). For the unfolding procedure described here, we introduce a comparison method between data and MC spectra which is able to distinguish significant shape differences. Exploiting the regularization function introduced before, it counts the events in data (N_d^{MC}) without including those corresponding to significant new structures. Dividing N_d^{MC} by the number of events in the MC (N_{MC}), one obtains the data/MC normalization factor. This procedure is especially useful when the differences between the two spectra consist of relatively narrow structures. Our normalization method allows a meaningful comparison of data and MC spectra and improves the convergence of the algorithm in this case. If the differences are widely distributed, they have smaller impact on the normalization factor, and the sensitivity of our method is weaker too.

2.3 The estimation of remaining fluctuations from background subtraction

Experimental spectra are generally obtained after background subtraction; this operation (performed before the unfolding) results in an increase in errors for the corresponding data points. Due to bin-to-bin or correlated fluctuations of the subtracted background, these points can fluctuate within their errors. These fluctuations can be important especially on distribution tails or dips, where the signal is weak and the background subtraction relatively large. Actually, it is only when the background subtraction produces a large increase in the uncertainties of the data points (well beyond their original statistical errors), that these fluctuations become a potential problem. The problematic regions of the spectrum

can be identified even before going to the unfolding. When computing the corrected distribution, the unfolding procedure has to take into account the size of the experimental errors, including those from background subtraction. At this step we identify large but not significant data-MC deviations. Not doing so could result in propagation of large fluctuations and uncertainties to more precisely known regions of the spectrum. Such an effect is to be avoided, and is a problem that we treat carefully. To the best of our knowledge, none of the previously published methods address this second type of problem and distinguish it from the previous one.

2.4 Folding and unfolding

In the MC simulation of the detector one can directly determine the number of events which were generated in the bin j and reconstructed in the bin i (A_{ij}). Provided that the transfer matrix A gives a good description of the detector effects, it is straightforward to compute the corresponding folding and unfolding matrix: $P_{ij} = \frac{A_{ij}}{\sum_{k=1}^{n_d} A_{kj}}, \tilde{P}_{ij} = \frac{A_{ij}}{\sum_{k=1}^{n_u} A_{ik}}$. Here, n_d is the number of bins in data (and reconstructed MC), while n_u is the one in the unfolding result (and true MC). The folding probability matrix, as estimated from the MC simulation, P_{ij} gives the probability for an event generated in the bin j to be reconstructed in the bin i. The unfolding probability matrix $\tilde{P}_{ij}$ corresponds to the probability for the "source" of an event reconstructed in the bin i to be situated in the bin j.

The folding matrix describes the detector effects, and one typically relies on the simulation in order to compute it. The quality of this simulation must therefore be the subject of dedicated studies within the analysis, and generally the transfer matrix can be improved before the unfolding. Systematic errors can be estimated for it and they are propagated to the unfolding result. The unfolding matrix depends not only on the description of detector effects but also on the quality of the model which was used for the true MC distribution. It is actually this model which can (and will) be iteratively improved, using the comparison of the true MC and unfolded distributions.

In order to perform the unfolding, one must first use the iterative procedure described in Sect. 2.2 to determine the MC normalization coefficient (N_d^{MC}/N_{MC}). One can then proceed to the unfolding, where, in the case of identical initial and final binnings, the result for $j \in [1; n_u]$ is given by:

$$u_j = t_j \cdot \frac{N_d^{MC}}{N_{MC}} + B_j^u + \sum_{k=1}^{n_d} f\left(|\Delta d_k|, \tilde{\sigma} d_k, \lambda\right) \Delta d_k \, \tilde{P}_{kj} + \left(1 - f\left(|\Delta d_k|, \tilde{\sigma} d_k, \lambda\right)\right) \Delta d_k \, \delta_{kj},$$

with $\Delta d_k = d_k - B_k^d - \frac{N_d^{MC}}{N_{MC}} \cdot r_k$. Here, for a given bin k, t_k is the number of true MC events, while $\tilde{\sigma} d_k$ is the uncertainty to be used for the comparison of the data (d_k) and the reconstructed MC (r_k). B is the (estimated) vector of the number of events in the data distribution which are associated to a fluctuation in the background subtraction. In the case of different binnings for the data and the unfolding, the Kroneker symbol δ must be replaced by a rebinning transformation R.

The first two contributions to the unfolded spectrum are given by the normalized true MC and the events potentially due to a fluctuation in background subtraction, which we do not transfer from one bin to another. Then one adds the number of events in the data minus the estimated effect from background fluctuations, minus the normalized reconstructed MC. A fraction f of these events are unfolded using the estimate of the unfolding probability matrix $\tilde{P}$, and the rest are left in the original bin. With the description of the regularization functions given in Section 2.1, it is clear that reducing λ would result in increasing the fraction of unfolded events, and reducing the fraction of events left in the original bin. Choosing an appropriate value for this coefficient provides one with a dynamical attenuation of spurious fluctuations, without reducing the performance of the unfolding itself.

2.5 The improvement of the unfolding probability matrix

As explained in the introduction, if the initial true MC distribution does not contain or badly describes some structures which are present in the data, one can iteratively improve it, and hence the transfer

matrix. This can be done by using a better (weighted) true MC distribution, with the same folding matrix describing the physics of the detector, which will yield an improved unfolding matrix.

The improvement is performed for one bin j at the time, exploiting the difference between an intermediate unfolding result and the true MC (Δu_j):

$$A'_{ij} = A_{ij} + f\left(|\Delta u_j|, \tilde{\sigma} u_j, \lambda_M\right) \Delta u_j \, P_{ij} \, \frac{N_{MC}}{N_d^{MC}}, \text{ for } i \in \{1; N_d\} \, . \tag{1}$$

Here, λ_M (for modification) stands for the regularization parameter used when modifying the matrix. Increasing λ_M would reduce the fraction of events in Δu_j used to improve the transfer matrix.

This method allows an efficient improvement of the folding matrix, without introducing spurious fluctuations. Actually, the amplification of small fluctuations can be prevented at this step of the procedure too.

3 The iterative unfolding strategy

In this section we describe a general unfolding strategy, based on the elements presented before. It works for situations presenting all the difficulties listed before, even in a simplified form, where some parameters are dropped and the corresponding steps get trivial. The strategy can be simplified even more, for less complex problems.

One will start with a null estimate of the fluctuations from background subtraction. A first unfolding, as described in Section 2.4, is performed, with a relatively large value of $\lambda = \lambda_L$. This step will not produce any important transfer of events from the regions with potential remaining background fluctuations (provided that λ_L is large enough), while the correction of resolution effects for the new structures in data will be limited too.

At this level one can start the iterations:

1) Estimation of the fluctuations in background subtraction

An estimate of the fluctuations in background subtraction can be obtained using the procedure described in Section 2.3. The parameter λ_S used here must be large enough, in order not to underestimate them. λ_S can however not be arbitrary large, as this operation must not bias initially unknown structures, by not allowing their unfolding.

2) Improvement of the unfolding probability matrix

Using the method described in Section 2.5, one can improve the folding matrix A. A parameter λ_M, small enough for an efficient improvement of the matrix, yet large enough not to propagate spurious fluctuations (if not eliminated at another step), must be used at this step.

3) An improved unfolding

A parameter λ_U will then be used to perform an unfolding following Section 2.4, exploiting the improvements done at the previous step. It must be small enough to provide an efficient unfolding, but yet large enough to avoid spurious fluctuations (if not eliminated elsewhere).

These three steps will be repeated until one gets a good agreement between data and reconstructed MC plus the estimate of fluctuations in background subtraction. Another way of proceeding (providing similar results) could consist in stopping the iterations when the improvement brought by the last one on the intermediate result is relatively small. The values of the λ parameters are to be obtained from (realistic) toy simulations, and some of these can be dropped by using trivial (null) values for them. The needed number of steps can also be estimated using these simulations.

4 A complex example for the use of the unfolding procedure

In the following we briefly present a rather complex, yet realistic test, proving the robustness of the method. It exhibits all the features discussed previously, which are simultaneously taken into account by the unfolding. For the clarity of the presentation, the structures and dips of the spectrum are separated.

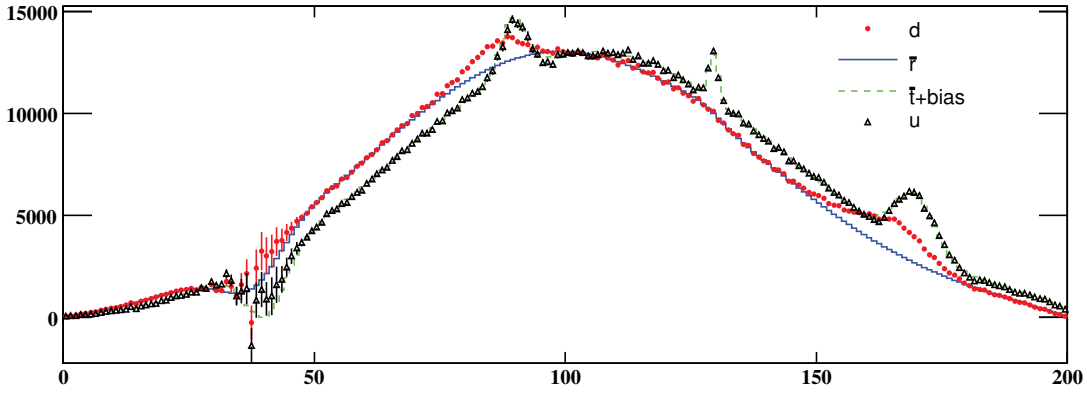


Fig. 2: The unfolding result after 65 iterations (u, triangles), compared to the data distribution (d, filled circles), the reconstructed MC in the model ($\bar{\mathrm{r}}$,solid line) and the true MC model plus the new structures ($\bar{\mathrm{t}} + \mathrm{bias}$, dashed line).

The structure around the bin 130 (see Fig. 2) is present both in data and simulation, while the ones at 90 and 170 are only in data. The dip around 40 is affected by large fluctuations due to background subtraction in data. All the structures are affected by resolution and a systematic transfer of events, from high to lower bin numbers.

The first unfolding step was performed with a very large value for λ_L and it corrects all the elements of the spectrum which are simulated in the MC, for both kinds of transfer effects (in spite of the fact that they are relatively important). The final unfolding result (after iterations) reconstructs well all the structures in the data model, without introducing important systematic effects due to the fluctuations in background subtraction (see Fig. 2). The errors of the unfolding result(s) were estimated using 100 MC toys, with fluctuated data and transfer matrix for the unfolding procedure.

Another example for the use of the IDS unfolding method, with less statistics available in the spectra, has been presented in Ref.. [8].

5 Conclusions

We have described a new iterative method of data unfolding, using a dynamical regularization. It allows us to treat several problems, like the effects of new structures in data and the large fluctuations from background subtraction, which were not considered before. This method has been tested for complex examples, and it was able to treat correctly all the effects mentioned before.

References

[1] A. Hocker and V. Kartvelishvili, Nucl. Instrum. Meth. A **372**, 469 (1996) [arXiv:hep-ph/9509307].

[2] V. Blobel, DESY 84-118, [arXiv:hep-ex/0208022].

[3] A. Kondor, JINR-E11-82-853.

[4] L. Lindemann and G. Zech, Nucl. Instrum. Meth. A **354**, 516 (1995).

[5] G. D'Agostini, Nucl. Instrum. Meth. A **362**, 487 (1995).

[6] P. D. Acton *et al.* [OPAL Collaboration], Z. Phys. C **59**, 1 (1993).

[7] B. Malaescu, arXiv:0907.3791 [physics.data-an].

[8] K. Tackmann's talk at this workshop.

SVD-based unfolding: implementation and experience

Kerstin Tackmann, Andreas Höcker
CERN, Geneva, Switzerland

Abstract
With the first year of data taking at the LHC by the experiments, unfolding methods for measured spectra are reconsidered with much interest. Here, we present a novel ROOT-based implementation of the Singular Value Decomposition approach to data unfolding, and discuss concrete analysis experience with this algorithm.

1 Introduction

The measured spectrum of a physical observable is usually distorted by detector effects, such as finite resolution and limited acceptance. A comparison of the measured spectrum with theoretical predictions requires a removal of these effects to obtain the true, underlying physical spectrum, or the folding of these effects into the theoretical prediction. Unfolding methods provide ways for correcting the measured distributions, where the difficulty lies in the statistical instability of the inversion problem, requiring regularization. One widely used method is based on a singular value decomposition (SVD) of the detector response matrix [1].

The unfolding problem can be formulated as a matrix equation, $\hat{A}_{ij}x_j = b_i$, where x is the true, physical distribution, b the measured distribution. $\hat{A}_{ij}$ is the probability for an event generated in bin j to be reconstructed in bin i and as such, $\hat{A}$ describes finite resolution and inefficiencies and can be obtained from the simulation (or appropriate control samples). The singular value decomposition of $\hat{A}$ serves both for shedding light on the underlying instability of the problem, as well as for providing a solution. Small singular values, which are often present in detector response matrices, are found to greatly enhance statistical fluctuations in the measured distribution. A suitably chosen regularization procedure dampens the enhanced fluctuations. Rewriting the above equation to $A_{ij}w_j = b_i$, where A_{ij} now contains numbers of events rather than probabilities, and w describes the ratio between the desired physical distribution and the underlying true distribution in the simulation (for example), allows for a better treatment of the statistical uncertainties in the detector matrix. At the same time, this allows for a physically motivated regularization via a discrete minimum-curvature condition on the ratio of the unfolded distribution and a simulated truth distribution, which corresponds to retaining the statistically significant contributions of w, shown to be related to the larger singular values in the decomposition of A.

This note presents a `C++` implementation of the SVD-based unfolding, discusses analysis experience with this algorithm, and provides a comparison to the iterative dynamically stabilized unfolding method (IDS) [2] for a concrete example.

2 ROOT-based implementation

A `C++` implementation of the SVD-based unfolding is provided by `TSVDUnfold`, which is part of the ROOT analysis framework [3] as of version 5.28. It can also be used through the `RooUnfold` framework [4], which is based on ROOT and comes with additional functionality.

`TSVDUnfold` provides access to the singular values of the detector response matrix and to the distribution of the $|d_i|$ (see Ref. [1]), which help to properly set the regularization strength parameter in the unfolding. `TSVDUnfold` also allows to propagate covariance matrices of the measured spectrum through the unfolding using pseudo experiments. In addition it provides the covariance matrix of the

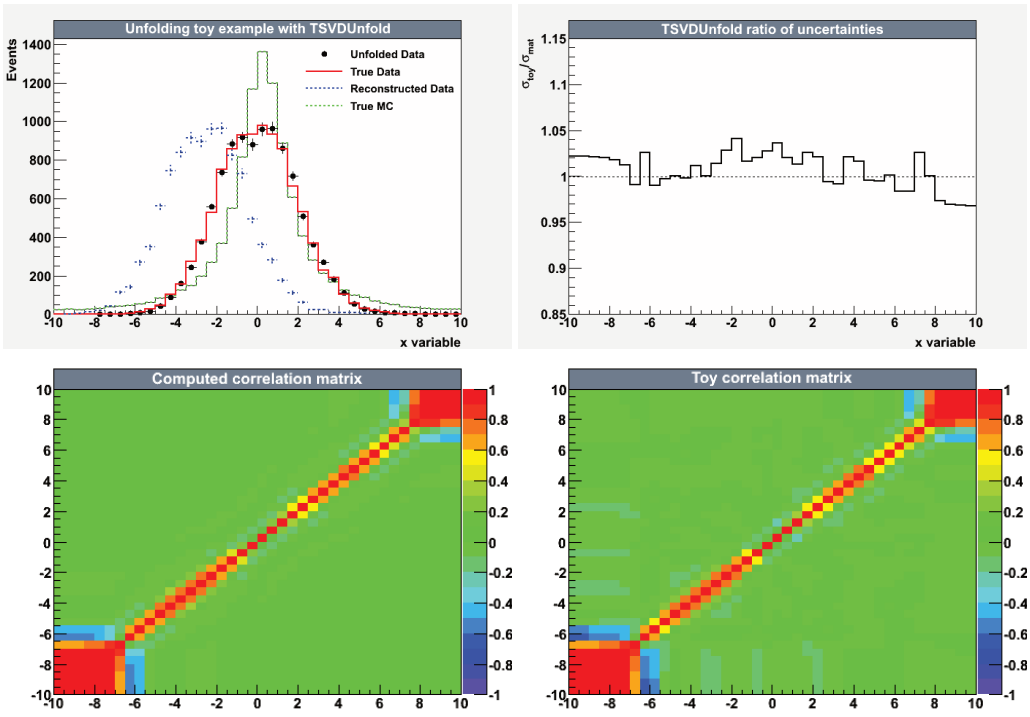


Fig. 1: Unfolding toy example. (Left top) Reconstructed and unfolded toy data, as well as the truth distributions for toy data and toy simulation. (Right top) Ratio of diagonal errors obtained by pseudo experiments and computed during the unfolding. (Left bottom) Correlation matrix on the unfolded spectrum as computed during the unfolding and (Right bottom) as obtained from pseudo experiments.

unfolded spectrum related to finite statistics in the simulation sample (or control sample) that is used to determine the detector response matrix, also making use of pseudo experiments.

More recently, `TSVDUnfold` has been extended to also provide the regularized covariance matrix and the inverse covariance matrix (not regularized) computed during the unfolding (see Eqs. (52,53) in Ref. [1]). In addition, the new version of `TSVDUnfold` implements the internal rescaling of the unfolding equations making use of the full covariance matrix of the measured spectrum (see Eq. (34) in Ref. [1]) rather than only its diagonal elements.

3 Covariance matrices

The covariance matrices of the unfolded spectrum as computed during the unfolding and as obtained from pseudo experiments, respectively, have been compared for a toy example (see Fig. 1) and have been found in good agreement. The uncertainties (taken from the diagonal elements of the covariance matrices) provided by the two methods agree to better than 4% and the correlations are well-reproduced. Even in the case of non-optimal regularization, the two methods provide compatible results: the uncertainties obtained with the two methods have been found to agree within 6% (11%) for a strongly under- (over-) regularized unfolding, with compatible correlation patterns.

4 Experience with SVD-unfolding in *BABAR*

The SVD-based unfolding has been used in numerous data analyses over the past 15 years, among which is the unfolding of the hadronic mass spectrum in inclusive, charmless, semileptonic B-meson decays, $B \to X_u \ell \nu$, at the *BABAR* experiment [5]. Due to the nature of the measured spectrum, its unfolding and in particular the determination of the appropriate regularization required careful studies.

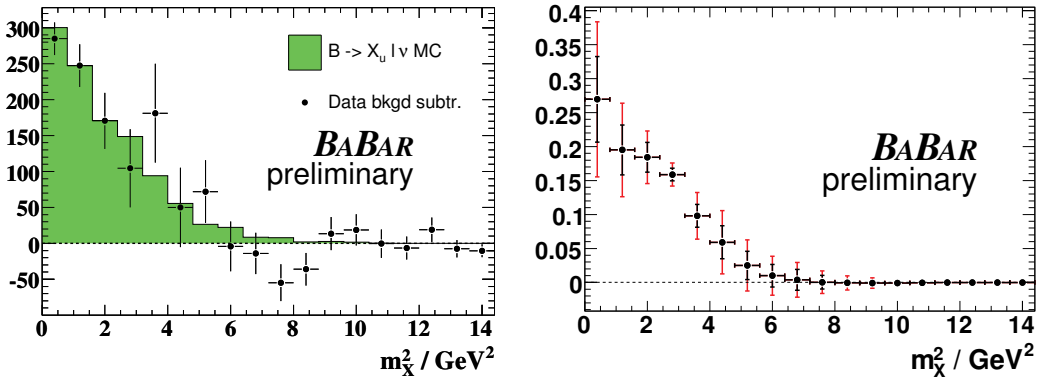


Fig. 2: Unfolding example from the *BABAR* experiment. (Left) Measured hadronic mass spectrum (statistical uncertainties only) and signal Monte Carlo simulation in $B \to X_u \ell\nu$ decays. (Right) Normalized unfolded hadronic mass spectrum with total (outer error bars) and statistical (inner error bars) uncertainties.

The relatively low statistics of estimated 1027 signal events and the subtraction of the dominant $B \to X_c \ell\nu$ backgrounds result in sizable statistical and systematic uncertainties. The size of the bins has been chosen to equal the hadronic mass resolution in signal events. Due to the large uncertainty in the reconstruction efficiency of the tagging method used, which results in a significantly better hadronic mass resolution, the unfolded spectrum is normalized to unit area, which results in increased bin-by-bin correlations.

The regularization has been determined with the use of pseudo experiments, where the toy data and toy simulation distributions and detector response differ in the assumed value of the b-quark mass, which determines the shape of the inclusive hadronic mass spectrum and is one of the primary results of the analysis. The regularization has been chosen such that the unfolding bias in the spectral moments of the unfolded spectrum, which are directly related to the b-quark mass, is small compared to their statistical errors.

A few observations have been made which are of relevance for the unfolding of spectra with sizable uncertainties. The unfolding gains stability when the internal rescaling of the equations that is performed by the SVD-based unfolding (see Ref. [1]) takes both the statistical and the systematic uncertainties into account, since this provides a better estimate of how well the different regions of the measured distribution are known. Moreover, the propagation of the covariance matrices of the measured spectrum to the unfolded spectrum shows a more linear behavior and is less likely to be affected by instabilities in the unfolding in the presence of sizable uncertainties when the covariance matrices related to different sources of uncertainties are propagated separately and then combined.

5 Comparison to interative dynamically stabilized unfolding

It is instructive to compare the results of different unfolding methods for the same example. Here, we present the results for the toy example of Sect. 3, using both SVD-based and IDS [2] unfolding.

