21 August 1997

PHYSICS LETTERS B

ELSEVIER

Physics Letters B 407 (1997) 73–78

Bayesian analysis of multi-source data

Pushpalatha C. Bhat [a], Harrison B. Prosper [b,1], Scott S. Snyder [c]

[a] *Fermi National Accelerator Laborarory, P.O. Box 500, Batavia, IL 60510, USA*

[b] *Department of Physics, Florida State University, Tallahassee, FL 32306, USA*

[c] *Brookhaven National Laboratory, Upton, NY 11973, USA*

Received 17 April 1997; revised manuscript received 4 June 1997

Editor: L. Montanet

Abstract

We present a simple method, based on Bayes' theorem, to fit binned data to one or more multi-source models. Assuming a Poisson probability for the count in each bin we can eliminate exactly the nuisance parameters from the likelihood function and arrive at a formula that can be broadly applied. We illustrate the method by showing how it can be used to estimate the top quark mass. © 1997 Published by Elsevier Science B.V.

PACS: 02.50.-r; 02.50.Cw; 07.05.kf; 14.65.Ha

Keywords: Bayesian analysis; Data analysis; Binned fitting; Likelihood; Multi-source data; Top quark mass

1. Introduction

A problem that arises frequently in experimental physics is to fit binned data to a model consisting of a sum of N sources, taking due account of known uncertainties. The prototypical example is a 2-source model consisting of a sum of signal plus background. Barlow and Beeston [1] have provided perhaps the best solution to that problem within the framework of frequentist statistics. In this paper, we suggest an alternative *Bayesian* method of analysis which, we believe, has conceptual and practical advantages.

First, we shall review briefly the conceptual basis of the method of Ref. [1], because it looks superficially similar to the one we propose and, therefore, one might be tempted to see no difference between the two. In frequentist statistics [2] a single data set is considered to be drawn from an ensemble of data sets. For the problem considered here each data set consists of a set of observed counts $\{D_i\}$ and N sets of source counts $\{A_{ji}\}$, where $i = 1, \ldots, M$ label the bins and $j = 1, \ldots, N$ the sources. We assume that the mean count in the ith bin, d_i, and the mean source counts, $\{a_{ji}\}$, are related by $d_i \equiv \sum_{j=1}^{N} p_j a_{ji}$. The quantity a_{ji} is the mean count for bin i of source j and p_j is the corresponding source strength, given as a fraction of the mean count $\sum_i a_{ji}$ of source j. Usually, the source counts $\{A_{ji}\}$ are the result of Monte Carlo calculations. If we assign a Poisson probability to the total count in each bin then we can write the likelihood function as

[1] E-mail: harry@hep.fsu.edu.

PII S0370-2693(97)00723-5

$$L(D|a,p) = \left(\prod_{i=1}^{M} \frac{\exp(-d_i) d_i^{D_i}}{D_i!} \right) \times \left(\prod_{i=1}^{M} \prod_{j=1}^{N} \frac{\exp(-a_{ji}) a_{ji}^{A_{ji}}}{A_{ji}!} \right). \quad (1.1)$$

The likelihood function is just the sampling distribution for the $M + N \times M$ counts. It contains $N \times M$ unknown parameters a_{ji}, plus N unknown parameters p_j; The parameters of interest are the source strengths p_j; the parameters a_{ji} are, in the present context, *nuisance* parameters which we must get rid of to make progress.

There is no general method to eliminate nuisance parameters from a likelihood function in the frequentist approach [2]. What is done, in practice, is to replace the nuisance parameters with their maximum likelihood estimates. Unfortunately, this does not guarantee their elimination from the sampling distribution of the estimates. Nor is there a guarantee that these estimates will always lie in the physical region.

The analysis method we suggest here provides a natural and consistent framework to overcome the aforementioned problems. After describing the method we show how it can be used to perform a straightforward analysis of top quark mass data.

2. The method

Let M be the number of bins into which the data are divided. For each multi-source model, labelled by the discrete parameter K, we shall assign a Poisson probability to the count per bin and take our likelihood function to be

$$L(D|a,p,K) = \prod_{i=1}^{M} \frac{\exp(-d_i) d_i^{D_i}}{D_i!}. \quad (2.1)$$

The likelihood function is the probability associated with the observed counts $\{D_i\}$. The second product in Eq. (1.1) is interpreted as an *informative* prior probability

$$Q(a,K) = \prod_{i=1}^{M} \prod_{j=1}^{N} \frac{\exp(-a_{ji}) a_{ji}^{A_{ji}}}{A_{ji}!} da_{ji}, \quad (2.2)$$

for the nuisance parameters a_{ji}.

The unknowns are the parameters p_j and a_{ji}. In order to make inferences about the former the nuisance parameters a_{ji} must be eliminated. According to probability theory [3] the general way to do this is to use Bayes' theorem

$$P(a,p,K|D) = \frac{L(D|a,p,K) Q(a,K) q(p,K)}{\sum_K \int_a \int_p L(D|a,p,K) Q(a,K) q(p,K)}, \quad (2.3)$$

and then marginalize (that is, integrate) the posterior probability $P(a,p,K|D)$ with respect to a, to obtain,

$$P(p,K|D) = \int_a P(a,p,K|D). \quad (2.4)$$

The function $q(p,K)$ $(\equiv f(p)dp)$ is, for a specified model K, the prior probability for the source strengths p_j, knowledge of which we assume is logically independent of knowledge of the parameters a_{ji}.