The regularization for the two methods has been determined independently. For the SVD-based unfolding, the distribution of the $|d_i|$ has been used to chose the regularization ($k = 16$). For the IDS unfolding, it has been determined using the toy data as well as the reconstructed improved toy simulation distributions. The unfolding results can be seen in Fig. 3. Neither unfolding result shows any obvious bias with the chosen regularization. However, the result of the IDS unfolding shows somewhat larger fluctuations around the true distributions as well as larger uncertainties, which points to a looser regularization than that used for the SVD-based unfolding. In addition, the observed pattern in the bin-by-bin correlations is very different. The result of the SVD-based unfolding shows positive correlations

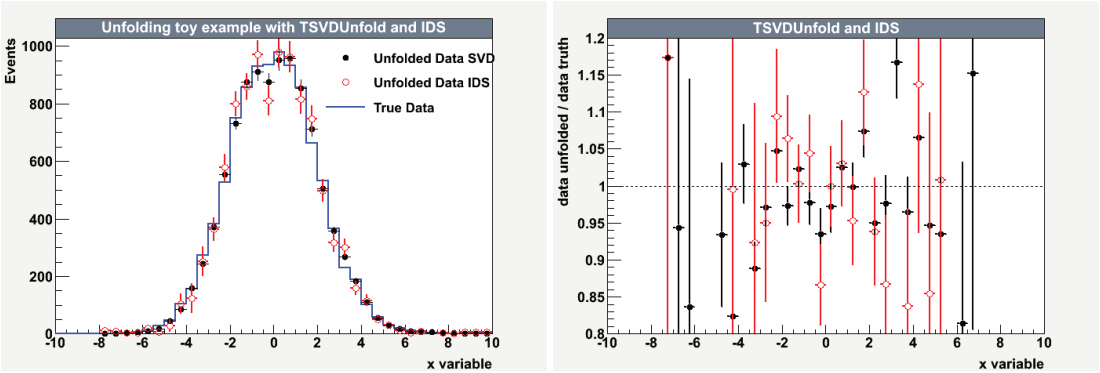


Fig. 3: Unfolding toy example. (Left) The results obtained with SVD-based unfolding and IDS unfolding are compared to the true toy distribution. (Right) The ratio of the unfolded distributions and the truth distribution.

between neighbouring bins, negative correlations in the medium range, and very small correlations in the long range. The result of the IDS unfolding in general shows smaller correlations, and neighbouring bins tend to be anti-correlated. In general, the stronger the regularization, the larger and broader are the positive correlations between adjacent bins. The difference in the correlations observed between the SVD-based and IDS unfolding results are due to the stronger regularization in the SVD-based unfolding, which is also apparent in the smaller diagonal errors.

6 Summary

`TSVDUnfold` provides a `C++` implementation of the SVD-based unfolding algorithm and is available as part of the ROOT analysis framework. Recently, it has been improved to take into account bin-by-bin correlations in the measured spectrum. SVD-based unfolding has been successfully used in many data analyses and a concrete example from the *BABAR* experiment has been presented, along with observations that are of relevance for unfolding spectra which are subject to large uncertainties. In addition, unfolding results for a toy example have been compared using SVD-based and IDS unfolding.

Acknowledgements

The authors wish to thank Vakhtang Kartvelishvili for discussions and advice related to the SVD-based unfolding, as well as Heiko Lacker for the collaboration in the implementation of `TSVDUnfold`. We furthermore wish to thank Bogdan Malaescu for general discussions on unfolding and providing the unfolded example spectrum using IDS unfolding. Thanks to Lorenzo Moneta `TSVDUnfold` is now distributed as part of the ROOT analysis framework, and thanks to Tim Adye it can be used also through `RooUnfold`.

References

[1] A. Höcker and V. Kartvelishvili, Nucl. Instrum. Meth., A **372** (1996), 469-481 [arXiv:hep-ph/9509307].
Also: V. Kartvelishvili, "Unfolding with SVD", these Proceedings.

[2] B. Malaescu, "An Iterative, Dynamically Stabilized (IDS) Method of Data Unfolding", this report.
Also: B. Malaescu, arXiv:0907.3791 [physics.data-an]

[3] R. Brun and F. Rademakers, Nucl. Instrum. Meth., A **372** (1996), 496.

[4] T. Adye, "Unfolding algorithms and tests using RooUnfold", this report.

[5] K. Tackmann [*BABAR* Collaboration], Eur. Phys. J. A **38** (2008) 137 [arXiv:0801.2985 [hep-ex]].

Regularization by Control of the Resolution Function

Michael Schmelling
MPI for Nuclear Physics, Heidelberg, Germany

Abstract
Unfolding based on Singular Value Decomposition is used to illustrate how regularization is related to control of the resolution function and thereby the interpretation of the unfolding result.

1 Introduction

A central problem in many data analyses is the correction of the raw measurements for distortions caused by the experimental setup. In the 1-dimensional case the mapping between a true physical density $b(y)$ and the experimentally observable one $a(x)$ is given by the integral equation

$$a(x) = \int dy\ g(x,y) \cdot b(y)\ . \tag{1}$$

In the high energy physics use case the unknown $b(y)$ often is proportional to a cross-section, i.e. non negative. In the following the response function $g(x,y)$, which for a given true value y is the probability density function for the actually observed quantity x, is assumed to be known. In typical applications the observable density $a(x)$ is estimated in a counting experiment. Since a finite number of events cannot fully determine a continuous function $a(x)$, it is obvious that some discretization has to be performed. A widely used and intuitive way to discretize density functions is by means of histograms, i.e. average densities over finite intervals rather than truly continuous functions. Equation (1) then becomes

$$a_k = \sum_{i=1}^{n} G_{ki} \cdot b_i \qquad \text{with} \qquad k = 1, 2, \ldots m \tag{2}$$

where a_k and b_i are the integrals over finite bins in x and y, respectively. The response function $g(x,y)$ translates into the response matrix G_{ki}. It is worth noting that for finite bin widths Δy the response matrix will general still depend on the unknown $b(y)$. Only in the limit of infinitesimal bin sizes the modeling of the detector response becomes truly independent of $b(y)$. In the following this kind of discretization errors will be ignored.

Before going further it is instructive to analyze the structure of the problem in terms of Fourier analysis [1]. Introducing an amplitude function $A(\omega)$ for the true distribution $b(y)$ and assuming a Gaussian response function $g(x,y)$ one schematically finds

$$b(y) = \int d\omega\ A(\omega) \cos(\omega y) \tag{3}$$

and

$$a(x) = \int dy\ \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-y)^2/2\sigma^2} \cdot b(y) = \int d\omega\ e^{-\omega^2\sigma^2} \cdot A(\omega) \cos(\omega x)\ . \tag{4}$$

The high frequency components of the observable density $a(x)$ are exponentially suppressed by the finite resolution σ of the response function, and thus can only be measured in the limit of very high statistics. It follows [2] that in practice unfolding will not be able to perform a complete correction but should rather be understood as "improvement of the resolution function". This shall be made more quantitative in the following.

2 Singular Value Decomposition

Singular Value Decomposition (SVD) highlights the properties of any matrix $Q[m,n]$ with m rows and n columns and $m \geq n$ by factorizing it in the form

$$Q[m,n] = U[m,n] \cdot W[n,n] \cdot V[n,n]^T \tag{5}$$

with a non-negative diagonal matrix W and orthogonal matrices U and V satisfying

$$U^T \cdot U = V^T \cdot V = V \cdot V^T = \mathbf{1}_n \; . \tag{6}$$

Here $\mathbf{1}_n$ denotes the unit matrix in n-dimensions. The diagonal elements of W are the "singular values" of Q, usually sorted in descending order.

SVD allows to analyze the nature of a given unfolding problem by diagonalizing it [3]. The first step is a linear transformation M of the measurements a to $a' = M\,a$ such that the covariance matrix of a' becomes the unit matrix, $C(a') = M\,C(a)\,M^T = \mathbf{1}_m$. SVD of the transformed unfolding problem then yields

$$a' = (M\,G)\,b = (U\,W\,V^T)\,b \tag{7}$$

and, introducing *normalized measurements* $u = U^T\,a'$ with again unit covariance matrix $C(u) = U^T\,C(a')\,U = \mathbf{1}_n$, the diagonalized unfolding problem can be expressed in the form

$$u = W\,V^T\,b \qquad \text{or} \qquad W^{-1}\,u = V^T\,b \; . \tag{8}$$

The first equation in eq.(8) is a discrete analog of eq.(4) for an arbitrary response matrix. The vector u is a representation of the measurements where the individual components are uncorrelated and have unit variance, and $V^T\,b$ the expansion of the discretized true distribution into the orthonormal basis provided by the rows of V^T. The connection between the two is given by the diagonal elements of W, which for typical problems has a steeply falling spectrum of singular values. The higher order expansion coefficients of b thus quickly reach expected values close to zero, which the measurements with unit errors cannot resolve.

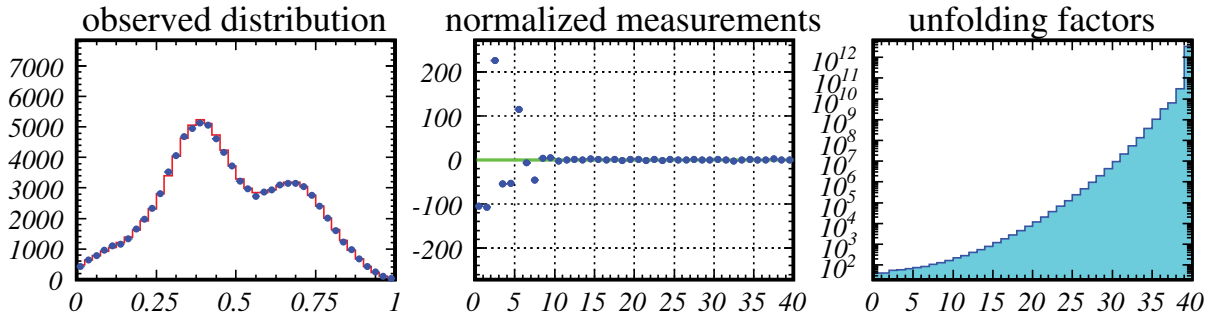


Fig. 1: Observed distribution with 10^5 events compared to the expectation (left), normalized measurements (center) and unfolding factors, i.e. the inverse of the singular values of the diagonalized unfolding problem (right).

This is illustrated by an example [1] with a true distribution $b(y)$ defined on the interval $y \in [0,1]$ and response function $g(x,y)$

$$\begin{aligned} b(y) &= \frac{20.334}{100 + (10y-2)^2} + \frac{2.0334}{1 + (10y-4)^2} + \frac{4.0668}{4 + (20y-15)^2} \\ g(x,y) &= \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(x - \left[y - \frac{1}{10}y^2\right]\right)^2\right) \cdot \left(1 - 2\left(y - \frac{1}{2}\right)^2\right) \; , \end{aligned} \tag{9}$$

the latter featuring a Gaussian resolution $\sigma = 0.05$, a quadratic bias, and a parabolic efficiency loss towards the phase space limit. The result of a discretization with 40 bins for the true and the observed distribution, assuming a total of 10^5 events in the measured distribution is shown in Fig. 1. One clearly sees that under the assumed conditions only the leading 10 to 15 coefficients of the solution are experimentally accessible. Using higher order coefficients to reconstruct the unfolded distribution will mainly amplify statistical fluctuations. So only the leading order coefficients of the expansion of the solution can be determined experimentally.

Not knowing the higher order coefficients implies that the fine structure cannot be fully resolved. There is a loss of resolution, and the question arises whether it is possible to quote some equivalent Gaussian resolution.

This can be answered heuristically by studying the expansion of a delta-pulse into a set of orthogonal functions. In the discrete case one simply has vector a b with components $b_i = \delta_{iI}$, i.e. zero everywhere except for component I, which is expanded into an orthogonal basis $v_k, k = 1, \ldots, n$. Such a basis is, for example, given by the columns of the matrix V introduced above. The expansion coefficients u_k are obtained by the scalar products $u_k = v_k \cdot b = v_{kI}$. Zeroing (i.e. discarding) the higher order coefficients in order to study what happens when those are not known, and re-synthesizing the content of bin I then yields

$$\hat{b}_I = \sum_{k=1}^{j} u_k \, v_{kI} = \sum_{k=1}^{j} v_{kI}^2 \; . \tag{10}$$

For $j = n$, i.e. in case all expansion coefficients are used, one recovers $\hat{b}_I = b_I = 1$. The truncated sum has $\hat{b}_I < 1$, but non-zero content in the neighboring bins.

The same happens when a delta-function is smeared by a Gaussian. The integral over the central bin is no longer unity, and part of the normalization ends up in the neighboring bins. This suggests to define an equivalent Gaussian resolution σ by the condition

$$\int_{-w/2}^{w/2} dx \, \frac{1}{\sqrt{2\pi}\sigma} \, e^{-x^2/\sigma^2} = \hat{b}_I \; , \tag{11}$$

with w the bin width and the parameter σ chosen such that the central bin has the same content as that of the truncated re-synthesis. One finds

$$\frac{\sigma}{w} = \frac{1}{\sqrt{8}\,\mathrm{erf}^{-1}(\hat{b}_I)} \oplus \frac{1}{\sqrt{12}} \; . \tag{12}$$

The term $1/\sqrt{12}$ (to be added in quadrature) has been put by hand to account for the fact that even for $b_I = 1$ the attainable resolution is limited by the finite bin width. Figure 2 shows for the above example that this heuristic approach gives a realistic estimate for the actual resolution of the truncated series.

3 Unfolding

The above example shows that only the leading order coefficients of the expansion of the solution into an orthonormal basis can be used for the construction of the unfolded distribution. Nevertheless, the resolution of the truncated result will usually be better than the resolution of the measurement. Also, biases and efficiency losses will be corrected.

Suppression or, more generally, adjustment of the insignificant components is also referred to as "regularization". The simplest scheme is to fix them to zero. Alternative choices, for example asking for minimal curvature or Maximum Entropy of the unfolding result can be employed to supply the missing coefficients. In general everything is allowed which is consistent with the measurements, including adjusting the well measured leading order coefficients within their uncertainties. In all cases the resolution of the result is determined by the number of coefficients which are used.

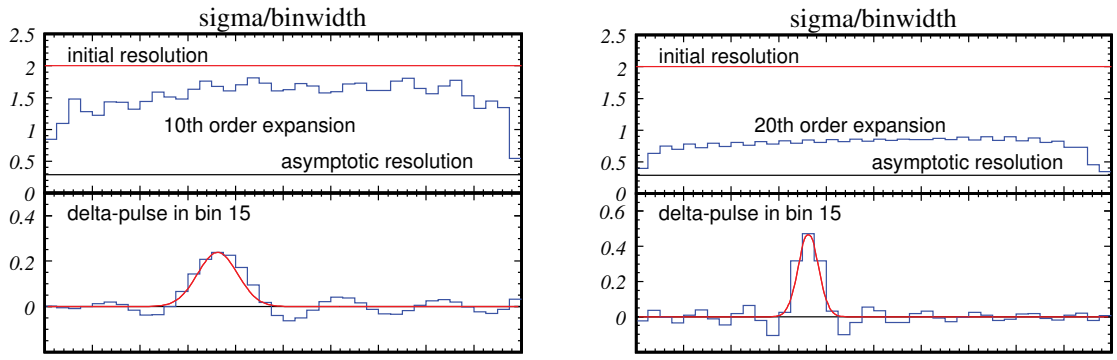


Fig. 2: Resolution of a truncated expansion into an orthogonal basis using the leading 10 (left) or 20 (right) from a total of 40 expansion coefficients. The actual values apply for the example discussed in the text. The resolution is given in units of the bin width, i.e. the asymptotic value is $1/\sqrt{12}$. The initial resolution of $\sigma = 0.05$ corresponds to the width of 2 bins. In the top parts the estimated resolution is shown as a function of the bin number, the lower parts compare the Gaussian estimates (continuous line) with the truncated synthesis (histogram) for a delta-pulse in bin 15.

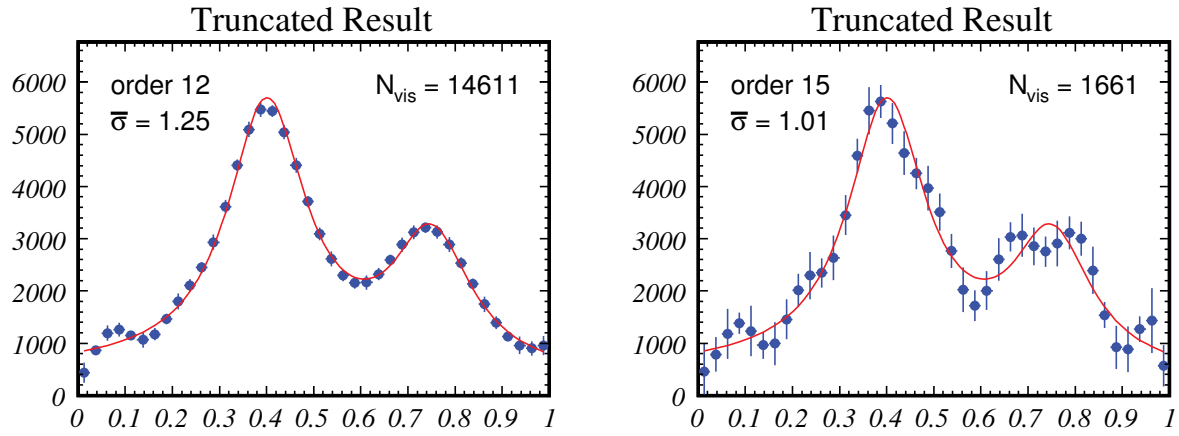


Fig. 3: Unfolding results based on the leading 12 (left) or 15 (right) coefficients of the expansion of the solution into an orthonormal basis. Also shown (continuous line) is the true distribution for the problem studied. Comparing the two results illustrates how better resolution (smaller $\bar{\sigma}$) implies larger uncertainties.

Figure 3 shows the unfolded distribution for the example introduced above when using either the leading 12 or the leading 15 coefficients. The errors are obtained by linear error propagation. Two figures-of-merit to judge the result are given, the average resolution $\bar{\sigma}$ in units of the bin width, obtained by averaging eq.(12) over all bins, and the visible statistics in the unfolded histogram defined as $N_{\rm vis} = (\sum b_i)^2 / \sum \sigma^2(b_i)$. The quantity $N_{\rm vis}$ is constructed such, that for independent poissonian errors it is identical to the number of entries in the histogram. While $\bar{\sigma}$ is proportional to correlation length between neighboring bins, $N_{\rm vis}$ is a simple measure for the smoothness of the result. For any quantitative statements regarding errors the full covariance matrix of the result has to be considered.

In both cases the unfolding result is consistent with the true distribution, which is overlayed in the plots. The bias and efficiency losses assumed in the model are corrected, and the resolution which initially was $\sigma/w = 2$ is improved to $\bar{\sigma} = 1.25$ and $\bar{\sigma} = 1.01$, respectively, Since the true distribution is sufficiently smooth, already the low order result looks good. The bins of the unfolded distribution, however, are correlated. When interpreting the result, one has to be aware that statements about average densities can only be made for regions like the FWHM of the effective resolution. To make this explicit the final results should be rebinned accordingly, taking into account the correlations between the bins.

4 Conclusions

Unfolding in general can only partly correct for distortions caused by an imperfect detector in the sense that the result still is a limited resolution image of the actual truth. This partial correction is referred to as regularization. For sufficiently smooth distributions it will render the unfolding result indistinguishable from the truth. The level of regularization can be characterized by the resolution of the unfolding result, which indicates the typical range over which reliable density estimates can be obtained. In this paper regularization was performed by a simple SVD-based truncation of insignificant components. More sophisticated methods, using also a priori information like positivity will in general perform better. Nevertheless, the basic features caused by information loss due to finite resolution are universal.

References

[1] V. Blobel, "Unfolding Methods in High Energy Physics Experiments", DESY 84-118.

[2] V. P. Zhigunov, "Improvement of resolution function as an inverse problem", NIM 216 (1983) 183.

[3] A. Höcker and V. Kartvelishvili, "SVD Aproach to Data Unfolding", NIM A 372 (1996) 469.

ARU – towards automatic unfolding of detector effects

H.P. Dembinski and M. Roth
IEKP & IK, KIT Karlsruhe, Germany

Abstract
This article presents the ARU algorithm, a general non-interactive algorithm for the unfolding of detector effects (resolution effects, efficiency, non-linear response) from one-dimensional data distributions. ARU uses an unbinned maximum-likelihood fit with a weighted regularization term, based on the relative information in the solution with respect to a reference distribution. The optimal regularization weight is found by minimizing the mean squared error of the solution. The algorithm's performance is demonstrated in a study of a toy data sets. The analysis shows that the bias on average is smaller than the statistical uncertainties which are properly estimated by the fit.

1 Introduction

The unfolding of detector effects from a measured distribution is a standard problem in particle physics, but a difficult one: the unfolding problem itself is ill-posed [1, 2] and has no unique solution. Several unfolding algorithms [3–7] are known to particle physicists, each with different strengths and weaknesses. This article describes the Automatic Regularized Unfolding (ARU) algorithm, strongly influenced by the works of Blobel [4] and Schmelling [6]. ARU is a regularized fit: a flexible parametrization with many free parameters is fitted to the data and softly constrained by a regularization term. The regularization term allows to reduce the freedom of the fit in a smooth way so that over-fitting is avoided.

ARU is a general non-parametric algorithm for unfolding one-dimensional data distributions that requires no user interaction. The unfolded solution that ARU chooses is optimal with respect to the principle of minimum mean squared error. The analysis of the data is completely unbinned with the advantage that no information is lost. The regularization adapts itself to the data distribution in order to minimize bias.

ARU's non-linear regularization term is a variant of the Kullback-Leibler divergence [8] between the solution and a reference distribution and measures the relative information in the solution with respect to the reference distribution. This choice is invariant to transformations and generalizes the principle of maximizing the entropy (minimizing the information) in the unfolded solution [2, 6]. The regularization introduces a bias as it pulls the solution to the reference. The bias is minimized by using a zero-order approximation of the solution as the reference distribution, obtained by correcting the original data distribution only for efficiency and calibration effects, but not for the critical resolution effects. This approach is invariant to coordinate transformations and self-consistent. If no resolution effects are present, the reference distribution becomes equal to the solution and the fit unbiased, despite the regularization.

In the following, we present the algorithm and close with a study of the performance of the algorithm on a larger number of toy Monte-Carlo data sets.

2 Basic concepts

ARU is a regularized fit. The unfolded solution is parametrized through a flexible and smooth function $b(x)$ which is forward-folded with the detector kernel $K(y, x)$ and fitted to the data under a soft constrain imposed by a regularization term. Many choices are possible for $b(x)$, but a parametrization with B-splines [9] is particularly suitable. A single B-spline is a unimodal, non-negative and piece-wise polynomial curve with finite support. It has non-zero derivatives up to a degree n and is defined on top

of a grid of m so-called knots x_i. The j-th B-spline is defined by the recursion

$$b_{j,0}(x) = \begin{cases} 1 & \text{if} \quad x_j \leq x < x_{j+1} \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

$$b_{j,n}(x) = \frac{x - x_j}{x_{j+n} - x_j} b_{j,n-1}(x) + \frac{x_{j+n+1} - x}{x_{j+n+1} - x_{j+1}} b_{j+1,n-1}(x), \quad j = 0, \ldots, m+n-2 \tag{2}$$

ARU uses B-splines with $n = 3$ and the index n will be omitted in the following. The solution $b(x)$, a general distribution function, is parametrized as

$$b(x) = \sum_j c_j b_j(x), \tag{3}$$

with coefficients c_j. None of the usual boundary conditions are enforced on the B-spline curve which therefore has $m + 2$ free parameters. The solution $b(x)$ is not normalized; the normalization is also determined from the data. The knots of $b(x)$ need to be narrow enough to pick up all features of the unfolded solution. Above this point, the number of knots and their positions become uncritical. It is not possible to have too many knots, but it will not change the result but increases the computing time of the regularized fit described below.

The solution $b(x)$ is then forward-folded with the detector kernel $K(y, x)$ which quantifies effects of non-linear calibration, efficiency, and limited resolution on the true quantity x

$$f(y) = \int K(y,x) b(x) \mathrm{d}x = \sum_j c_j \int K(y,x) b_j(x) \mathrm{d}x = \sum_j c_j f_j(y), \tag{4}$$

yielding folded basis functions $f_j(y)$ of the folded solution $f(y)$ in the space of the observations. The folded basis functions $f_j(y)$ are computed numerically by the algorithm. Thanks to the linearity of the parametrization, this expensive step needs to be performed only once per application.

The folded solution $f(y)$ is then fitted to the data with the extended maximum-likelihood method [2, 10], i.e., by minimizing the negative log-likelihood function $L_1(\boldsymbol{c})$

$$L_1(\boldsymbol{c}) = \sum_j c_j F_j - \sum_i \ln f(y_i), \tag{5}$$

under the constraint $c_j > 0$, whereas $F_j = \int \mathrm{d}y f_j(y)$ is the total integral of $f_j(y)$ and the y_i denote the observations.

Without an additional constraint, the parameters c_j will have a huge variance and the solution $b(x)$ will be dominated by oscillations that mainly represent noise. In order to compensate for this the combination $L(\boldsymbol{c}) = L_1(\boldsymbol{c}) + wL_2(\boldsymbol{c})$ is minimized for a given w with the regularization term

$$L_2(\boldsymbol{c}) = \int b(x) \ln \frac{b(x)}{g(x)} \mathrm{d}x - \sum_j c_j B_j \tag{6}$$

and $B_j = \int \mathrm{d}x\, b_j(x)$. The regularization term L_2 is a variant of the Kullback-Leibler divergence ($b(x)$ and $g(x)$ are not normalized). If only L_2 is minimized, $b(x)$ approaches $g(x)$. This narrows down the solution space but also introduces a bias. The bias can be reduced by choosing $g(x)$ properly, which will be discussed in the next section.

Minimizing $L(\boldsymbol{c})$ appears to be difficult since L_1 and L_2 are non-linear functions of $\boldsymbol{c}$. However, one can show that the curvature of both terms is always positive and thus only a single global minimum exists. Standard non-linear minimization algorithms with always converge to it, independent of the starting point.

3 Choice of the reference distribution $g(x)$

Eq. (6) vanishes if $g(x)$ is equal to the true solution [2]. Since the true solution is unknown, $g(x)$ can be made only as close as possible to the correct solution. The simplest choice is $g_{(0)}(x)$, a uniform distribution with the correct normalization of the final result

$$g_{(0)}(x) = \frac{1}{x_{m-1} - x_0} \sum_i \epsilon^{-1}(y_i), \tag{7}$$

where $\epsilon(y)$ is the efficiency, y_i are the data points, and x_0 and x_{m-1} the first and last knot positions. With this choice our regularization becomes equivalent to the maximum entropy approach [2, 6].

An iterative approach comes to mind. The unfolding is started with $g_{(0)}(x)$ to obtain a solution $b_{(0)}(x)$, which is then used as $g_{(1)}(x) := b_{(0)}(x)$ to obtain $b_{(1)}(x)$, and so forth. Unfortunately, this approach enhances artificial fluctuations and cannot be used. Instead, we propose to unfold once with $g_{(0)}(x)$ and then use the folded solution $f_{(0)}(y)$ as $g_{(1)}(x)$

$$g_{(1)}(x) = f_{(0)}(\bar{y}(x)) \frac{1}{\epsilon\big(\bar{y}(x)\big)} \frac{\partial \bar{y}}{\partial x} \tag{8}$$

where $\epsilon(y)$ describes the efficiency and $\bar{y}(x)$ the (possibly non-linear) average response of the detector. By doing so, we get an approximation that includes all effects except the resolution.

4 Choice of regularization weight

The choice of the regularization weight w is the choice of the trade-off between bias and variance. We minimize the mean integrated squared error (MISE) of the folded solution $f(y)$ to get an optimal compromise

$$\text{MISE}\big(f(y)\big) = \int \mathrm{d}y \, E\big[\big(f(y) - f_{\text{true}}(y)\big)^2\big] = \int \mathrm{d}y \, \big\{V[f(y)] + \big(f(y) - f_{\text{true}}(y)\big)^2\big\}. \tag{9}$$

The variance $V[f(y)]$ can be derived from the covariance matrix $\boldsymbol{V}[\boldsymbol{c}]$ of the coefficient vector $\boldsymbol{c}$

$$V[f(y)] \simeq \sum_i \sum_j \frac{\partial f(y)}{\partial c_i} \frac{\partial f(y)}{\partial c_j} V[\boldsymbol{c}]_{ij} = \sum_i \sum_j f_i(y) \, f_j(y) \, V[\boldsymbol{c}]_{ij}. \tag{10}$$

The analytical calculation of $\boldsymbol{V}[\boldsymbol{c}]$ is shown in the next section.

What remains is to estimate the bias. Since $f_{\text{true}}(y)$ is unknown, we apply the plug-in principle[1] [11] and replace $f_{\text{true}}(y)$ by the empirical distribution $f_{\text{emp}}(y) = \sum_i \delta(y - y_i)$ of the observations y_i. The empirical distribution $f_{\text{emp}}(y)$ is a maximum likelihood estimate of $f_{\text{true}}(y)$ if no other information is available. With these insights we can transform Eq. (9) after some steps into

$$\text{MISE}\big(f(y)\big) = \sum_k \sum_l \big(V[\boldsymbol{c}]_{kl} + c_k \, c_l\big)\phi_{kl} - 2 \sum_i f(y_i) + \text{const.} \tag{11}$$

The last term does not depend on w and therefore is irrelevant for the minimization. The matrix $\phi_{kl} = \int \mathrm{d}y \, f_k(y) f_l(y)$ is computed numerically once per application of the algorithm.

The minimization of Eq. (11) as a function of the regularization weight w is carried out numerically. The coefficients c_i and the covariance matrix $\boldsymbol{V}[\boldsymbol{c}]$ are re-calculated in each step by minimizing $L(\boldsymbol{c})$.

[1] Physicists use the plug-in principle (unintentionally) whenever they approximate the Poisson uncertainty of a count n by $\sqrt{n}$. In this case the unknown mean λ is replaced by its empirical value n.

5 Uncertainty of the solution

The variance of $b(x)$ is computed from the covariance matrix $\boldsymbol{V}[\boldsymbol{c}]$ of the coefficients analogue to Eq. (10). The covariance matrix is a sum of two contributions, one arising from the statistical term L_1 and one from the regularization term L_2 since we estimate the reference distribution $g(x)$ also from the data

$$\boldsymbol{V}[\boldsymbol{c}] = \boldsymbol{V}_1[\boldsymbol{c}] + \boldsymbol{V}_2[\boldsymbol{c}], \tag{12}$$

but $\boldsymbol{V}_2[\boldsymbol{c}]$ is usually negligible. Its contribution is only expected to be significant in ranges with poor detector efficiency $\epsilon(y)$.

In order to derive $\boldsymbol{V}_1[\boldsymbol{c}]$, we follow Blobel's discussion [1,4] which also gives useful insight into the effect of the regularization. It starts with a Taylor expansion of $L(\boldsymbol{c})$ around a point $\boldsymbol{c}_0$ close to the minimum

$$L(\boldsymbol{c}) \simeq L(\boldsymbol{c}_0) + \boldsymbol{c}^T\left(\boldsymbol{h}_1 + w\boldsymbol{h}_2\right) + \frac{1}{2}(\boldsymbol{c}-\boldsymbol{c}_0)^T\boldsymbol{H}_1(\boldsymbol{c}-\boldsymbol{c}_0) + w\frac{1}{2}(\boldsymbol{c}-\boldsymbol{c}_0)^T\boldsymbol{H}_2(\boldsymbol{c}-\boldsymbol{c}_0), \tag{13}$$

with gradients $\boldsymbol{h}_{1,2}$ and Hesse matrices $\boldsymbol{H}_{1,2}$ of $L_{1,2}(\boldsymbol{c})$ evaluated at $\boldsymbol{c}_0$.

Eq. (13) can be simplified by dropping all constant terms, this does not change the position of the minimum. Since $\boldsymbol{H}_1$ and $\boldsymbol{H}_2$ are symmetric, we can simplify to

$$L(\boldsymbol{c}) = \boldsymbol{c}^T\left(\boldsymbol{h}_1 + w\boldsymbol{h}_2 - \left(\boldsymbol{H}_1 + w\boldsymbol{H}_2\right)\boldsymbol{c}_0\right) + \frac{1}{2}\boldsymbol{c}^T\boldsymbol{H}_1\boldsymbol{c} + w\frac{1}{2}\boldsymbol{c}^T\boldsymbol{H}_2\boldsymbol{c}. \tag{14}$$

We now change into another coordinate system with the transformation matrix $\boldsymbol{M}$ in which Eq. (14) takes its simplest form

$$\boldsymbol{c}^T = \bar{\boldsymbol{c}}^T\boldsymbol{M}^T = \bar{\boldsymbol{c}}^T\boldsymbol{U}_1^T\boldsymbol{D}_2^{-1/2}\boldsymbol{U}_2^T. \tag{15}$$

The matrices $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$ are rotations. The first rotation $\boldsymbol{U}_2$ is chosen such that $\boldsymbol{H}_2$ becomes the diagonal matrix $\boldsymbol{D}_2$. The matrix $\boldsymbol{D}_2^{-1/2}$ is also diagonal and defined as

$$D_{2_{jj}}^{-1/2} = 1\Big/\sqrt{D_{2jj}}. \tag{16}$$

This scaling transformation turns $\boldsymbol{D}_2$ into the unit matrix. The matrix $\boldsymbol{D}_2^{-1/2}$ always exists, because all diagonal elements of $\boldsymbol{D}_2$ are positive. The unit matrix is invariant to any further rotation. These transformations applied to $\boldsymbol{H}_1$ lead to a another symmetric and positive-definite matrix $\tilde{\boldsymbol{H}}_1$. The last rotation $\boldsymbol{U}_1$ is chosen such that $\tilde{\boldsymbol{H}}_1$ turns into the diagonal matrix $\boldsymbol{S}$, and so we get

$$L(\bar{\boldsymbol{c}}) = \bar{\boldsymbol{c}}^T\boldsymbol{M}^T\left(\boldsymbol{h}_1 + w\boldsymbol{h}_2 - \left(\boldsymbol{H}_1 + w\boldsymbol{H}_2\right)\boldsymbol{c}_0\right) + \frac{1}{2}\bar{\boldsymbol{c}}^T\boldsymbol{S}\,\bar{\boldsymbol{c}} + w\frac{1}{2}\bar{\boldsymbol{c}}^T\bar{\boldsymbol{c}}. \tag{17}$$

The minimum of $L(\bar{\boldsymbol{c}})$ can now be calculated by solving $\nabla L(\bar{\boldsymbol{c}}) = 0$. The solution $\bar{\boldsymbol{c}}$ is compactly expressed as a combination of the two solutions $\bar{\boldsymbol{c}}_1$ and $\bar{\boldsymbol{c}}_2$ of the unregularized problem ($w = 0$, $L \equiv L_1$) and the purely regularized problem ($w \to \infty$, $L \equiv L_2$), respectively:

$$\bar{\boldsymbol{c}}_1 = \boldsymbol{S}^{-1}\boldsymbol{M}^T(\boldsymbol{H}_1\boldsymbol{c}_0 - \boldsymbol{h}_1) \tag{18}$$

$$\bar{\boldsymbol{c}}_2 = \boldsymbol{M}^T(\boldsymbol{H}_2\boldsymbol{c}_0 - \boldsymbol{h}_2) \tag{19}$$

$$\bar{\boldsymbol{c}} = (\boldsymbol{S} + w\boldsymbol{1})^{-1}\left(\boldsymbol{S}\bar{\boldsymbol{c}}_1 + w\bar{\boldsymbol{c}}_2\right). \tag{20}$$

It turns out that the transformed solution $\bar{\boldsymbol{c}}$ is a component-wise linear interpolation of the two extreme cases, since $\boldsymbol{S}$ is diagonal. The mixture for each $\bar{c}_i$ depends on the relative size of the weight w and $S_{ii} = \sigma_i^{-2}$, the inverse of the variance of the corresponding coefficient $\bar{c}_{1i}$. The regularization weight w effectively dampens coefficients with a large variance ($S_{ii} \ll w$).

Eq. (15) and Eq. (20) show the relation between the coefficients $\bar{c}_{1k}$, for which the variance S_{kk}^{-1} is known, and the coefficients c_i. We can obtain after successive error propagation

$$V_1[\boldsymbol{c}]_{ij} = \sum_k \frac{\partial c_i}{\partial \bar{c}_k} \frac{\partial \bar{c}_k}{\partial \bar{c}_{1k}} \frac{\partial c_j}{\partial \bar{c}_k} \frac{\partial \bar{c}_k}{\partial \bar{c}_{1k}} S_{kk}^{-1} = \sum_k M_{ik} M_{jk} \frac{S_{kk}}{(S_{kk} + w)^2}. \tag{21}$$

The second contribution $\boldsymbol{V}_2[\boldsymbol{c}]$ to the total variance is also obtained from a different kind of error propagation. It starts with the known variance $\boldsymbol{V}[\boldsymbol{d}]$ of the coefficient vector $\boldsymbol{d}$ of $g(x) = \sum_l d_l b_l(x)$ and uses Eq. (A.2) from the appendix. The change $\delta \boldsymbol{H}$ after a variation $\delta \boldsymbol{d}$ is conveniently zero. For $\delta \boldsymbol{h}$, we get

$$\delta h_k = \frac{\partial h_k}{\partial d_l} \delta d_l = -\int \mathrm{d}x \, \frac{b_k(x)\, b_l(x)}{g(x)} \delta d_l, \tag{22}$$

and thus finally obtain

$$V_2[\boldsymbol{c}]_{ij} = \sum_k \sum_l \sum_m \sum_p H_{ik}^{-1} H_{jl}^{-1} \frac{\partial h_k}{\partial d_m} \frac{\partial h_l}{\partial d_p} V[\boldsymbol{d}]_{mp}. \tag{23}$$

6 Monte-Carlo study

The method is demonstrated with a simple toy Monte-Carlo. The true distribution $t(x)$ is given by the sum of two Gaussians $N(\mu, \sigma)$ in the range $x \in [0, 1]$

$$t(x) = 0.3N(0.3, 0.1) + 0.7N(0.7, 0.05), \tag{24}$$

which form a shallow peak next to a sharp peak. The true distribution $t(x)$ is smeared out with a Gaussian kernel $K(y, x) = N(x - y, 0.1)$. ARU is applied to this data using 20 evenly spaced knots in the interval $[0, 1]$. This number is enough to pick up all features of the unfolded solution and further increasing the number does not change the result. Figure 1 shows the unfolding applied to a sample of 1000 events. The method picks up the true features and does not introduce artificial ones. Some bias is visible and expected, but it is of the same order as the estimated statistical uncertainty.

In order to show ARU's average performance, 2000 data sets are generated and unfolded, half of them with 100 events and the other half with 10000 events each. The results are shown in Fig. 2. The estimated uncertainty agrees well with the observed standard deviation. The average bias is comparable to the statistical uncertainty of the solution.

7 Conclusions and outlook

We presented ARU, an automatic algorithm to unfold detector effects from one-dimensional distributions. ARU tries to improve existing algorithms in several ways. It is a completely unbinned approach with a regularization term that is a variant of the Kullback-Leibler divergence between the solution and a reference distribution. The reference distribution is a smooth approximation to the final solution, obtained by correcting the data only for calibration and efficiency effects. We argue that this choice reduces the bias with respect to a maximum entropy regularization which we generalized with our approach. Our algorithm is based only on statistical information and therefore invariant to transformations of the data. The optimal regularization weight is found by minimizing the mean integrated squared error of the solution. An application of the algorithm to a large sample of toy data sets demonstrates the correct estimation of the statistical uncertainty and a regularization bias that is of the order of the statistical uncertainty.

The source code of ARU can be downloaded from Hepforge [12]. Future work will focus on the generalization of the algorithm to multi-dimensional distributions.

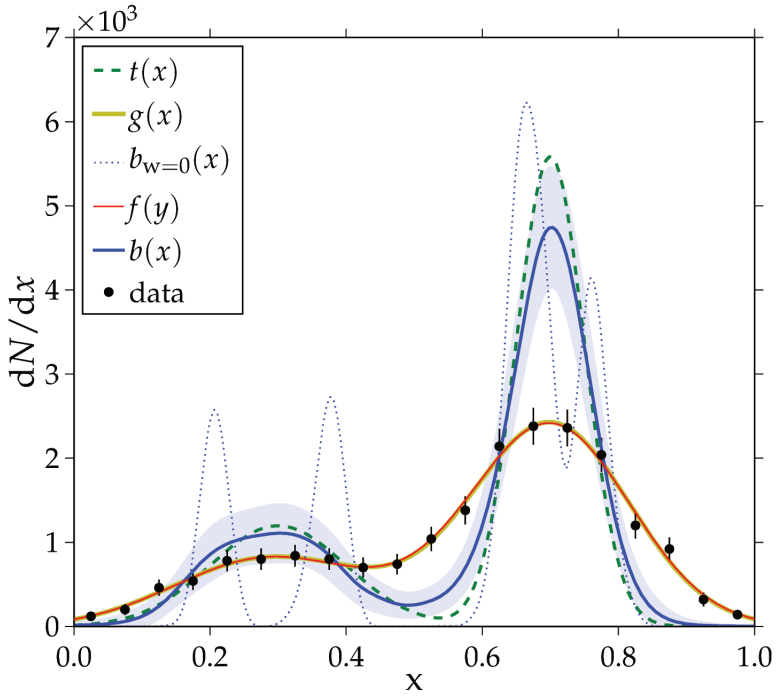


Figure 1: Unfolding of a toy data set of 1000 events. $t(x)$ is the true distribution, the points show a histogram of the smeared data. In this case, the folded solution $f(y)$ is on top of the reference distribution $g(x)$ used for regularization. The regularized solution $b(x)$ shows no undesired oscillations, in contrast to the solution $b_{w=0}(x)$, which is obtained if no regularization is applied.

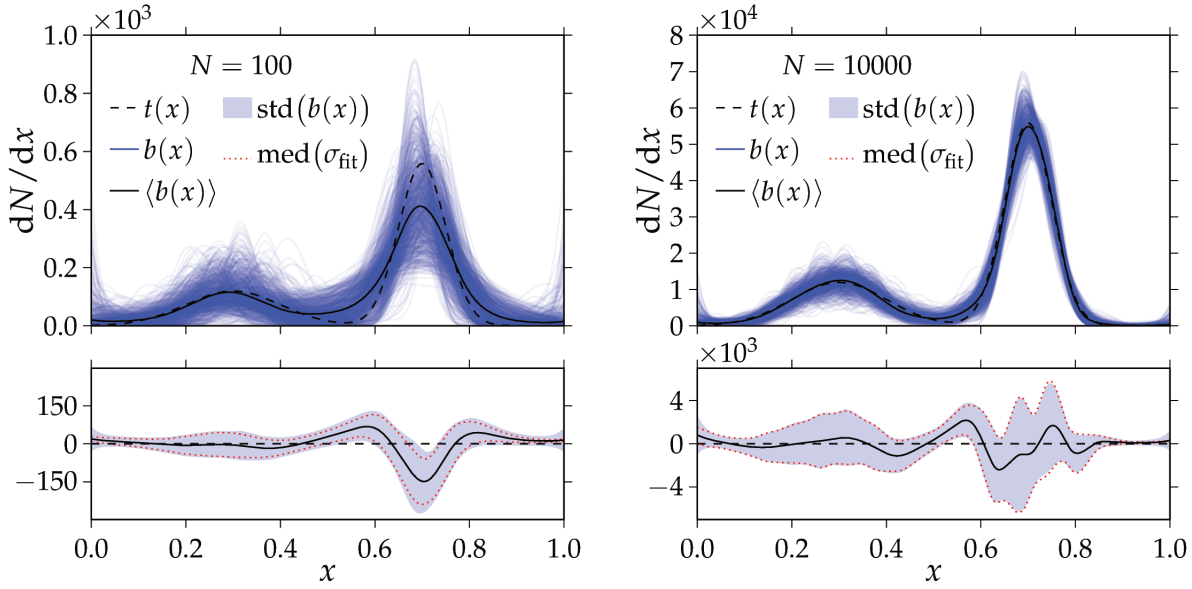


Figure 2: Monte-Carlo study of the unfolding method applied 1000 times to random toy data sets of 100 events (left) and 10000 events (right) each. The top plots show the true distribution $t(x)$ and the unfolded solutions $b(x)$ plotted transparently on top of each other. The bottom plots show the bias $\langle b(x) - t(x) \rangle$ of the unfolding, the standard deviation $\mathrm{std}\big(b(x)\big)$ of the unfolded set, and the median $\mathrm{med}(\sigma_{\mathrm{fit}})$ of the estimated uncertainty of the solution.

Acknowledgements

We are indebted to M. Schmelling for valuable discussions about unfolding concepts and his friendly support of this work.

Bibliography

[1] V. Blobel and E. Lohrmann, *Statistische und numerische Methoden der Datenanalyse* (Teubner Verlag, Wiesbaden, Germany, 1998).

[2] G. Cowan, *Statistical Data Analysis* (Clarendon Press, Oxford, 1998).

[3] R. Gold, Technical Report No. ANL-6984, Argonne National Laboratory, Argonne, Illinois (1964).

[4] V. Blobel, *Unfolding methods in high energy physics experiments*, Proceedings of the 1984 CERN School of Computing (1984).

[5] G. D'Agostini, Nucl. Instrum. Meth. **A362**, 487 (1992).

[6] M. Schmelling, Nucl. Instrum. Meth. **A340**, 400 (1994).

[7] A. Hoecker and V. Kartvelishvili, Nucl. Instrum. Meth. **A372**, 469 (1996), arXiv hep-ph/9509307v2.

[8] S. Kullback and R. A. Leibler, Ann. Math. Stat. **22**, 79 (1951).

[9] C. de Boor, *A Practical Guide to Splines* (Springer Verlag, New York, Heidelberg, Berlin, 1978).

[10] R. Barlow, Nucl. Instrum. Meth. **A297**, 496 (1990).

[11] B. Efron and R. Tibshirani, *An introduction to the bootstrap* (Chapman & Hall, London, 1993).

[12] `http://projects.hepforge.org/aru`.

Appendix

A Propagation of model uncertainties into a maximum-likelihood estimate

We will derive a formula that allows to propagate small changes in the likelihood function, for example due to systematic variations in the explanatory model, into its maximum likelihood estimate.

Let $L(\boldsymbol{c})$ be a general log-likelihood function. We expand it in a Taylor series up to second order with gradient $\boldsymbol{h}$ and Hesse matrix $\boldsymbol{H}$ around a point $\boldsymbol{c}_0$. If $\boldsymbol{c}_0$ is close to the minimum of $L(\boldsymbol{c})$, the Taylor series will be a good approximation for $L(\boldsymbol{c})$ and the position of the minimum $\boldsymbol{c}$ can be analytically calculated as

$$\boldsymbol{c} = \boldsymbol{c}_0 - \boldsymbol{H}^{-1}\boldsymbol{h}. \tag{A.1}$$

We now regard a slightly changed $\tilde{L} = L + \delta L$ with according changes in gradient $\tilde{\boldsymbol{h}} = \boldsymbol{h} + \delta\boldsymbol{h}$ and Hesse matrix $\tilde{\boldsymbol{H}} = \boldsymbol{H} + \delta\boldsymbol{H}$. If $\boldsymbol{c}$ is the minimum of the undisturbed function $L(\boldsymbol{c})$, Eq. (A.1) can be used to calculate the corresponding shift in the solution generated by the disturbance:

$$\begin{aligned}
\delta\boldsymbol{c} = \tilde{\boldsymbol{c}} - \boldsymbol{c} &= -\tilde{\boldsymbol{H}}^{-1}\tilde{\boldsymbol{h}} \\
&= -(\boldsymbol{H} + \delta\boldsymbol{H})^{-1}(\boldsymbol{h} + \delta\boldsymbol{h}) \\
&= -(\boldsymbol{H}^{-1} - \boldsymbol{H}^{-1}\boldsymbol{H}^{-1}\delta\boldsymbol{H} + O(\delta\boldsymbol{H}^2))\delta\boldsymbol{h} \\
&= \boldsymbol{H}^{-1}(\boldsymbol{H}^{-1}\delta\boldsymbol{H} - \mathbf{1})\delta\boldsymbol{h} + O(\delta\boldsymbol{H}^2\delta\boldsymbol{h}),
\end{aligned} \tag{A.2}$$

where the Taylor expansion of $\tilde{\boldsymbol{H}}^{-1}$ was used and the fact that $\boldsymbol{h}$ vanishes at $\boldsymbol{c}$. The last term $O(\delta\boldsymbol{H}^2\delta\boldsymbol{h})$ is of higher order and can be neglected for small disturbances in $L(\boldsymbol{c})$. The truncated formula is a powerful tool which allows to propagate systematic uncertainties of the explanatory model into the result without resorting to Monte-Carlo techniques in many cases. An example application is shown in Section 5.

Unfolding at CMS

Matthias Weber on behalf of the CMS Collaboration
Institute for Particle Physics, ETH Zurich, 8093 Zurich, Switzerland

Abstract
The unfolding techniques used by the CMS collaboration on the 2010 data analyses are presented. Each method is discussed on the basis of an experimental measurement. The main focus is on studying the sensitivity to different models used in the determination of the response matrix and the propagation of statistical errors.

1 Unfolding of experimental distributions

In order to allow a *direct* comparison of experimental measurements with theoretical predictions, the measurements must be unfolded for detector effects. This also permits the direct comparison of distributions from different experiments without knowledge of the detector response for each experiment. Moreover, the automated tuning of Monte Carlo generators using multiple measurements is greatly facilitated using unfolded data. In principle the measurements could be presented without corrections for detector effects together with a detector response matrix. The smearing of the theory distribution with a published response matrix for each single measurement and each experiment is far more complicated in Monte Carlo tuning efforts.

The measurements discussed here are based on data collected in proton-proton collisions with the Compact Muon Solenoid (CMS) experiment at the Large Hadron Collider (LHC). A detailed description of the CMS detector can be found elsewhere [1]. The detector response matrix R in the unfolding procedure is usually derived using simulated Monte Carlo (MC) samples. The CMS detector response is derived via simulation based on GEANT4 [2]. The inversion of the response matrix can lead to unacceptable solutions, since small statistical fluctuations can lead to large effects in the solution. Even negative entries in the unfolded distribution are possible. This oscillating behaviour can be reduced by imposing the requirement that the true distribution is smooth. This smoothing procedure is known as regularization. The regularized unfolding methods used and investigated by CMS are iterative Bayesian unfolding [3], the SVD method [4] and Tikhonov regularization [5].

2 Iterative Unfolding: Charged Particle Multiplicities

The iterative Bayesian unfolding method is used in the measurement of the charged particle multiplicities [6]. The observed charged multiplicity $O = (O_1, \ldots, O_N)$ will in general be different from the true multiplicity $T = (T_1, \ldots, T_M)$ due to track reconstruction inefficiencies, the presence of secondary particles and decay products of long-lived hadrons. In this method the inversion is done in a stepwise iterative procedure. The unfolded distribution for the iteration step k is:

$$T_j^{(k)} = \sum_i \frac{R_{ij} T_j^{(k-1)}}{\sum_s R_{il} T_s^{(k-1)}} \cdot O_i. \tag{1}$$

After each iteration k the χ^2 between $T^{(k)}$ and $T^{(k-1)}$ is calculated. The procedure is stopped once the χ^2 converges and a stable solution T is found. The response matrix R is derived from samples generated with PYTHIA6 [7] with the tune D6T. It is checked that the final solution T does not depend on initial ansatz for the true distribution, using in one case a flat distribution and in the other case the MC generator level distribution. In a second step the robustness of the unfolding procedure is tested, using the PYTHIA6 DW and a PHOJET [8] response matrix to unfold pseudodata generated by PYTHIA6 tune D6T.

The observed differences shown in Fig. 1 are small over a large range, for higher multiplicity bins they are dominated by statistical fluctuations.

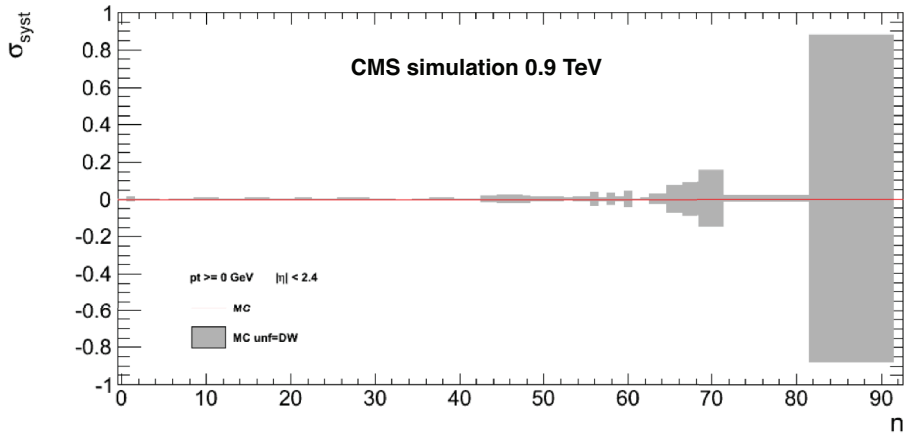


Fig. 1: The robustness of the unfolding procedure, unfolding PYTHIA6 D6T pseudodata distribution of the charged particle multiplicity n with the PYTHIA6 DW response matrix. σ_{syst} represents the relative difference between the unfolded result and MC thruth.

The covariance matrix of the unfolded spectrum is derived using a resampling technique. As shown in Fig. 2 the statistical errors are very dependent on the number of iterations. The errors increase as a function of the iterations, while the statistical bias decreases. For a small number of iterations the errors are smaller than $\sqrt{N}$, where N is the number of entries.

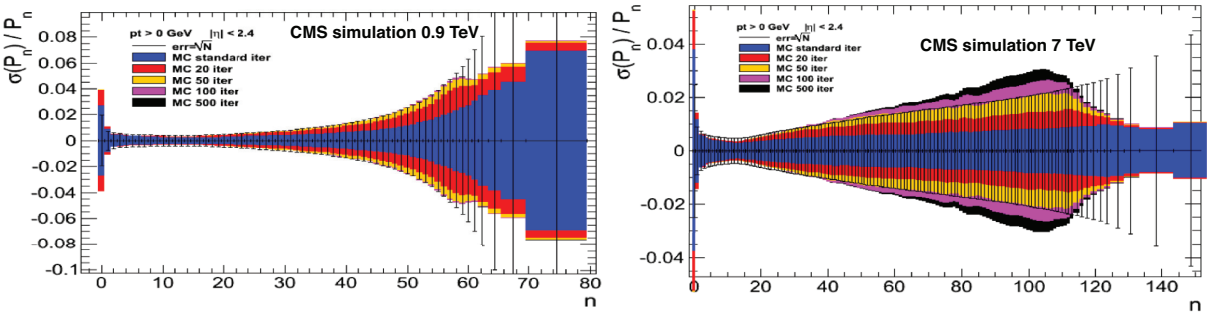


Fig. 2: The dependence of the relative statistical errors of the charged multiplicity spectrum on the number of iterations in the unfolding. The pseudodata is simulated with PYTHIA6 at $\sqrt{s} = 900$ GeV (left) and $\sqrt{s} = 7$ TeV (right).

3 SVD Unfolding: Hadronic Event Shapes

The SVD method is a special implementation of the Tikhonov regularization method, based on the singular value decomposition of the response matrix. The regularization method works as low-pass filter and suppresses small singular values, which would lead to oscillating behaviour. This method is used in the measurement of hadronic event-shape variables in 7 TeV proton-proton jet data [9]. The variable which we will use in the following is the central transverse thrust, which is defined as

$$\tau_{\perp,\mathcal{C}} \equiv 1 - \max_{\hat{n}_{\mathrm{T}}} \frac{\sum_i |\vec{p}_{\perp,i} \cdot \hat{n}_{\mathrm{T}}|}{\sum_i p_{\perp,i}}, \tag{2}$$

where $p_{\perp,i}$ are the jet transverse momenta. Well balanced dijet events have low thrust values close to 0, spherical multijet events have high values. The measurement is performed in several bins of the leading

jet transverse momentum $p_{\perp,1}$. The unfolded measured distributions are compared to predictions from the MC generators PYTHIA6, HERWIG++ [10], ALPGEN [11], MADGRAPH [12] and PYTHIA8 [13]. The measurement can be used in further tuning of MC generators.

The jet energy and jet position resolutions of the detector distort the event-shape distributions. Especially in the lower range of event shape values the off-diagonal elements of the response matrix are sizable (20-30%); only for higher event-shape values the diagonal element is around 90%. The response matrix is determined using simulated PYTHIA6 D6T QCD events. The consistency between the unfolded pseudodata distribution and the MC generator level distribution is checked for each generator using in all cases the PYTHIA6 D6T response matrix. Fig. 3 shows that for all other generators a good closure can be observed. The regularization parameter is chosen such that the χ^2 value between the unfolded and the generator level distribution is minimal (In this test we consider all generators but PYTHIA6). The full covariance matrix is used in the χ^2 calculation.

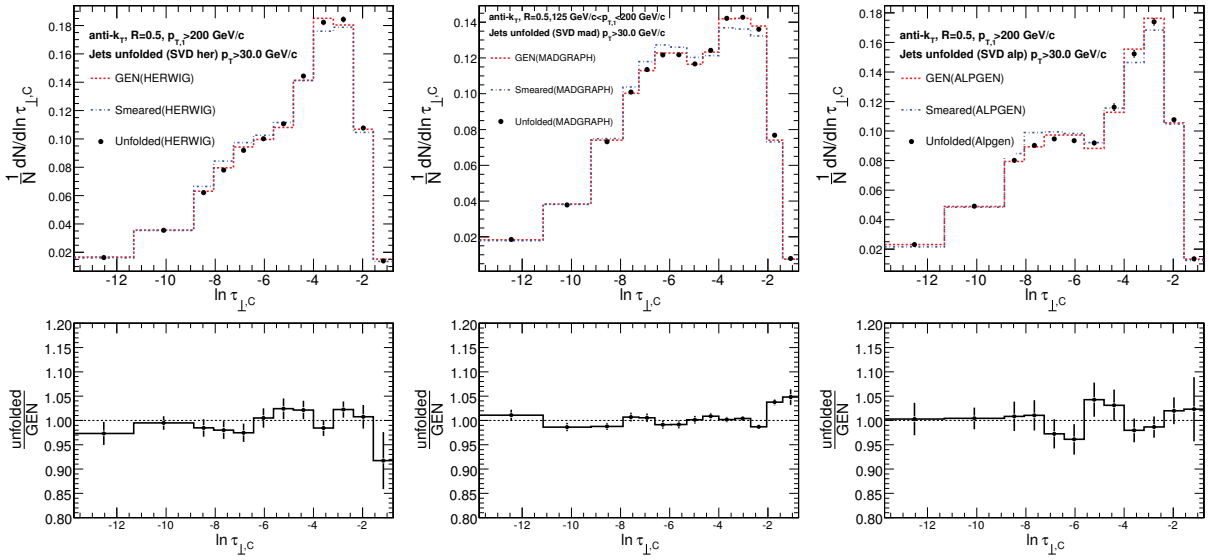


Fig. 3: The closure of the SVD unfolding procedure for the central transverse thrust distribution for HERWIG++ (left), MADGRAPH (middle) and ALPGEN (right) pseudodata. The ratios show the deviations from the generator level distribution.

Unfolding the data with a response matrix determined from MADGRAPH instead of PYTHIA6 gives consistent results. We check that no preference for one of the generators is introduced by the unfolding procedure. In a first step the χ^2 between the simulated pseudodata and the data distribution prior to unfolding is calculated. These χ^2 values are compared to χ^2 values between the generator level distributions and the data distributions after unfolding. The ordering is the same before and after unfolding and the values are similar. The iterative Bayesian unfolding is applied as a further cross-check. The resulting unfolded distribution agrees within 1% with the distribution of the SVD unfolding.

The covariance matrix of the SVD method is non diagonal with large bin-to-bin correlations in the errors (Fig. 4, left). The statistical error of a bin i is taken as the square root of corresponding diagonal element of the covariance matrix $\sqrt{C_{ii}}$. As illustrated in Fig. 4, right using PYTHIA6 pseudodata, the relative statistical errors prior to unfolding are for some regularization choices bigger than the relative statistical errors after the unfolding for almost all bins. With stronger regularization (small regularization parameter, e.g. Reg. 7), the errors can be smaller than for the raw data. The correlation between the bins of the unfolded distribution is bigger and the diagonal element of the covariance matrix smaller. In those

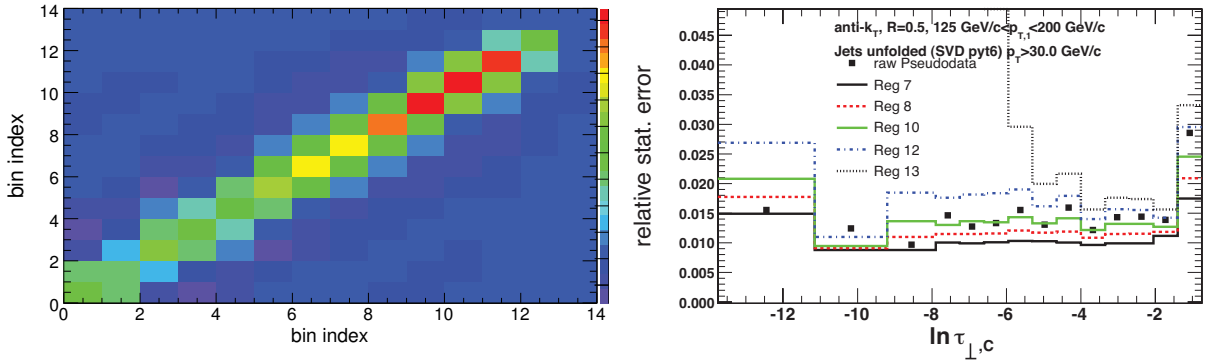


Fig. 4: The covariance matrix after unfolding using as regularization value 7 (left). Comparison of the relative statistical errors prior to unfolding and after the unfolding using several strengths of regularization (right). In both cases PYTHIA6 has been used to generate the pseudodata.

cases the statistical error is smaller than $\sqrt{N}$, where N is the bin entry in these bins after unfolding the distributions. For small regularization (higher regularization value. e.g Reg. 13) the errors approximate the errors of the matrix inversion and can be sizable.

The jet energy resolution uncertainty of 10% is treated in the measurement as uncertainty in the response matrix, i.e. the unfolding is repeated with a new response matrix determined from PYTHIA6 D6T with jet energies smeared by an additional 10%.

4 Tikhonov Regularization: Inclusive Jet Cross Section

The Tikhonov regularization method is investigated in the context of the inclusive jet cross section measurement. The true spectrum is distorted by the finite detector energy resolution. The inclusive jet transverse momentum spectrum is a steeply falling distribution covering a range of many orders of magnitude and thus more challenging than e.g. event shape distributions. The determination of the response matrix is difficult, since a huge phase space with dramatically different cross sections needs to be covered with sufficient statistics in all corners. A further complication is the fact that one particle level jet might be reconstructed as two jets in the detector or vice versa. The response matrix is thus calculated using a theory curve and smearing it with measured jet resolution functions. Cross-checks on Monte Carlo show that the solution of the Tikhonov method also depends on the choice of the number of bins for theory and data. The solution is found using the quasi-optimal approach [14]. In this approach the regularization parameter τ is varied in fine steps starting with larger parameters τ. For each step k the maximum deviation between the contents of all bins j of the unfolded histograms $\Delta(\tau) = \max_j |O_j(\tau_k) - O_j(\tau_{k-1})|$ is determined. Minima of Δ are stable solutions; the first minima depends on starting conditions of the iteration procedure and should be disregarded. In general the deepest minimum is preferred as solution, since it corresponds to the most stable solution as a function of the regularization parameter τ.

The solution of the method shows a good closure using the deepest minimum. The effective input unfolding correction factor is correctly reproduced by this solution. The error propagation of the regularized solution by the matrix inversion can lead to large estimates of the statistical uncertainties as shown in Fig. 5. In data the inclusive jet spectrum uses several jet trigger streams, involving also heavily prescaled low p_{T} triggers. This can lead to artefacts in the unfolded distribution and the error propagation. The Tikhonov regularization is also used in the measurement of jet shapes.

5 Summary and Conclusion

Several unfolding methods used in CMS 2010 data analyses are presented: iterative Bayesian unfolding, SVD unfolding, Tikhonov regularization and matrix inversion. The usual tests performed in the unfolding

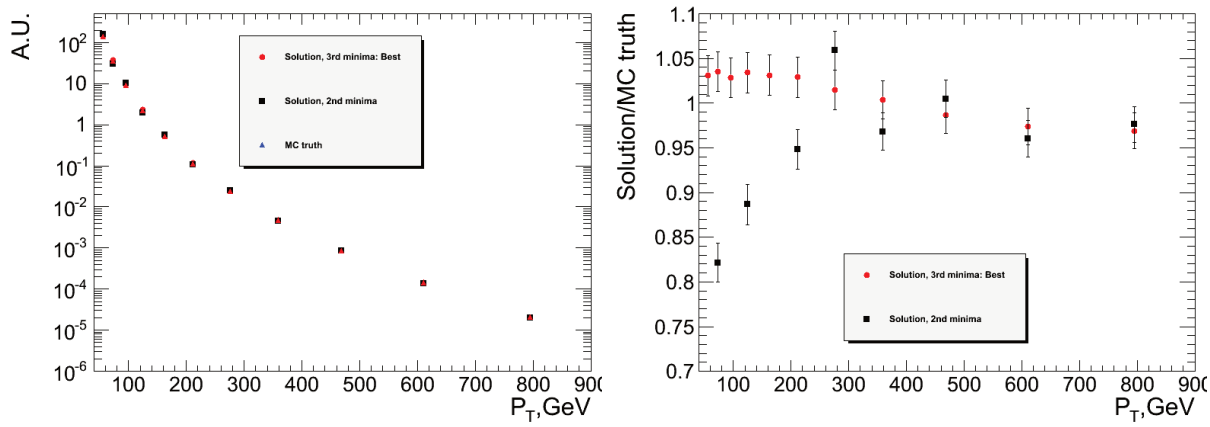


Fig. 5: The closure of the unfolding correction in the analysis of the inclusive jet cross section. On the left the MC truth spectrum and two unfolded spectra are shown. The ratio between the unfolded solutions and MC truth is shown on the right.

procedure involve closure tests and the model dependency. Uncertainties in the modelling of the response matrix are examined. The interpretation of the error propagation is discussed, especially the fact that errors are sometimes smaller than $\sqrt{N}$.

References

[1] CMS Collaboration, "The CMS experiment at the CERN LHC", *JINST* **3** (2008) S08004.

[2] S. Agostinelli et al., "GEANT4: A simulation toolkit", *Nucl. Instrum. Meth.*, **A**506 (2003) 250.

[3] G. D'Agostini, "A Multidimensional unfolding method based on Bayes' theorem", *Nucl. Instrum. Meth.* **A362** (1995) 487.

[4] A. Höcker and V. Kartvelishvili, "SVD Approach to Data Unfolding", *Nucl. Instrum. Meth.* **A372** (1996) 46.

[5] A.N. Tikhonov, "On the solution of improperly posed problems and the method of regularization", *Sov. Math.* **5** (1964) 1035.

[6] CMS Collaboration, "Charged particle multiplicities in pp interactions at $\sqrt{s}$ =0.9, 2.36, and 7 TeV", *JHEP* **1** (2011) 79.

[7] T. Sjöstrand, S. Mrenna, and P. Z. Skands, "PYTHIA 6.4 Physics and Manual", *JHEP* **05** (2006) 026.

[8] R. Engel and J. Ranft, "Hadronic photon-photon interactions at high-energies", *Phys. Rev.* **D54** (1996) 4244, arXiv:hep-ph/9509373.

[9] CMS Collaboration, "Measurement of Hadronic Event Shapes in pp Collisions at $\sqrt{s} = 7$ TeV", *Phys. Lett. B* **699** (2011) 48.

[10] M. Bahr et al., "Herwig++ Physics and Manual", *Eur. Phys. J.* **C58** (2008) 639.

[11] M.L. Mangano et al., "ALPGEN, a generator for hard multiparton processes in hadronic collisions", *JHEP* **07** (2003) 001.

[12] J. Alwall et al., "MadGraph/MadEvent v4: The New Web Generation", *JHEP* **09** (2007) 028.

[13] T. Sjöstrand, S. Mrenna, and P. Z. Skands, "A Brief Introduction to PYTHIA8.1", *Comput. Phys. Commun.* **178** (2008) 852.

[14] V.B. Glasko, "Inverse problems of Mathematical Physics", *American Institute of Physics translation series* (1988). (2008) 852.

Unfolding in ATLAS

Georgios Choudalakis
University of Chicago, on behalf of the ATLAS Collaboration

Abstract
This article presents the unfolding techniques used so far in ATLAS. Two representative examples are discussed in detail; one using bin-by-bin correction factors, and the other iterative unfolding.

1 Introduction

The distribution of any observable is distorted due to experimental limitations. Unfolding is the procedure of estimating the "truth-level" spectrum, i.e., the spectrum that would be measured with an ideal detector and infinite event statistics. A general introduction to unfolding, and details about various methods are given in other contributions to this workshop. The focus here will be on real life examples of unfolding in ATLAS analyses.

As of early 2011, ATLAS has used two unfolding methods:

i) bin-by-bin correction factors;
ii) the iterative method by D'Agostini [1].

One representative example will be presented from each method. In Section 2, bin-by-bin correction is presented through the inclusive jet p_T spectrum measurement [2]. In Section 3, D' Agostini's iterative method [1] is presented, as it was used to estimate the spectrum of charged particle multiplicity in minimum bias interactions [3].

Both methods have drawbacks. An insightful overview can be found in [4], and in other contributions to this workshop. Bin-by-bin correction has been particularly criticized for not dealing carefully with bin correlations, among other things. ATLAS is considering methods beyond bin-by-bin in the next round of analyses where this method was used.

In searches for new physics, ATLAS does not apply any unfolding, because it is unnecessary for making a discovery, or for setting a limit to some model, or for estimating model parameters. Unfolding can be regarded as useful when the distribution itself (or a binned version thereof) is regarded as the set of parameters of interest.

2 Bin-by-bin correction factors

Several ATLAS analyses have used the method of bin-by-bin correction factors [2, 5–7], mostly because of its simplicity. The example of inclusive jet p_T measurement [2] will be discussed. The main result of this measurement is shown in Fig. 1. In this analysis truth-level corresponds to hadron-level.

2.1 Method description

Let T_i be the expected number of events in bin i of the truth-level p_T spectrum, which is obtained from Monte Carlo (MC). Leading order PYTHIA [8] QCD MC was used in the case of [2], where no event selection was applied. The truth-level p_T spectrum contains jets reconstructed after hadronization, applying the anti-k_T clustering algorithm on stable hadrons produced after fragmentation and hadronization. Detector simulation is not involved in the truth-level spectrum.

Let R_i be the expected number of events in bin i of the measured p_T spectrum, which suffers from detector smearing, after event selection which includes trigger requirements, jet reconstruction

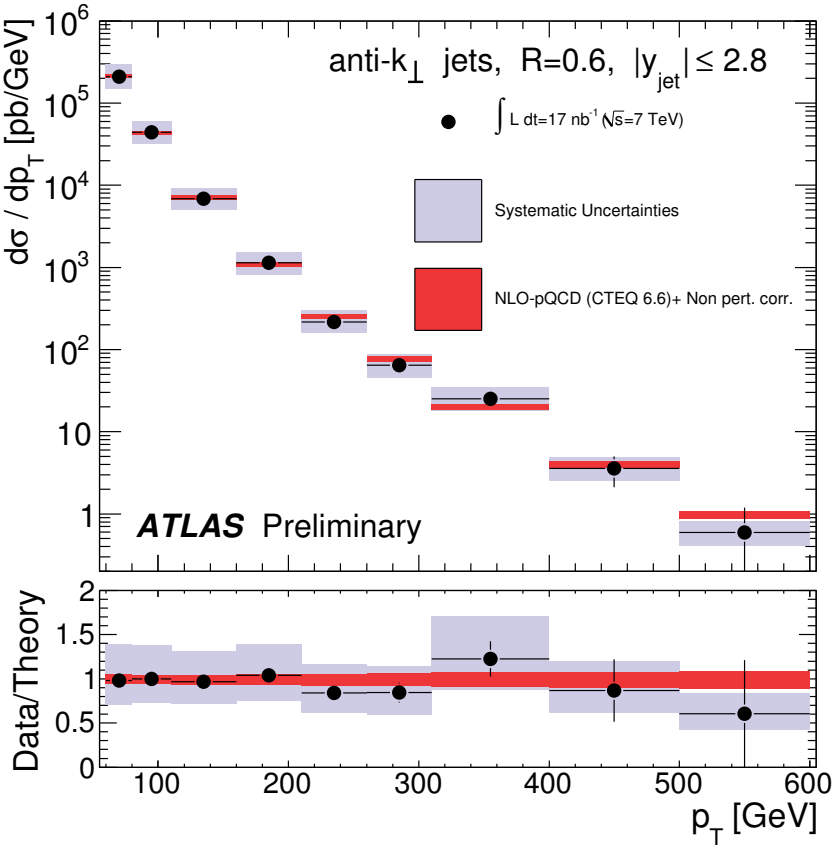


Fig. 1: The estimated truth-level spectrum of inclusive jet p_T (filled markers) from [2], obtained using bin-by-bin correction factors, compared to the theoretical truth-level QCD prediction (red band). The black error bars represent the statistical uncertainty of the estimated spectrum, and the blue band the total systematic uncertainty, which is obtained by summing in quadrature individual systematic uncertainties. The dominant contribution comes from the jet energy scale uncertainty. In each bin the estimated truth-level spectrum has been divided by the width of the bin and by the integrated luminosity, whose uncertainty (11%) is not included in the blue error band.

inefficiency at low p_T, primary vertex requirements, jet quality criteria etc. The same PYTHIA QCD MC is used as before, after ATLAS detector simulation, to obtain R_i. Jets are reconstructed by applying the same anti-k_T algorithm on topological clusters of energy deposited in the calorimeter [9].

Let D_i be the actually observed number of events in bin i of the measured p_T spectrum. Whereas T_i and R_i are both real numbers after normalizing the MC samples to the integrated luminosity of the available dataset, D_i can only take integer values, because the observed events are discrete. If it is assumed that R_i is the result of an ideal simulation of all physical processes that occur at the proton collisions[1] and of the ATLAS detector, then D_i is a random integer that follows a Poisson distribution with mean R_i.

$$C_i \equiv \frac{T_i}{R_i}, \tag{1}$$

be the correction factor corresponding to bin i of the observed p_T spectrum. The correction factors used in [2] are shown in Fig. 2.

[1] Obviously this is not a good assumption when one acknowledges the possibility of new physics, but in measurements such as the one we discuss here it is presumed that what is measured is just QCD.

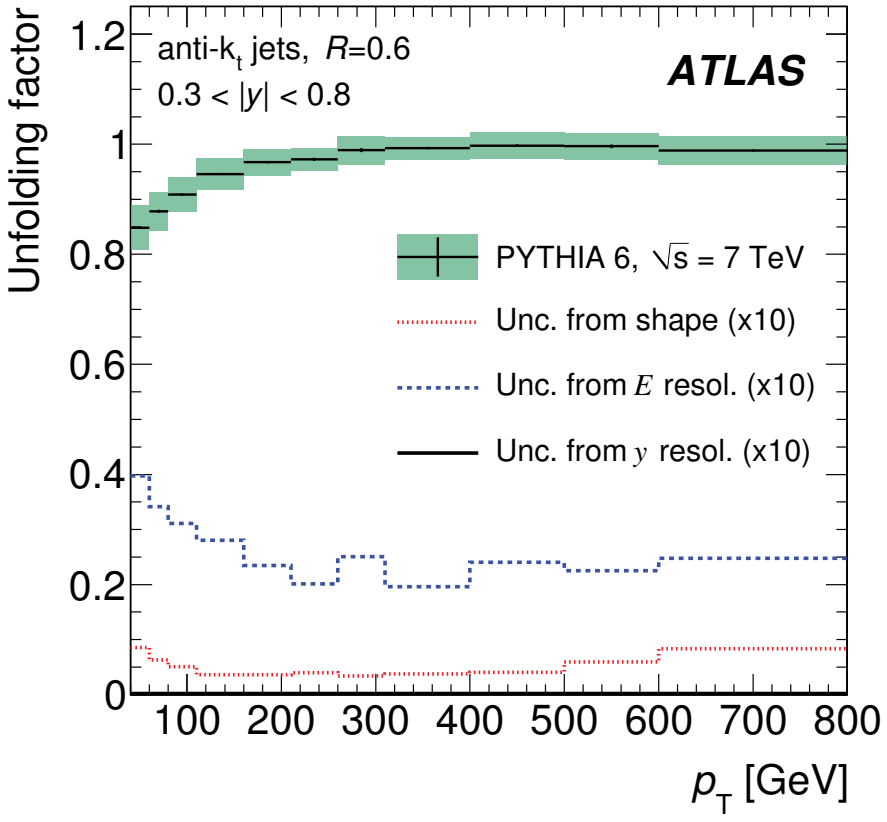


Fig. 2: The correction factors (C_i) used in [2]. The statistical uncertainties (black crosses) are invisibly small. The green band represents the total systematic uncertainty, except for the part which is due to jet energy scale, which is discussed in Section 2.3.4.

The answer returned for bin i of the truth-level p_T spectrum after bin-by-bin correction is

$$U_i \equiv C_i \cdot D_i. \tag{2}$$

U_i is the estimator of T_i.

2.1.1 Bias

The estimator U_i has a bias that is easy to compute.

Let's consider the possibility that the truth-level spectrum is actually T_i', which may differ from the assumed T_i. This could happen, for example, if sizable processes other than those included in PYTHIA QCD are occurring in nature, or if the modeling of QCD by PYTHIA is unrealistic. Let's also assume that the actual expected spectrum at detector level is R_i', which may differ from R_i for the above reasons, as well as due to unrealistic modeling of the detector response and of the quantities involved in event selection. The bias of the estimator U_i then is

$$\langle U_i - T_i' \rangle = \langle \frac{T_i}{R_i} D_i - T_i' \rangle = \frac{T_i}{R_i} \langle D_i \rangle - T_i' = \frac{T_i}{R_i} R_i' - T_i' = \left(\frac{T_i}{R_i} - \frac{T_i'}{R_i'} \right) R_i'. \tag{3}$$

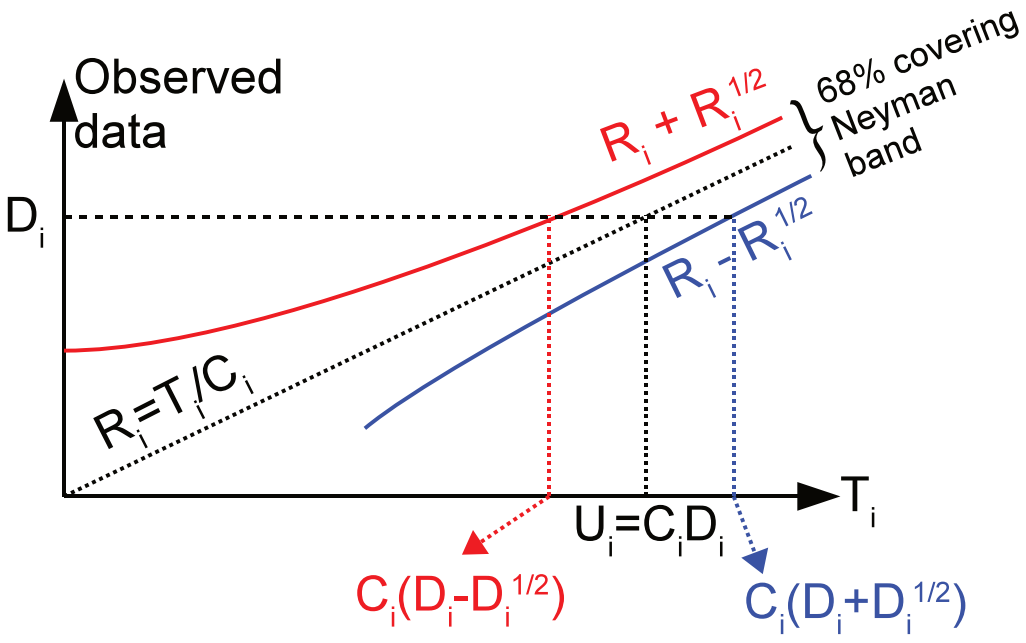


Fig. 3: Sketch of the Neyman construction used to correspond an observed number of data events D_i to a 68% confidence interval for T_i in the bin-by-bin correction factor method. The definitions of D_i, T_i, R_i, C_i and U_i are given in Sec. 2.1. For large values of T, the upper and lower bounds of the Neyman band follow asymptotically $R_i + \sqrt{R_i}$ and $R_i - \sqrt{R_i}$ respectively. At low values of T, where Poisson is not well-approximated by a Gaussian, the Neyman band is not symmetric around R_i, which is the reason that in this sketch the bounds of the Neyman band are obscured at low T.

2.2 Statistical uncertainty

The Neyman construction shown in Fig. 3 is effectively used to obtain a confidence interval for T_i, given C_i and the data D_i. Having observed D_i, the 68% confidence interval (CI) for U_i is approximately

$$C_i(D_i \pm \sqrt{D_i}). \tag{4}$$

This is a fair approximation when D_i is large, in which case the Poisson distribution of D_i with mean R_i is similar to a Gaussian of mean R_i and standard deviation $\sqrt{R_i}$. Although this approximation fails in bins with few data, the same formula was used in all p_T bins, so for all bins it was assumed that the statistical uncertainty of U_i is symmetric and equal to

$$\sigma_{U_i} = C_i\sqrt{D_i}. \tag{5}$$

This is the size of the black error bars in Fig. 1.

2.3 Systematic uncertainty

The following main sources of systematic uncertainty were identified in [2]:

i) the correction factor C_i is subject to statistical fluctuations due to finite MC event statistics;
ii) the amount of p_T smearing in detector simulation may be unrealistic;
iii) the used spectrum of T_i may be unrealistic;
iv) jet energy scale uncertainty.

The following paragraphs describe how each systematic uncertainty was propagated to the final estimator of the truth-level spectrum.

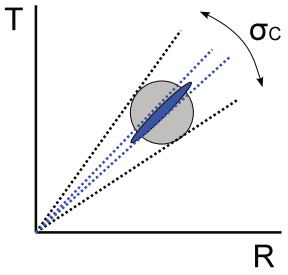


Fig. 4: The effect of covariance between T_i and R_i in the variance of the correction factor C_i. The gray circle indicates the case of zero covariance, and the dark ellipse the case of highly positive correlation.

2.3.1 Finite MC event statistics

The correction factors C_i have some uncertainty due to random fluctuations of the finite MC event statistics available to determine T_i and R_i. When T_i fluctuates above its mean, so does R_i, so the two are highly correlated (Fig. 4). The statistical uncertainty in C_i was computed taking this correlation into account, as follows:

The MC events which compose R_i ($N(R_i)$) are separated into those coming from the same truth-level bin ($N(R_i \wedge T_i)$) and those coming from different truth-level bins ($N(R_i \wedge \neg T_i)$).[2] Similarly, the $N(T_i)$ MC events which contribute to T_i are separated into those that end up in the same bin after detector simulation and event selection ($N(T_i \wedge R_i)$), and those that migrate to different bins ($N(T_i \wedge \neg R_i)$). The variables $N(T_i \wedge R_i)$ and $N(R_i \wedge T_i)$ are identical. So, C_i can be expressed as a function of three statistically independent random variables:

$$C_i = \frac{T_i}{R_i} = \frac{N(T_i \wedge R_i) + N(T_i \wedge \neg R_i)}{N(T_i \wedge R_i) + N(R_i \wedge \neg T_i)}. \tag{6}$$

Since C_i is expressed as a function of three statistically uncorrelated variables, error propagation can be used where covariance terms are zero. Each one of the three MC event populations has standard deviation $\sqrt{N}$.

2.3.2 Jet p_T resolution uncertainty

A relative systematic uncertainty of 15% in jet p_T resolution was assumed, based on the results of in-situ studies [10].

To model the effect of a different p_T resolution on C_i, the jets in MC events were smeared by an additional amount α, which varied from 0 to 20% of the nominal smearing that is present in ATLAS MC. For each amount of extra smearing, the values of R_i change, while T_i is not affected. As a result each correction factor C_i has a dependence on the amount of extra smearing. It was furthermore observed that in all bins i the correction factor C_i varied linearly with α.

It is possible to increase the smearing of jet p_T by adding to it a random offset of appropriate variance, but it is not possible to do the opposite, i.e., to reduce the amount of smearing that is nominally present in the ATLAS MC. This complicates the task of determining the uncertainty on C_i, because the resolution uncertainty of 15% is symmetric; the jet p_T resolution could be 15% worse or 15% better than its nominal value. The observation that C_i depends linearly on the extra smearing justifies the assumption that, if the resolution improved, C_i would still vary linearly.

[2] The symbol $\wedge$ is the logical "and", while $\neg$ is the logical "not". So, $R_i \wedge \neg T_i$ means belonging in R_i *and not* in T_i.

2.3.3 *Uncertainty in spectrum shape*

The correction factors C_i depend on the choice of T_i, which affects also R_i. If, for example, PYTHIA QCD does not provide a realistic model of the true spectrum, that can bias U_i, unless R_i and T_i are simultaneously wrong in such a way that T_i/R_i remains equal to the (unknown) actual ratio T'_i/R'_i in Eq. 3. The use of bins quite wider than the amount of smearing makes it more likely that, even if T_i is not modeled correctly, the ratio T_i/R_i in each bin will be approximately correct. In [2] the bins are safely wider than jet p_T resolution, and their edges are driven by experimental constraints, such as trigger thresholds.

To assess the uncertainty from possible wrong modeling of T_i, the MC events used to determine C_i were re-weighted in multiple ways. Their re-weighting was determined by functions smooth in jet p_T, chosen so as to bracket the variation observed by varying parton density functions, by including next-to-leading-order corrections to QCD, as well as the difference observed between D_i and R_i. For each set of re-weighted MC events both T_i and R_i were re-computed, and so was C_i for each bin i. The largest variation observed in each C_i was taken as a systematic uncertainty.

2.3.4 *Jet energy scale uncertainty*

By far the dominant uncertainty in the final U_i comes from the uncertainty in jet energy scale (JES). All previous uncertainties, added in quadrature, amount to about 5% of relative uncertainty in C_i, which is the error band shown in Fig. 2. The rest $\sim$40% of uncertainty in the final answer comes from the JES uncertainty, and it dominates the blue error band in Fig. 1.

To propagate the JES uncertainty, the reconstructed p_T of all jets in MC events is shifted by ± 1 standard deviation, the exact size of which is a function of jet p_T and pseudo-rapidity η. That affects R_i strongly, while T_i doesn't change, therefore C_i varies significantly. By applying on D_i the two alternative values of C_i, from the positive and the negative JES shift, two extreme U_i values are obtained for each bin i, whose distance is considered as the JES uncertainty on U_i.

3 Iterative unfolding

ATLAS used D' Agostini's iterative unfolding [1] in the study of minimum bias pp collisions [3]. The example to be shown is the estimation of the truth-level distribution of the multiplicity of charged particles. The result of this analysis is shown in Fig. 5.

3.1 Method description

The full method is clearly described in the original article [1] by D' Agostini. This paragraph will make a connection between the quantities in [3] and the notation used in [1].

Let n_{ch} be the number of charged particles produced in a pp collision. This is the truth-level quantity whose distribution needs to be estimated. It corresponds to the "cause" C mentioned in [1].

Let n_{trk} be the number of reconstructed tracks in a pp collision, which satisfy the selection criteria listed in [3]. It corresponds to the "effect" E mentioned in [1].

The reconstructed tracks are typically fewer than the actual charged particles, due to tracking inefficiency, therefore typically $n_{trk} \leq n_{ch}$. Therefore the migrations matrix is highly non-diagonal, and schematically looks like Fig. 6.

Re-writing the basic formulas from [1], substituting $C \to n_{ch}$ and $E \to n_{trk}$, we get

$$\hat{N}(n_{ch}) = \frac{1}{\epsilon(n_{ch})} \sum_{n_{trk} \geq 2} N(n_{trk}) P(n_{ch}|n_{trk}) \qquad \epsilon(n_{ch}) \neq 0, \tag{7}$$

$$P(n_{ch}|n_{trk}) = \frac{P(n_{trk}|n_{ch}) P_0(n_{ch})}{\sum_{n_{ch} \geq 1} P(n_{trk}|n_{ch}) P_0(n_{ch})}, \tag{8}$$

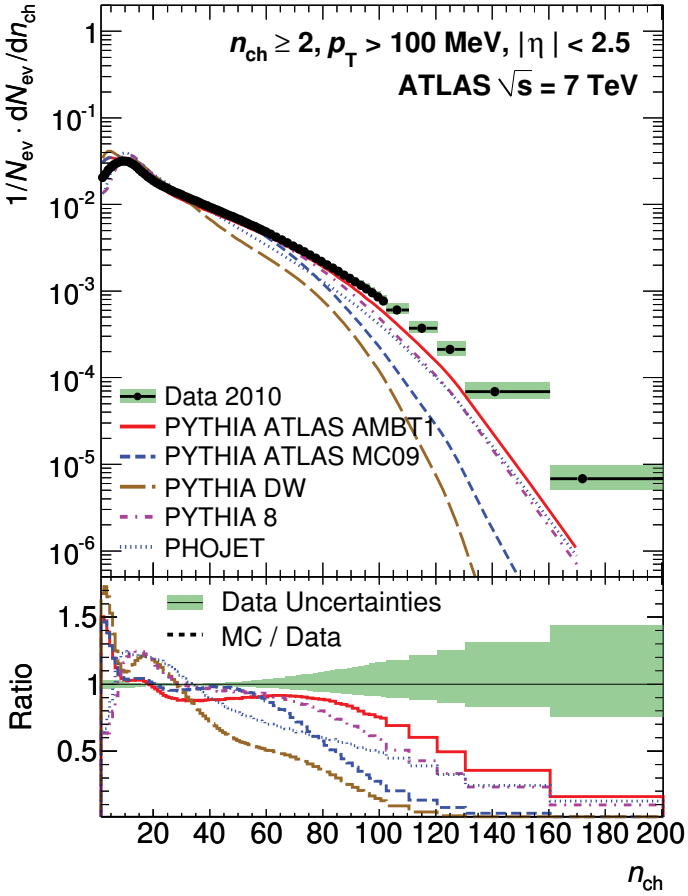


Fig. 5: The estimated spectrum of charged particle multiplicity (filled circles) in minimum bias pp interactions, from [3]. The statistical uncertainty is smaller than the marker size, and the asymmetric color band represents the total systematic uncertainty. Various theoretical predictions are overlaid for comparison.

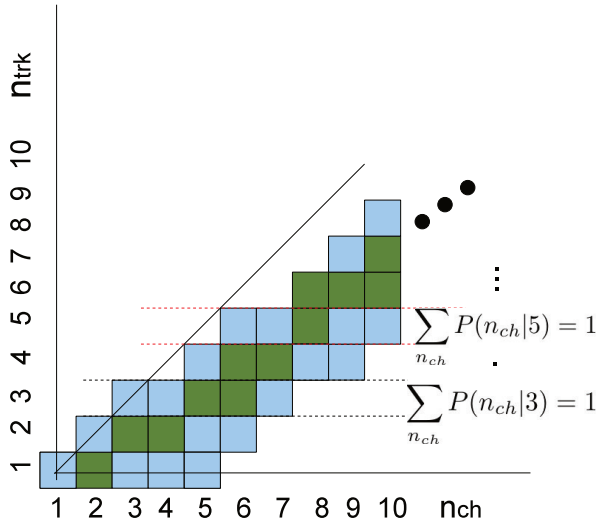


Fig. 6: Schematic representation of migrations matrix. The dark green squares represent higher probability than the light blue. Initially each matrix element equals the probability of MC events to contain n_{ch} charged particles and to have n_{trk} reconstructed tracks that satisfy the criteria listed in [3]. Then, the elements of each row, corresponding to a fixed n_{trk}, are normalized to have sum 1.

where the efficiency $\epsilon(n_{ch})$ corresponds to the probability of reconstructing at least two tracks, a requirement related to having a reliable primary vertex reconstruction, for a given number of charged particles:

$$\epsilon(n_{ch}) = P(n_{trk} \geq 2 | n_{ch}). \tag{9}$$

The term $P_0(n_{ch})$ is an arbitrary initial distribution for the truth-level quantity n_{ch}. The symbol $N(n_{trk})$ denotes the population of events where n_{trk} tracks were reconstructed, and $\hat{N}(n_{ch})$ is the estimator of the population of events with n_{ch} charged particles at truth-level.

3.1.1 Initial distribution and iterations

In [3], the initial distribution was defined to be the n_{ch} spectrum predicted by PYTHIA minimum bias MC. The reason is that the PYTHIA prediction has been tuned to data from various past experiments, so it is a reasonable starting point.

In iterative unfolding the number of iterations is decided arbitrarily. Too many iterations result in bin-by-bin fluctuations in the unfolded spectrum, similar to what one may get from simple migration matrix inversion [4]. Too few iterations increase too much the influence of the initial distribution on the final answer.

In [3], a convergence criterion was defined to determine when to stop iterating. The criterion was

$$\frac{\chi^2}{N_{bins}} < 1, \tag{10}$$

where

$$\chi^2 \equiv \sum_{i=1}^{N_{bins}} \left(\frac{n_{ch}^{i,current} - n_{ch}^{i,previous}}{\sqrt{n_{ch}^{i,previous}}} \right)^2 . \tag{11}$$

Namely, iterations continued until the latest unfolded spectrum ($n_{ch}^{current}$) remained statistically consistent with the spectrum from the previous iteration ($n_{ch}^{previous}$). It was found that 4 iterations were enough to meet this convergence criterion.

3.1.2 The term $\epsilon(n_{ch})$

In principle one should extract $\epsilon(n_{ch})$ defined in Eq. 9, directly from the MC events used to populate the migrations matrix (Fig. 6). However, a decision was made in [3] to use instead a parametric approximation of $\epsilon(n_{ch})$.

Making the simplification that each charged particle has the same "average effective" probability ϵ_{eff} of being reconstructed as a track, the probability of having at least two reconstructed tracks is given by

$$f(n_{ch}) = 1 - (1 - \epsilon_{eff})^{n_{ch}} - n_{ch}(1 - \epsilon_{eff})^{(n_{ch}-1)}\epsilon_{eff}. \tag{12}$$

The unknown parameter ϵ_{eff} was adjusted so as make $f(2)$ equal to the $\epsilon(n_{ch} = 2)$ obtained from MC. The resulting value for ϵ_{eff} is within 4% from the average probability of track reconstruction that is determined from MC simulation, which indicates that $f(n_{ch})$ matches well the MC-driven $\epsilon(n_{ch})$ even for $n_{ch} > 2$.

After adjusting ϵ_{eff} as described, the quantity $f(n_{ch})$ from Eq. 12 substitutes $\epsilon(n_{ch})$ in Eq. 7. Practically this efficiency becomes $\simeq 1$ for $n_{ch} > 4$, and that is true regardless of using $f(n_{ch})$ or $\epsilon(n_{ch})$.

3.2 Statistical uncertainty

In Eq. 7, the estimator $\hat{N}(n_{ch})$ depends on the measured $N(n_{trk})$, which are the result of independent Poisson fluctuations. Simple error propagation would lead to the following standard deviation for $\hat{N}(n_{ch})$:

$$\sigma_{\hat{N}(n_{ch})} = \sqrt{\sum \left(\frac{1}{\epsilon(n_{ch})} P(n_{ch}|n_{trk}) \right)^2 N(n_{trk})}. \tag{13}$$

The way statistical uncertainty was actually calculated in [3] was

$$\sigma_{\hat{N}(n_{ch})} = \sqrt{\hat{N}(n_{ch})}. \tag{14}$$

Either way, the statistics in all bins of n_{trk} are high enough to make the statistical uncertainty negligible. In Fig. 5, the statistical error bars are invisible.

3.3 Systematic uncertainty

The following main sources of systematic uncertainty will be discussed:

i) The choice of initial distribution $P_0(n_{ch})$;
ii) The uncertainty in track reconstruction efficiency;
iii) The uncertainty in MC spectrum.

3.3.1 Choice of initial distribution

The stability of the answer under different choices of initial distribution $P_0(n_{ch})$ was tested by assuming a "flat" initial distribution $P_0(n_{ch}) = 1$, and repeating the iterative unfolding procedure. This choice is obviously physically absurd; its purpose was only to show that even under extreme choices of $P_0(n_{ch})$ the answer $\hat{N}(n_{ch})$ doesn't change much.

Starting from a flat initial distribution, the number of iterations required to converge (Eq. 10) increased from 4 to 7. The final answer changed by less than 2% in all bins of n_{ch}, which was taken as a systematic uncertainty in $\hat{N}(n_{ch})$.

3.3.2 Track reconstruction efficiency uncertainty

The main effect this unfolding is correcting is the inefficiency of tracking. This inefficiency is reflected in the probabilities of Eq. 7, and is obtained from MC simulation. If tracking inefficiency in MC is wrong, so is the obtained spectrum after unfolding.

Fig. 7 shows the track reconstruction efficiency (ϵ_{trk}) in ATLAS simulation.

To propagate the uncertainty of ϵ_{trk} into $\hat{N}(n_{ch})$, the natural thing to do would be to shift systematically ϵ_{trk}, thus changing $P(n_{ch}|n_{trk})$, and see how much $\hat{N}(n_{ch})$ would change. Instead, what was done in [3] was to keep the migration probabilities fixed, and modify the data ($N(ch_{trk})$) on which iterative unfolding was applied. The way in which the data were modified is described next.

Assume an event in data has n_{trk} tracks. Take one of these tracks. Its p_T corresponds to some efficiency ϵ_{trk} (Fig. 7). For the sake of clarity, let's say it corresponds to $\epsilon_{trk} = 0.80 \pm 0.05$. This ϵ_{trk} gets reduced by 1 standard deviation, so it is brought down to 0.75. For this reduced ϵ_{trk}, the expected number of tracks is $\frac{1}{0.80} \times 0.75 \simeq 0.94$. The track is then randomly kept, with probability 0.94, or discarded, with probability 0.06. This procedure of efficiency reduction and random removal is repeated for all n_{trk} tracks of the event. In the end, the event is left with n'_{trk}, where $n'_{trk} \leq n_{trk}$.

The above procedure is repeated for all data events, reducing n_{trk} to n'_{trk} in each event. Then, the distribution $N(n'_{trk})$ is unfolded instead of $N(n_{trk})$, which results in $\hat{N}'(n_{ch})$ instead of $\hat{N}(n_{ch})$.

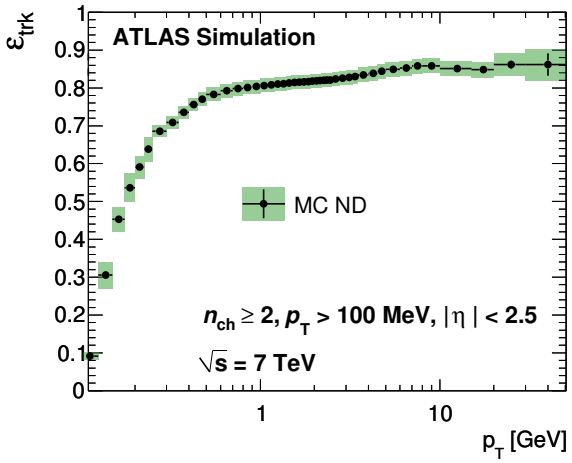


Fig. 7: Track reconstruction efficiency ϵ_{trk} in ATLAS simulation. The error band represents its systematic uncertainty.

The above procedure could only remove tracks, not create any. However, the ϵ_{trk} uncertainty is symmetric, which means that the actual ϵ_{trk} could be also greater than its nominal value. For this reason, the difference between $\hat{N}(n_{ch})$ and $\hat{N}'(n_{ch})$ is symmetrized, and used as a systematic uncertainty in $\hat{N}(n_{ch})$. That means, for example, that if in a bin of n_{ch} the $\hat{N}'(n_{ch})$ was 5% greater than $\hat{N}(n_{ch})$, the uncertainty is set to $\pm 5\%$.

3.3.3 Uncertainty due to spectrum shape

The observed spectrum of track transverse momentum (p_T^{trk}) disagrees with the MC prediction after full ATLAS detector simulation, as shown in Fig. 8. This discrepancy is related to the unfolding from n_{trk} to n_{ch}, because ϵ_{trk} is a function of p_T^{trk} (Fig. 7). If the p_T^{trk} is not realistically modeled, neither is ϵ_{trk}.

The way this was treated was the following: In each bin of n_{trk}, the mean ϵ_{trk} was found by looping through all data events in the bin, and corresponding each observed p_T^{trk} to the value of ϵ_{trk} obtained from MC (Fig. 7). The same was then done for MC events in bins of n_{trk}, again corresponding the p_T^{trk} in MC events to values of ϵ_{trk} from Fig. 7. In each bin of n_{trk}, the average track reconstruction efficiency $\langle \epsilon_{trk} \rangle$ from data was compared to the same quantity from MC. The same $p_T^{trk} \rightarrow \epsilon_{trk}$ correspondence was used for both data and MC tracks, therefore the difference in the resulting $\langle \epsilon_{trk} \rangle$ is due to the different p_T^{trk} distributions.

In each bin of n_{trk}, if $\langle \epsilon_{trk} \rangle$ is larger in data than in MC, then the efficiency in data gets reduced by the observed difference, in the same stochastic way described in Section 3.3.2. This results in a different number of tracks $n'_{trk} \leq n_{trk}$ for each event in the data. The iterative unfolding is then applied to $N(n'_{trk})$, and a different estimator of the truth-level spectrum is obtained ($\hat{N}'(n_{ch})$). The difference between the nominal $\hat{N}(n_{ch})$ and $\hat{N}'(n_{ch})$ is found, and is used as a one-sided systematic uncertainty in $\hat{N}(n_{ch})$.

In n_{trk} bins where the $\langle \epsilon_{trk} \rangle$ in data is smaller than in MC, one would ideally wish to increase the ϵ_{trk} of the data, but it is not possible to create tracks, as explained in Section 3.3.2. Instead, the ϵ_{trk} of data is *reduced* by the observed difference, as if the data had greater $\langle \epsilon_{trk} \rangle$ than the MC. This results in a different data spectrum $N(n'_{trk})$, which after unfolding results in a different estimator $\hat{N}'(n_{ch})$. The difference between the nominal $\hat{N}(n_{ch})$ and $\hat{N}'(n_{ch})$ is found, and instead of using it directly as a one-sided systematic uncertainty in $\hat{N}(n_{ch})$, we use its opposite, to take into account the fact that the data ϵ_{trk} was reduced instead of increased.

The above procedure results in an asymmetric systematic uncertainty. The upper and lower uncer-

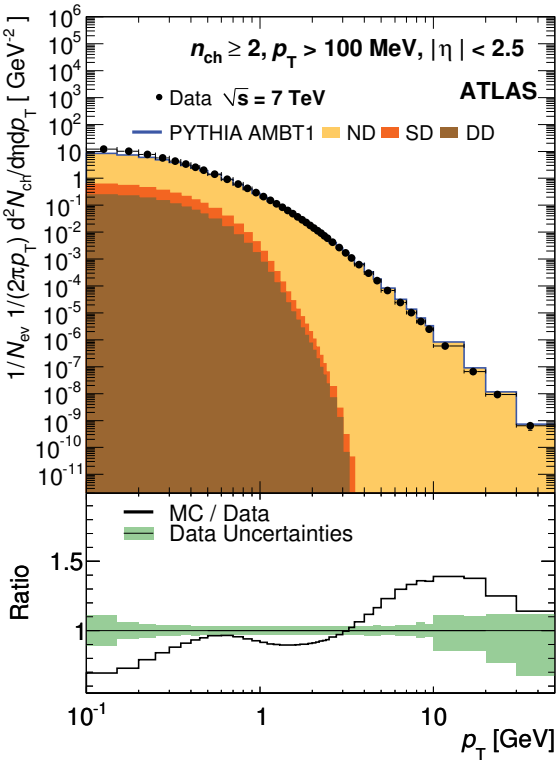


Fig. 8: The observed spectrum of p_T^{trk}, compared to the spectrum expected from MC simulation.

tainty are separately added in quadrature with the other systematic uncertainties, which are symmetric. This results in two unequal total systematic uncertainties, one suggesting that $N(n_{ch})$ could be above and the other below its nominal value. This asymmetric systematic uncertainty appears as a colored band in Fig. 5.

4 Concluding remarks

Two examples were shown of how unfolding has been used in ATLAS. They were chosen to be representative of different cases; one is using the bin-by-bin factors to correct the spectrum of a continuous observable (jet p_T), whereas the other uses iterative unfolding to estimate the truth-level distribution of a discrete variable (n_{ch}). The systematic uncertainties are quite different, as one analysis deals mainly with energy smearing, and the other with tracking inefficiency.

As of the time of this workshop, ATLAS has used extensively bin-by-bin correction factors, and in some cases iterative unfolding. More methods are being considered for future iterations of some of these analyses.

Unfolding has been used only in analyses where the goal was to estimate a truth-level distribution. Unfolding has been deliberately avoided in searches for new physics, where bias in bins with low statistics can not be afforded, where it can not be assumed that the data are consistent with the MC prediction as is silently assumed in some stages of unfolding, and where Poisson-distributed data are simpler to evaluate than estimators resulting from unfolding procedures after a series of arbitrary regularization choices. There are several unfolding methods, in some of which anomalies due to new physics could even be reduced, whereas the observed data are unique. Any inference is possible using directly the data, without unfolding, and a theoretical prediction that either includes full detector simulation, or at least an approximation of it that amounts to the inverse of unfolding, namely folding, which can be done with no need for regularization. The only task for which unfolding is strictly needed is the estimation of a truth-

level spectrum, whose later use to make statistical inferences becomes complicated by non-Poissonian statistics, biases that are hard to estimate, and bin-to-bin correlations.

In none of the analyses where unfolding was used was the full covariance matrix provided. The latter would be necessary to correctly compare a truth-level theoretical prediction to the result of unfolding. In most cases the comparison between the result of unfolding and the truth-level Standard Model prediction is made qualitatively, avoiding to provide a p-value that would require proper use of the covariances between bins. When a more quantitative comparison is attempted, like in [6], a χ^2 is used only as a metric to determine if one theory agrees with the data more than another, but not as a test statistic to compute a p-value.

Acknowledgements

I thank the organizers of PHYSTAT2011 for the opportunity to have this discussion, and my ATLAS collaborators for their feedback in preparation for this workshop.

References

[1] G. D' Agostini, *A Multidimensional Unfolding Method Based On Bayes' Theorem* Nucl. Instr. and Meth. in Phys. Res. A362 (1995) 487

[2] ATLAS Collaboration, *Measurement of inclusive jet and dijet cross sections in proton-proton collisions at 7 TeV centre-of-mass energy with the ATLAS detector*, Eur.Phys.J.C71:1512,2011, arXiv:1009.5908v2 [hep-ex]

[3] ATLAS Collaboration, *Charged-particle multiplicities in pp interactions measured with the ATLAS detector at the LHC*, arXiv:1012.5104, accepted by New Journal of Physics.

[4] Glen Cowan, *A Survey Of Unfolding Methods For Particle Physics*, available at the URL `http://www.ippp.dur.ac.uk/old/Workshops/02/statistics/proceedings/cowan.pdf`

[5] ATLAS Collaboration, *Measurement of the inclusive isolated prompt photon cross section in pp collisions at $\sqrt{s}$ = 7 TeV with the ATLAS detector*, arXiv:1012.4389v2 [hep-ex], submitted to Phys. Rev. D.

[6] ATLAS Collaboration, *Study of Jet Shapes in Inclusive Jet Production in pp Collisions at $\sqrt{s}$ = 7 TeV using the ATLAS Detector*, arXiv:1101.0070v1 [hep-ex], submitted to Phys. Rev. D.

[7] ATLAS Collaboration, *Measurement of the production cross section for W-bosons in association with jets in pp collisions at $\sqrt{s}$ = 7 TeV with the ATLAS detector*, arXiv:1012.5382v2 [hep-ex], submitted to Physics Letters B.

[8] T. Sjostrand, S. Mrenna and P. Z. Skands, JHEP **0605**, 026 (2006) [arXiv:hep-ph/0603175].

[9] ATLAS Collaboration, *Properties of Jets and Inputs to Jet Reconstruction and Calibration with the ATLAS Detector Using Proton-Proton Collisions at $\sqrt{s}$=7 TeV*, ATLAS-CONF-2010-053, available at the URL `http://cdsweb.cern.ch/record/1281310`

[10] ATLAS Collaboration, *Jet energy resolution and selection efficiency relative to track jets from in-situ techniques with the ATLAS Detector Using Proton-Proton Collisions at a Center of Mass Energy $\sqrt{s}$ = 7 TeV*, ATLAS-CONF-2010-054, available at the URL `http://cdsweb.cern.ch/record/1281311`

Comments on Unfolding Methods in ALICE

Jan Fiete Grosse-Oetringhaus for the ALICE collaboration
CERN, Geneva, Switzerland

Abstract
This paper discusses the unfolding methods presently used within ALICE and gives examples of practical issues with unfolding from an experimentalist's point of view.

1 Unfolding Methods

The main unfolding methods used within ALICE are χ^2 minimization with regularization [1] and iterative unfolding based on Bayes' theorem [2, 3]. Both use a detector response matrix which is determined with a Monte Carlo simulation.

In χ^2 minimization with regularization, the binned unfolded spectrum U is found by minimizing

$$\hat{\chi}^2(U) = \sum_m \left(\frac{M_m - \sum_t R_{mt} U_t}{e_m} \right)^2 + \beta F(U), \tag{1}$$

where R is the response matrix, M is the measured spectrum, e is the estimated measurement error, and $\beta F(U)$ is a regularization term that suppresses high-frequency components in the solution. The regularization term imposes assumptions on the shape of the corrected spectrum which are kept to a minimum by just requiring that the corrected spectrum is smooth. The smoothness is imposed by the choice

$$F(U) = \sum_t \frac{(U'_t)^2}{U_t} = \sum_t \frac{(U_{t-1} - U_t)^2}{U_t}, \tag{2}$$

which minimizes the fluctuations with respect to a constant constraint imposed by first derivatives or by

$$F(U) = \sum_t \frac{(U''_t)^2}{U_t} = \sum_t \frac{(U_{t-1} - 2U_t + U_{t+1})^2}{U_t}, \tag{3}$$

which minimizes the fluctuations with respect to a linear constraint imposed by second derivatives.[1] The regularization coefficient β is chosen such that, after minimization, the contribution of the first term in Eq. 1 is of the same order as the number of degrees of freedom (the number of bins in the unfolding).

The second unfolding method is an iterative procedure based on Bayes' theorem using the relations:

$$\tilde{R}_{tm} = \frac{R_{mt} \cdot P_t}{\sum_{t'} R_{mt'} P_{t'}}, \qquad U_t = \sum_m \tilde{R}_{tm} M_m, \tag{4}$$

with an *a priori* distribution P. The result U of an iteration is used as a new *a priori* P distribution for the following iteration. Optionally a smoothing is applied, averaging over adjacent bins:

$$\hat{U}_t = (1 - \alpha) \cdot U_t + \alpha \cdot \frac{1}{3}(U_{t-1} + U_t + U_{t+1}) \tag{5}$$

with $0 \leq \alpha \leq 1$ deciding the level of smoothing. The smoothing as well as limiting the number of iterations regularizes the distribution and reduces the influence of high-frequency oscillation in the solution [4].

[1] The contribution of Eq. (2) and Eq. (3) to Eq. (1) vanishes for a constant and linear solution, respectively.

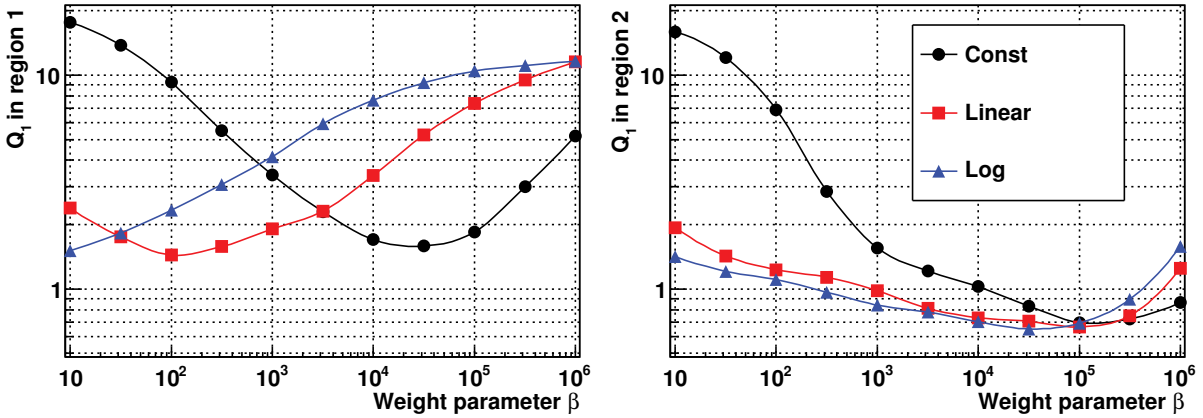


Fig. 1: Performance of the χ^2-minimization. The figure shows the difference between an input distribution and the unfolded distribution ($Q_1 = \langle \frac{|T_t - U_t|}{e_t} \rangle_t$) for a region where the slope changes quickly (left panel) and where the slope is rather constant (right panel) for three different regularizations (Const: Eq. (2), linear: Eq. (3), and log (equation not shown here, see [6]), as function of regularization weight β. The lines are drawn only to guide the eye. Figure from [6].

The bias of the regularization on the unfolded result is evaluated with the prescription given in [5]:

$$b_t = \sum_m \frac{\partial T_t}{\partial M_m} \left(\sum_t R_{mt} U_t - M_m \right) \tag{6}$$

where the derivative is evaluated numerically by unfolding several times while changing M.

1.1 Optimal parameters

While evaluating the regularization term $F(U)$ and the magnitude of β in χ^2 minimization as well as the number of iterations and the smoothing parameter in iterative unfolding, it is important to consider the expected shape of the distribution to be measured. If the distribution has different qualitative behavior in different regions, it is important to study them separately. For the example of the multiplicity distribution, there is a region where the slope changes quickly (at low multiplicity) and a region where the slope is rather constant (at intermediate multiplicities); a third qualitatively different region is at high multiplicities where the statistics is low. Figure 1 shows a MC study where different regularizations are evaluated as a function of β. It can be seen that the optimal parameter β is different for the different regions but only one parameter value can be used in the unfolding for the final distribution. Further, the optimal β for the two shown regions is closer in case of the regularization Eq. (2) than for Eq. (3). Such a conclusion, however, has to be studied using the expected shape of the unfolded distribution. In practice, a range of β values have to be used and the sensitivity to the specific choice has to be addressed in the systematic uncertainties. More details about this evaluation can be found in [6].

2 Applications

Unfolding methods are used among other analyses for the measurement of multiplicity distributions, p_T distributions and the jet spectra. For illustration, the response matrices for the unfolding of the multiplicity distribution and the jet p_T spectrum are shown in Figure 2.

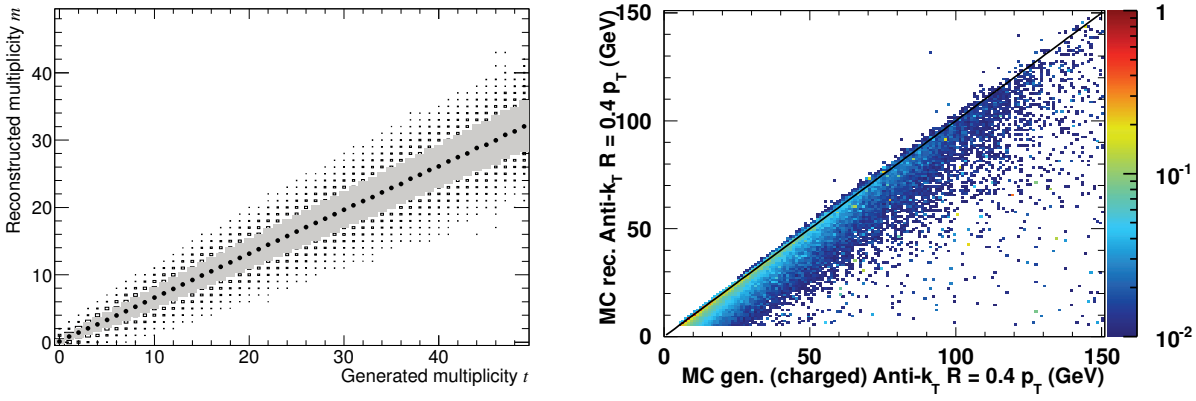


Fig. 2: Graphical representation of detector response matrices. Left panel: number of found tracks (m) vs the number of generated primary particles in $|\eta| < 1.0$ (t). The distribution of the measured tracks multiplicity for a given generated multiplicity shown with its most probable value (dots), r.m.s. (shaded areas), and full spread (squares). Figure from [7]. Right panel: reconstructed vs generated jet p_T found with the anti-k_T algorithm [8] (cone size $R = 0.4$).

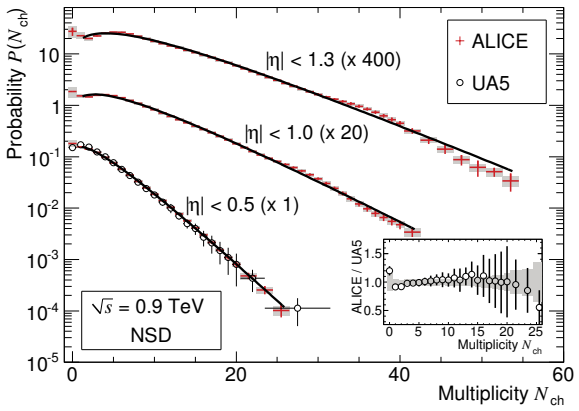


Fig. 3: Multiplicity distributions in three pseudorapidity (η) ranges for non single diffractive events at $\sqrt{s} = 900\,\mathrm{GeV}$. The solid lines show fits with negative binominal distribution. Figure from [7].

3 Residual structures in unfolded results

As discussed earlier regularization reduces the influence of high-frequency fluctuations. Obviously, such structures are not removed completely and depending on the analyzed sample they may still appear quite significant. Figure 3 shows the multiplicity distribution measured by ALICE at $\sqrt{s} = 900\,\mathrm{GeV}$ in different η-regions [7]. The distributions are fitted with negative-binomial distributions which emphasizes small wavy fluctuations in the two larger pseudorapidity intervals at multiplicities above 25. Visually these may appear to be significant, but one needs to note that the errors in the deconvoluted distribution are correlated over a range comparable to the multiplicity resolution (see the left panel of Figure 2).

It is interesting to study if one can attribute such a structure in the unfolded distribution to a structure in the measured distribution. This is done by assuming an exponential shape of the corrected distribution in the corresponding multiplicity range and studying the effect on the residuals. Figure 4 shows in the left panel the residuals of the distribution of Figure 3 in $|\eta| < 1$. The right panel shows the same residuals when an exponential shape of the unfolded distribution in the region around 30 is enforced. One sees that removing the structure from the unfolded result leads to the appearance of a

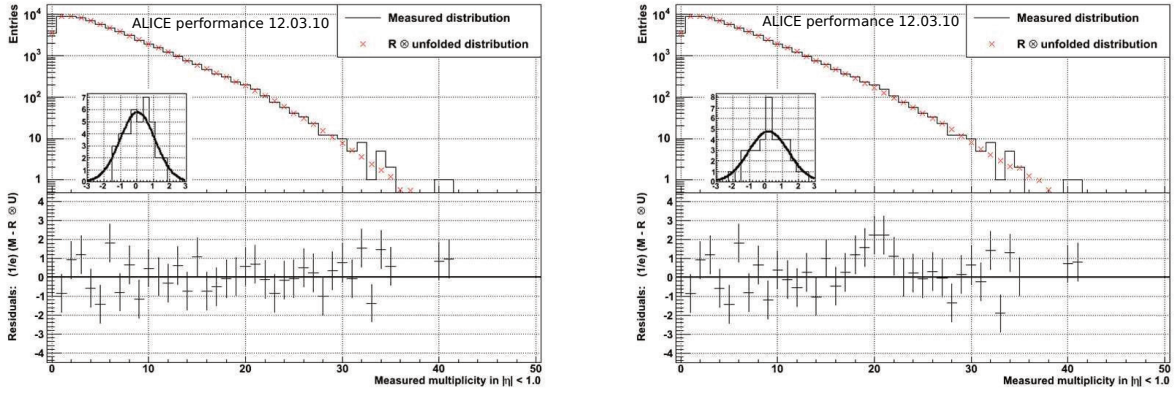


Fig. 4: Measured raw multiplicity distribution (elements of vector M, histogram), superimposed on the convolution $R \otimes U$ of the unfolded distribution with the response matrix (crosses), for $|\eta| < 1.0$ (upper plot). The error bars are omitted for visibility. Normalized residuals, i.e. the difference between the measured raw distribution and the corrected distribution folded with the response matrix divided by the measurement error (lower plot). The inset shows the distribution of these normalized residuals fitted with a Gaussian. The left panel shows the *normal* result while in the right panel an exponential shape is enforced. The small wavy fluctuation in the unfolded distribution (Fig. 3) reappears in the residuals when the exponential is enforced (right panel). For more details see text.

structure in the residuals. However, it is found only in a very few multiplicity bins, clearly less than in the unfolded distributions. This is expected because the response matrix is quite wide (see Figure 2, left panel) and therefore a measured bin can contribute to many unfolded bins. In summary one can learn from this example that a small fluctuation in the measured data can lead to a visually quite appealing structure in the unfolded result. Similar observations for a different deconvolution method were made by UA5 in [9].

References

[1] V. Blobel, in *8th CERN School of Comp. – CSC'84*, Aiguablava, Spain, 9–22 Sep. 1984, CERN-85-09, 88, (1985)

[2] G. D'Agostini, Nucl. Instrum. Meth. A **362**, 487 (1995)

[3] G. D'Agostini, CERN Report CERN-99-03 (1999)

[4] V. Blobel, arXiv:hep-ex/0208022 (2002).

[5] G. Cowan, in *Advanced Statistical Techniques in Particle Physics*, Durham, England, 18-22 Mar 2002, Durham Univ., 248 (2002)

[6] J.F. Grosse-Oetringhaus, PhD thesis, University of Münster, Germany, CERN-THESIS-2009-033 (2009)

[7] ALICE Collaboration, K. Aamodt et al., Eur. Phys. J. C **68** (2010) 89

[8] M. Cacciari and G. P. Salam, Phys. Lett. B **641** (2006) 57

[9] UA5 Collaboration, R.E. Ansorge et al., Z. Phys. C **43**, 357 (1989)

Unfolding algorithms and tests using RooUnfold

Tim Adye
Rutherford Appleton Laboratory, Science and Technology Facilities Council,
Harwell Science and Innovation Campus, Didcot OX11 0QX, United Kingdom.

Abstract
The RooUnfold package provides a common framework to evaluate and use different unfolding algorithms, side-by-side. It currently provides implementations or interfaces for the Iterative Bayes, SVD, and TUnfold methods, as well as bin-by-bin and matrix inversion reference methods. Common tools provide covariance matrix evaluation and multi-dimensional unfolding. A test suite allows comparisons of the performance of the algorithms under different truth and measurement models. Here I outline the package, the unfolding methods, and some experience of their use.

1 RooUnfold package aims and features

The RooUnfold package [1] was designed to provide a framework for different unfolding algorithms. This approach simplifies the comparison between algorithms and has allowed common utilities to be written.

Currently RooUnfold implements or interfaces to the Iterative Bayes [2, 3], Singular Value Decomposition (SVD) [4–6], TUnfold [7], bin-by-bin correction factors, and unregularized matrix inversion methods.

The package is designed around a simple object-oriented approach, implemented in C++, and using existing ROOT [8] classes. RooUnfold defines classes for the different unfolding algorithms, which inherit from a common base class, and a class for the response matrix. The response matrix object is independent of the unfolding, so can be filled in a separate 'training' program.

RooUnfold can be linked into a stand-alone program, run from a ROOT/CINT script, or executed interactively from the ROOT prompt. The response matrix can be initialized using existing histograms or matrices, or filled with built-in methods (these can take care of the normalization when inefficiencies are to be considered). The results can be returned as a histogram with errors, or a vector with full covariance matrix. The framework also takes care of handling multi-dimensional distributions (with ROOT support for 1–, 2–, and 3–dimensional histograms), different binning for measured and truth distributions, variable binning, and the option to include or exclude under- and over-flows.

It also supports different methods for calculating the errors that can be selected with a simple switch: bin-by-bin errors with no correlations, the full covariance matrix from the propagation of measurement errors in the unfolding, or the covariance matrix calculated using Monte Carlo (MC) toys.

All these details are handled by the framework, so don't have to be implemented for each algorithm. However different bin layouts may not produce good results for algorithms that rely on the global shape of the distribution (SVD).

A toy MC test framework is provided, allowing selection of different MC probability density functions (PDF) and parameters, comparing different binning, and performing the unfolding with the different algorithms and varying the unfolding regularization parameters. Tests can be performed with 1D, 2D, and 3D distributions. The results of a few such tests are presented in section 4.

2 C++ classes

Figure 1 summarizes how the ROOT and RooUnfold classes are used together. The RooUnfoldResponse object can be constructed using a 2D response histogram (TH2D) and 1D truth and measured projections

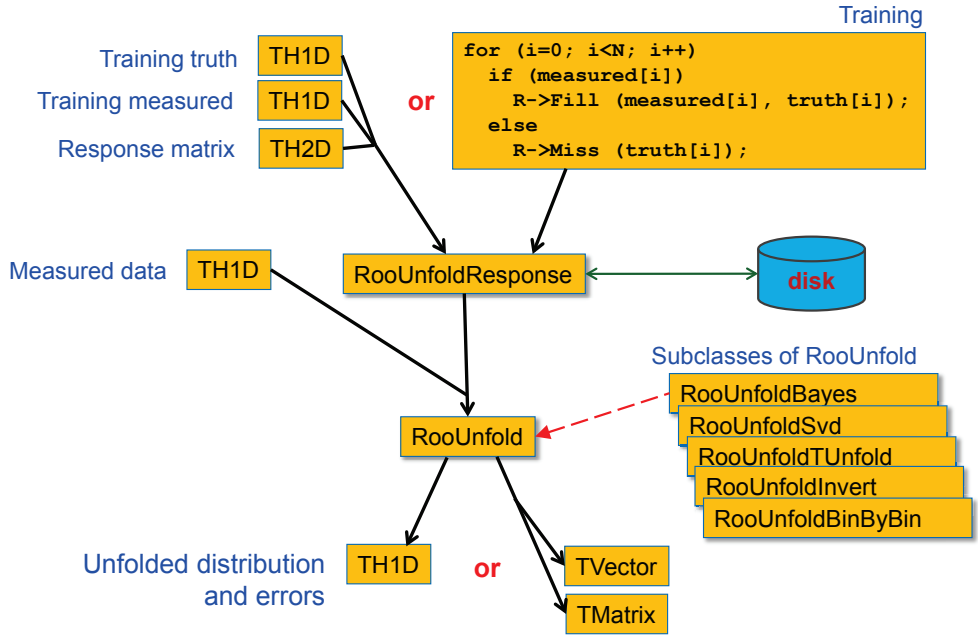


Fig. 1: The RooUnfold classes. The training truth, training measured, measured data, and unfolded distributions can also be given as TH2D or TH3D histograms.

(these are required to determine the effect of inefficiencies). Alternatively, RooUnfoldResponse can be filled directly with the `Fill`(x_{measured}, x_{true}) and `Miss`(x_{true}) methods, where the `Miss` method is used to count an event that was not measured and should be counted towards the inefficiency.

The RooUnfoldResponse object can be saved to disk using the usual ROOT input/output streamers. This allows the easy separation in separate programs of MC training from the unfolding step.

A RooUnfold object is constructed using a RooUnfoldResponse object and the measured data. It can be constructed as a RooUnfoldBayes, RooUnfoldSvd, RooUnfoldTUnfold, (etc) object, depending on the algorithm required.

The results of the unfolding can be obtained as ROOT histograms (TH1D, TH2D, or TH3D) or as a ROOT vector (TVectorD) and covariance matrix (TMatrixD). The histogram will include just the diagonal elements of the error matrix. This should be used with care, given the significant correlations that can occur if there is much bin-to-bin migration.

3 Unfolding algorithms

3.1 Iterative Bayes' theorem

The RooUnfoldBayes algorithm uses the method described by D'Agostini in [2]. Repeated application of Bayes' theorem is used to invert the response matrix. Regularization is achieved by stopping iterations before reaching the 'true' (but wildly fluctuating) inverse. The regularization parameter is just the number of iterations. In principle, this has to be tuned according to the sample statistics and binning. In practice, the results are fairly insensitive to the precise setting used.

RooUnfoldBayes takes the training truth as its initial prior, rather than a flat distribution, as described by D'Agostini. This should not bias the result once we have iterated, but could reach an optimum after fewer iterations.

This implementation takes account of errors on the data sample but not, by default, uncertainties in the response matrix due to finite MC statistics. That calculation can be very slow, and usually the training sample is much larger than the data sample.

RooUnfoldBayes does not normally do smoothing, since this has not been found to be necessary and can, in principle, bias the distribution. Smoothing can be enabled with an option.

3.2 Singular Value Decomposition

RooUnfoldSvd provides an interface to the TSVDUnfold class implemented in ROOT by Tackmann [6], which uses the method of Höcker and Kartvelishvili [4]. The response matrix is inverted using singular value decomposition, which allows for a linear implementation of the unfolding algorithm. The normalization to the number of events is retained in order to minimize uncertainties due to the size of the training sample. Regularization is performed using a smooth cut-off on small singular value contributions ($s_i^2 \rightarrow s_i^2/(s_i^2 + s_k^2)$, where the kth singular value defines the cut-off), which correspond to high-frequency fluctuations.

The regularization needs to be tuned according to the distribution, binning, and sample statistics in order minimize the bias due to the choice of the training sample (which dominates at small k) while retaining small statistical fluctuations in the unfolding result (which grow at large k).

The unfolded error matrix includes the contribution of uncertainties on the response matrix due to finite MC training statistics.

3.3 TUnfold

RooUnfoldTUnfold provides an interface to the TUnfold method implemented in ROOT by Schmitt [7]. TUnfold performs a matrix inversion with 0-, 1-, or 2-order polynomial regularization of neighbouring bins. RooUnfold automatically takes care of packing 2D and 3D distributions and creating the appropriate regularization matrix required by TUnfold.

TUnfold can automatically determine an optimal regularization parameter (τ) by scanning the 'L-curve' of $\log_{10}\chi^2$ vs $\log_{10}\tau$.

3.4 Unregularized algorithms

Two simple algorithms, RooUnfoldBinByBin, which applies MC correction factors with no inter-bin migration, and RooUnfoldInvert, which performs unregularized matrix inversion with singular value removal (TDecompSVD) are included for reference.

4 Examples

Examples of toy MC tests generated by RooUnfoldTest are shown in Figs. 2–4. These provide a challenging test of the procedure. Completely different training and test MC models are used: a single wide Gaussian PDF for training and a double Breit-Wigner for testing. In both cases these are smeared, shifted, and a variable inefficiency applied to produce the 'measured' distributions.

5 Unfolding errors

Regularization introduces inevitable correlations between bins in the unfolded distribution. To calculate a correct χ^2, one has to invert the covariance matrix:

$$\chi^2 = (\mathbf{x}_{\mathrm{measured}} - \mathbf{x}_{\mathrm{true}})^{\mathrm{T}}\mathbf{V}^{-1}(\mathbf{x}_{\mathrm{measured}} - \mathbf{x}_{\mathrm{true}}) \tag{1}$$

However, in many cases, the covariance matrix is poorly conditioned, which makes calculating the inverse problematic. Inverting a poorly conditioned matrix involves subtracting large, but very similar numbers, leading to significant effects due to the machine precision.

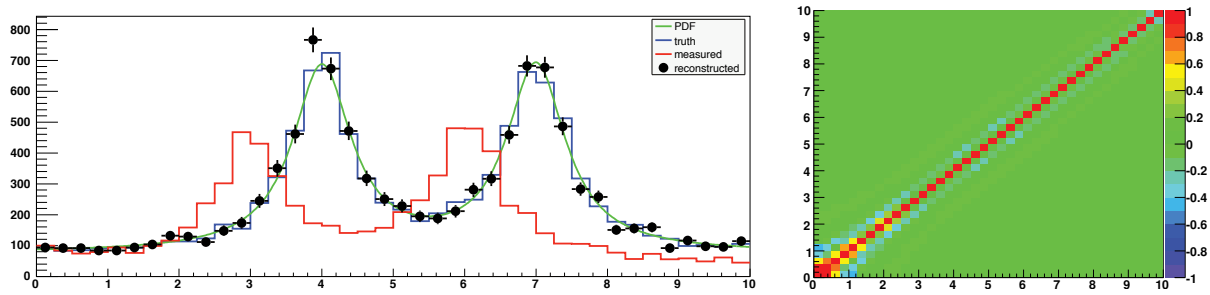


Fig. 2: Unfolding with the Bayes algorithm. On the left, a double Breit-Wigner PDF on a flat background (green curve) is used to generate a test 'truth' sample (upper histogram in blue). This is then smeared, shifted, and a variable inefficiency applied to produce the 'measured' distribution (lower histogram in red). Applying the Bayes algorithm with 4 iterations on this latter gave the unfolded result (black points), shown with errors from the diagonal elements of the error matrix. The bin-to-bin correlations from the error matrix are shown on the right.

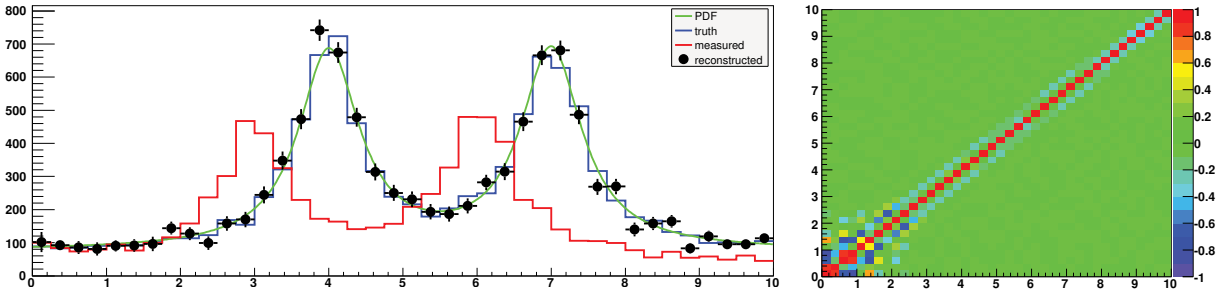


Fig. 3: Unfolding with the SVD algorithm ($k = 30$) on the same training and test samples as described in Fig. 2.

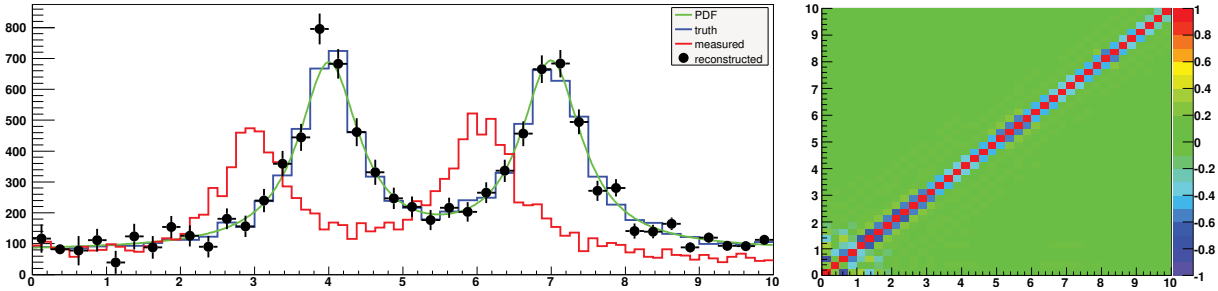


Fig. 4: Unfolding with the TUnfold algorithm ($\tau = 0.004$) on the same training and test samples as described in Fig. 2. Here we use two measurement bins for each truth bin.

5.1 Unfolding errors with the Bayes method

As shown on the left-hand side of Fig. 5, the uncertainties calculated by propagation of errors in the Bayes method were found to be significantly underestimated compared to those given by the toy MC. This was found to be due to an omission in the original method outlined by D'Agostini ([2] section 4).

The Bayes method gives the unfolded distribution ('estimated causes'), $\hat{n}(\mathrm{C}_i)$, as the result of applying the unfolding matrix, M_{ij}, to the measurements ('effects'), $n(\mathrm{E}_j)$:

$$\hat{n}(\mathrm{C}_i) = \sum_{j=1}^{n_\mathrm{E}} M_{ij} n(\mathrm{E}_j) \quad \text{where} \quad M_{ij} = \frac{P(\mathrm{E}_j|\mathrm{C}_i) n_0(\mathrm{C}_i)}{\epsilon_i f_j} \tag{2}$$

$P(\mathrm{E}_j|\mathrm{C}_i)$ is the response matrix, $\epsilon_i \equiv \sum_{j=1}^{n_\mathrm{E}} P(\mathrm{E}_j|\mathrm{C}_i)$ are efficiencies, and $f_j \equiv \sum_{l=1}^{n_\mathrm{C}} P(\mathrm{E}_j|\mathrm{C}_l) n_0(C_l)$ is the folded prior distribution, $n_0(C_l)$ — initially arbitrary (eg. flat or MC model), but updated on subsequent iterations.

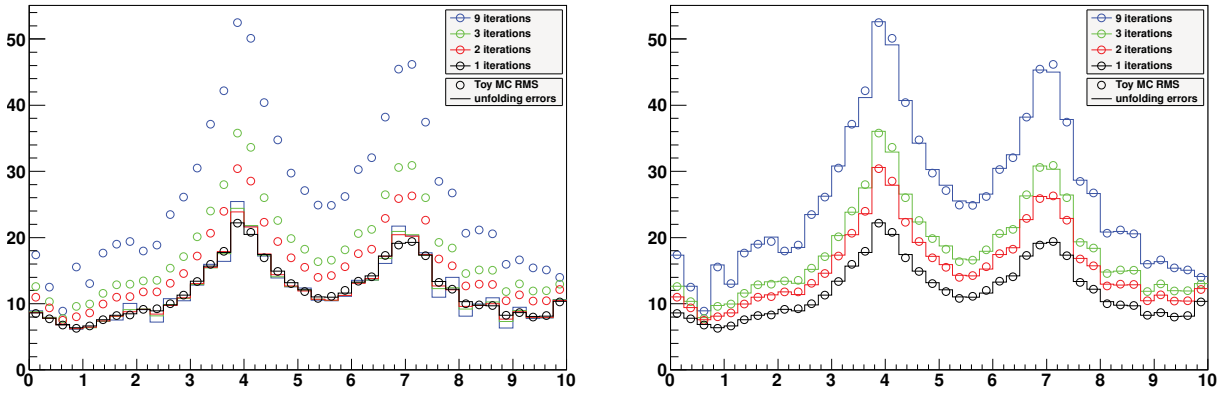


Fig. 5: Bayesian unfolding errors (lines) compared to toy MC RMS (points) for 1, 2, 3, and 9 iterations. The left-hand plot shows the errors using D'Agostini's original method, ignoring any dependence on previous iterations (only the M_{ij} term in Eq. (3)). The right-hand plot shows the full error propagation.

The covariance matrix, which here we call $V(\hat{n}(\mathrm{C}_k), \hat{n}(\mathrm{C}_l))$, is calculated by error propagation from $n(\mathrm{E}_j)$, but M_{ij} is assumed to be itself independent of $n(\mathrm{E}_j)$. That is only true for the first iteration. For subsequent iterations, $n_0(\mathrm{C}_i)$ is replaced by $\hat{n}(\mathrm{C}_i)$ from the previous iteration, and $\hat{n}(\mathrm{C}_i)$ depends on $n(\mathrm{E}_j)$ (Eq. (2)).

To take this into account, we compute the error propagation matrix

$$\frac{\partial \hat{n}(\mathrm{C}_i)}{\partial n(\mathrm{E}_j)} = M_{ij} + \sum_{k=1}^{n_\mathrm{E}} M_{ik} n(\mathrm{E}_k) \left(\frac{1}{n_0(\mathrm{C}_i)} \frac{\partial n_0(\mathrm{C}_i)}{\partial n(\mathrm{E}_j)} - \sum_{l=1}^{n_\mathrm{C}} \frac{\epsilon_l}{n_0(\mathrm{C}_l)} \frac{\partial n_0(\mathrm{C}_l)}{\partial n(\mathrm{E}_j)} M_{lk} \right) \tag{3}$$

This depends upon the matrix $\frac{\partial n_0(\mathrm{C}_i)}{\partial n(\mathrm{E}_j)}$, which is $\frac{\partial \hat{n}(\mathrm{C}_i)}{\partial n(\mathrm{E}_j)}$ from the previous iteration. In the first iteration, the second term vanishes ($\frac{\partial n_0(\mathrm{C}_i)}{\partial n(\mathrm{E}_j)} = 0$) and we get $\frac{\partial \hat{n}(\mathrm{C}_i)}{\partial n(\mathrm{E}_j)} = M_{ij}$.

The error propagation matrix can be used to obtain the covariance matrix on the unfolded distribution

$$V(\hat{n}(\mathrm{C}_k), \hat{n}(\mathrm{C}_l)) = \sum_{i,j=1}^{n_\mathrm{E}} \frac{\partial \hat{n}(\mathrm{C}_k)}{\partial n(\mathrm{E}_i)} V(n(\mathrm{E}_i), n(\mathrm{E}_j)) \frac{\partial \hat{n}(\mathrm{C}_l)}{\partial n(\mathrm{E}_j)} \tag{4}$$

from the covariance matrix of the measurements, $V(n(\mathrm{E}_i), n(\mathrm{E}_j))$.

Without the new second term in Eq. (3), the error is underestimated if more than one iteration is used, but agrees well with toy MC tests if the full error propagation is used, as shown in Fig. 5.

6 Status and plans

RooUnfold was first developed in the *BABAR* software environment and released stand-alone in 2007. Since then, it has been used by physicists from many different particle physics, particle-astrophysics, and nuclear physics groups. Questions, suggestions, and bug reports from users have prompted new versions with fixes and improvements.

Last year I started working with a small group hosted by the Helmholtz Alliance, the Unfolding Framework Project [9]. The project is developing unfolding experience, software, algorithms, and performance tests. It has adopted RooUnfold as a framework for development.

Development and improvement of RooUnfold is continuing. In particular, determination of the systematic errors due to uncertainties on the response matrix, and due to correlated measurement bins will be added. The RooUnfold package will be incorporated into the ROOT distribution, alongside the existing TUnfold and TSVDUnfold classes.

References

[1] The RooUnfold package and documentation are available from
http://hepunx.rl.ac.uk/~adye/software/unfold/RooUnfold.html

[2] G. D'Agostini, "A Multidimensional unfolding method based on Bayes' theorem," Nucl. Instrum. Meth. A **362** (1995) 487.

[3] K. Bierwagen, "Bayesian Unfolding," these proceedings.

[4] A. Hocker and V. Kartvelishvili, "SVD Approach to Data Unfolding," Nucl. Instrum. Meth. A **372** (1996) 469.

[5] V. Kartvelishvili, "Unfolding with SVD," these proceedings.

[6] K. Tackmann, "SVD-based unfolding: implementation and experience," these proceedings.

[7] The TUnfold package is available in ROOT [8] and documented in
http://www.desy.de/~sschmitt/tunfold.html

[8] R. Brun and F. Rademakers, "ROOT: An object oriented data analysis framework," Nucl. Instrum. Meth. A **389** (1997) 81. See also http://root.cern.ch/.

[9] For details of the Unfolding Framework Project, see
https://www.wiki.terascale.de/index.php/Unfolding_Framework_Project

Appendix

Program Committee

J. Berger (Duke)
V. Blobel (DESY)
B. Cousins (UCLA)
D. Cox (Oxford)
G. Cowan (Royal Holloway)
K. Cranmer (NYU)
L. Demortier (Rockefeller)
A. de Roeck (CERN)
B. Efron (Stanford)
G. Flucke (DESY)
E. Gross (Weizmann)
D. Hand (Imperial College)
J. Linnemann (MSU)
R. Lockhart (Simon Fraser)
L. Lyons (Imperial College)
M.L. Mangano (CERN)
S. Schmitt (DESY)
M. Williams (Imperial College)

List of Participants

ADYE, Tim	RAL, Didcot, UK
ANDARI, Nansi	Lab. de l'Accelerateur Lineaire, Paris, FRANCE
APERIO BELLA, L.	Laboratoire d'Annecy-le-Vieux, FRANCE
APOLLE, Rudi	RAL/Oxford, UK
ASSAMAGAN, Ketevi A.	BNL, USA
BACKES, Moritz	Université de Genève, SWITZERLAND
BARILE, Francesco	INFN, Bari, ITALY
BARLOW, Roger	University of Manchester, UK
BAYARRI, M.J.	Universitat de València, SPAIN
BEAUCHEMIN, Pierre-Hugues	University of Oxford, UK
BEAUJEAN, Frederik	Max Planck Institute, Munich, GERMANY
BEHNKE, Olaf	DESY, GERMANY
BELL, William H.	Université de Genève, SWITZERLAND
BERGER, James	Duke University, USA
BERNARDO, Jose M.	Universitat de València, SPAIN
BEVAN, Adrian	Queen Mary, University of London, UK
BIANCHIN, Chiara	Università degli Studi di Padova & INFN, ITALY
BIERWAGEN, Katharina	Georg-August-Universitaet, Göttingen, GERMANY
BITYUKOV, Sergey	Institute for high energy physics, Protvino, RUSSIA
BLOBEL, Volker	UHH - Universitaet Hamburg, GERMANY
BORRONI, Sara	INFN, Sezione di Roma I, ITALY & CERN
BRUN, Rene	CERN, SWITZERLAND
BUCKINGHAM, Ryan M.	University of Oxford, UK
CAFFARRI, Davide	Universita degli Studi di Padova INFN, ITALY
CALDWELL, Allen	Max Planck Institute, Munich, GERMANY
CALVET, David	LPC, Clermont Ferrand, FRANCE
CAMACHO TORO, Reina C.	LPC, Clermont Ferrand Université Blaise Pascal, FRANCE
CANO, Javier	URJC, Madrid, SPAIN
CARDINALE, Roberta	University of Genova - INFN, ITALY
CASADEI, Diego	New York University, USA
CERNY, Karel	Charles University, Prague, CZECH REPUBLIC
CHOUDALAKIS, Georgios	University of Chicago, Enrico Fermi Institute Chicago, USA
CONWAY, John	University of California, Davis, USA
COWAN, Glen	Royal Holloway, University of London, UK
COX, David	Nuffield College, Oxford, UK
CRANMER, Kyle	New York University, USA
CZYCZULA, Zofia	University of Oslo, NORWAY
DAVIES, Gavin	Imperial College, London, UK
DE ROECK, Albert	CERN, SWITZERLAND
DELMASTRO, Marco	CERN, SWITZERLAND

DEMBINSKI, Hans	KIT, Karlsruhe, GERMANY
DEMORTIER, Luc	Rockefeller University, USA
DENG, Jianrong	University California, Irvine, USA
DRASAL, Zbynek	IPNP, Prague, CZECH REPUBLIC
FASSI, Farida	CC-IN2P3/CNRS, Villeurbanne/Lyon, FRANCE
FAYARD, Louis	LAL, Orsay, FRANCE
FEBBRARO, Renato	LPC, Clermont-Ferrand, FRANCE
FELCINI, Marta	IFCA & UCLA, USA
FERRETTO PARODI, Andrea	INFN, Genova, ITALY
FONSECA DE SOUZA, Sandro	Universidade do Estado do Rio De Janeiro, BRAZIL
FORTE, Stefano	Milan University, ITALY
FRUHWIRTH, Rudolf	Institute of High Energy Physics, Vienna, AUSTRIA
FUJIWARA, Makoto	TRIUMF / Univ. of Calgary, CANADA
GANDINI, Paolo	University of Oxford, UK
GIACOBINO, Caroline	Université de Genève, SWITZERLAND
GONZALEZ SEVILLA, Sergio	DPNC, University of Geneva, SWITZERLAND
GRACHOV, Oleg	University of Kansas, USA
GRAZIANI, Giacomo	INFN, Sezione di Firenze-Università, ITALY
GROSSE-OETRINGHAUS, Jan	CERN, SWITZERLAND
GROSS, Eilam	Weizmann Institute of Science, Rehovot, ISRAEL
GUENTHER, Jaroslav	Institute of Physics - Acad. of Sciences, CZECH REPUBLIC
HANSEN, Jorgen	Niels Bohr Institute, Copenhagen, DENMARK
HAREL, Amnon	University of Rochester, USA
HAYS, Jonathan	Imperial College, London, UK
HERNANDO MORATA, Jose	Universidade de Santiago de Compostela, SPAIN
HOECKER, Andreas	CERN, SWITZERLAND
HRISTOV, Peter	CERN, SWITZERLAND
IANNI, Aldo	LNGS & INFN, Assergi, ITALY
ISSEVER, Cigdem	Oxford University, UK
JACHOLKOWSKI, Adam	Universita di Catania, SPAIN
JAMES, Fred	CERN, SWITZERLAND
JOHNSON, Valen	University of Texas, M.D. Anderson Cancer Center, USA
KALOGEROPOULOS, Alexis	Vrije Universiteit, Brussel, BELGIUM
KANAKI, Kalliopi	University of Bergen, NORWAY
KARBACH, Till Moritz	Dortmund, GERMANY
KARTVELISHVILI, Vato	Lancaster University, UK
KHOO, Teng Jian	University of Cambridge, UK
KIRN, Malina	University of Maryland, USA
KLEIN, Benjamin	Universiteit Gent, BELGIUM
KOVALSKYI, Dmytro	University of California, Santa Barbara, USA
KRASNIKOV, Nikolai	Institute for Nuclear Research, Moscow, RUSSIA
KUEMPEL, Daniel	Bergische Universitaet Wuppertal, GERMANY
KVASNICKA, Peter	Charles University, Prague, CZECH REPUBLIC

KVITA, Jiri	CERN, SWITZERLAND
LAHAV, Ofer	University College London, UK
LANDSBERG, Greg	Brown University, USA
LIE, Ki	University of Illinois, Urbana-Champaign, USA
LINNEMANN, Jim	Michigan State University, USA
LISTER, Alison	University of Geneva, SWITZERLAND
LIVERMORE, Sarah	University of Oxford, UK
LYONS, Louis	Imperial College, UK
MAGNAN, Anne-Marie	Imperial College, UK
MALAESCU, Bogdan	CERN, SWITZERLAND
MAL, Prolay K.	University of Arizona, USA
MANGANO, Michelangelo	CERN, SWITZERLAND
MARCISOVSKY, Michal	IoP ASCR Prague, CZECH REPUBLIC
MARZIN, Antoine	University of Oklahoma, USA
MASSIRONI, Andrea	INFN and University of Milano, ITALY
MILLAN MEJIAS, Barbara	Zurich University, SWITZERLAND
MINOT, Ariana Sage	LAL, Orsay, FRANCE
MITSOU, Vasiliki	University of València, SPAIN
MONETA, Lorenzo	CERN, SWITZERLAND
MURRAY, Bill	Rutherford Appleton Laboratory, UK
NIKOLOPOULOS, Konstantinos	Brookhaven National Laboratory, USA
NYGAARD, Casper	Niels Bohr Institute, Copenhagen, DENMARK
PANARETOS, Victor	EPFL, Lausanne, SWITZERLAND
PASHAPOUR, Shabnaz	Georg-August-Universitaet, Göttingen, GERMANY
PATRIZII, Laura	INFN, Bologna, ITALY
PEDRAZA MORALES, Maria I.	University of Wisconsin, USA
PETERSEN, Troels	Niels Bohr Institute, Copenhagen, DENMARK
PIERINI, Maurizio	CERN, SWITZERLAND
POLCI, Francesco	LPSC, Grenoble, FRANCE
POLIFKA, Richard	Universita Karlova, Praha, CZECH REPUBLIC
PRICE, Darren	Indiana University, USA
PROSPER, Harrison B.	Florida State University, USA
PRZEDZINSKI, Tomasz	Jagellonian University, Krakow, POLAND
PUESCHEL, Elisa	University of Massachusetts, USA
RAMMENSEE, Michael	Albert-Ludwigs-Universitaet Freiburg, GERMANY
RANUCCI, Gioacchino	INFN, Milano, ITALY
RAUTENBERG, Julian	Bergische Universitaet, Wuppertal, GERMANY
READ, Alexander	University of Oslo, NORWAY
REECE, Ryan D.	University of Pennsylvania, USA
REZVANI, Reyhaneh	University of Toronto, CANADA
ROLKE, Wolfgang	University of Puerto Rico, Mayaguez, USA
ROSZKOWSKI, Leszek	University of Sheffield, UK & SINS, Warsaw, POLAND
ROTHENBERG, Allan	CERN, SWITZERLAND

ROEVER, Christian	MPI für Gravitationsphysik, Hannover, GERMANY
SALEMI, Francesco	AEI - Hannover, Hannover, GERMANY
SARDY, Sylvain	Université de Genève, SWITZERLAND
SCHAARSCHMIDT, Jana	LAL, Orsay, FRANCE
SCHMELLING, Michael	MPI for Nuclear Physics, Heidelberg, GERMANY
SCHMITT, Stefan	DESY, Hamburg, GERMANY
SCHOTT, Grégory	KIT, Karlsruhe, GERMANY
SIMAK, Vladislav	Acad. of Sciences, Prague, CZECH REPUBLIC
SMESTAD, Lillian	University of Oslo, NORWAY
STANCO, Luca	INFN, Padova, ITALY
STEINKAMP, Olaf	Universitaet Zuerich, SWITZERLAND
STOBER, Fred	KIT, Karlsruhe, GERMANY
TACKMANN, Kerstin	CERN, SWITZERLAND
TARRADE, Fabien	Brookhaven National Laboratory, USA
TROTTA, Roberto	Imperial College, UK
TUA, Alan	University of Sheffield, UK
TUPPUTI, Salvatore A.	CERN, SWITZERLAND
VAN DYK, David	University of California, Irvine, USA
VEVERKA, Jan	Caltech, USA
VITELLS, Ofer	Weizmann Institute of Science, ISRAEL
WAGNER, Boris	University of Bergen, NORWAY
WALSH, S.	Universiteit Gent, BELGIUM
WEBER, Matthias A.	ETH, Zuerich, SWITZERLAND
WILLIAMS, J. Michael	Imperial College, UK
WOUDSTRA, Martin	University of Massachusetts, USA
YACOOB, Sahal	University of the Witwatersrand, SOUTH AFRICA
YAMAMOTO, Kyoko	Iowa State University, USA
ZECH, Guenter	Universitaet Siegen, GERMANY
ZEMAN, Martin	Acad. of Sciences, Prague, CZECH REPUBLIC
ZEVI DELLA PORTA, Giovanni	Harvard University, USA
ZHUKOV, Valery	University of Karlsruhe, GERMANY