It is convenient to define the global likelihood function $l(D|p,K)$ by

$$l(D|p,K) \equiv \int_a L(D|a,p,K) Q(a,K), \quad (2.5)$$

and write Eq. (2.4) as

$$P(p,K|D) = \frac{l(D|p,K)\, q(p,K)}{\sum_K \int_p l(D|p,K)\, q(p,K)}. \quad (2.6)$$

With our choices for the prior probability, Eq. (2.2), and the likelihood function, Eq. (2.1), it is possible to perform the $N \times M$-dimensional integral in Eq. (2.4) exactly and obtain the formula

$$l(D|p,K) = \prod_{i=1}^{M} \sum_{k_1,\ldots,k_N=0}^{D_i} \prod_{j=1}^{N} \binom{A_{ji}+k_j}{k_j} \times \frac{p_j^{k_j}}{(1+p_j)^{A_{ji}+k_j+1}}, \quad (2.7)$$

where the indices k_j satisfy the multinomial constraint $\sum_{j=1}^{N} k_j = D_i$. The calculation is outlined in the appendix.

We again stress the importance of being clear about the conceptual basis of the method; in particular, we should understand what $P(p,K|D)$ is and what it is

not. The function $P(p, K|D)$ describes, in a probabilistic manner, what we know about the parameters after having acquired a particular data set $\{D_i\}$ and after having performed a particular set of Monte Carlo calculations, leading to a particular distribution of bin counts $\{A_{ji}\}$. It does *not* describe the sampling distribution of the parameters p_j. The source strengths p_j are presumed to have fixed values, albeit unknown. If nonetheless we wish to interpret $P(p, K|D)$ in frequency terms we would have to posit an ensemble of hypothetical universes each with differing sets of fixed source strengths. We see, however, neither a conceptual nor a practical advantage in this artifice over simply interpreting $P(p, K|D)$ as a weight between zero and one that describes how well we know the parameters p_j after we have acquired a particular data set. Likewise, the prior probability $q(p, K)$ is a weight we assign consistent with whatever pertinent information we might have about the parameters p_j, irrespective of the information provided by the data set D.

3. Estimating physical quantities

When we have several models, each labelled by the parameter K, we can calculate the probability of each model K by marginalizing $P(p, K|D)$ with respect to p:

$$P(K|D) = \int_p P(p, K|D). \tag{3.1}$$

An interesting application of Eq. (3.1) is when K labels the elements of a set of models that differ in the value of some physical quantity; for example, the top quark mass. In the case of top anti-top events, formed in the reaction $p\bar{p} \to t\bar{t}$, $P(K|D)$ would pick out the background plus signal model with the top quark mass that best fits the data. Moreover, an optimal estimate of the physical quantity, in the sense that the mean squared deviation from the true value is minimized, is the mean of the posterior probability. Therefore, we would expect to obtain a good estimate of the top quark mass and an estimate of the uncertainty from

$$\hat{m} = \sum_K m_K P(K|D),$$

$$\sigma^2_{\hat{m}} = \sum_K m_K^2 P(K|D) - \hat{m}^2, \tag{3.2}$$

where m_K is the assumed top quark mass for model K and $\sigma_{\hat{m}}$ is one (of many) measures of the width of the posterior distribution (assuming uniformly spaced m_K). This measure of uncertainty does not have a frequency interpretation because $P(K|D)$ is not a sampling distribution.

An alternative way to proceed would be to use the *evidence* procedure [4] which, for our case, entails inserting the maximum likelihood estimates of p, $\hat{p}$ (obtained by maximizing the global likelihood $l(p, K|D)$), into the posterior probability $P(p, K|D)$ and then using either the mean of $P(\hat{p}, K|D)$, or the position of its peak as an estimate. The global likelihood $l(p, K|D)$ is maximized at the point $\hat{p}$ such that

$$\sum_{j=1}^{N} \hat{p}_j \left(\sum_{i=1}^{M} A_{ji} + M \right) = \sum_{i=1}^{M} D_i. \tag{3.3}$$

4. Systematic uncertainty

There is no well-founded procedure to deal with systematic uncertainty in frequentist statistics. One either resorts to the artifice mentioned earlier or one abandons conceptual consistency and grafts Bayesian notions onto frequentist procedures [5]. In the Bayesian approach systematic uncertainty can be treated in a unified consistent manner. It is also straightforward and requires merely a reinterpretation of the label K.

To render the discussion more concrete let us suppose that we have generated a series of models K that differ not only in the physical quantity of interest, here the top quark mass, but also in the value of the renormalization scale used to calculate the models. The usual practice is to calculate the models at a small number of different scales. The renormalization scale is an example of a nuisance parameter that is unphysical and whose value is arbitrary. To the degree that calculations are sensitive to the renormalization scale the arbitrariness of the latter will introduce further uncertainty in the models. That uncertainty, however, can be accounted for by simply summing Eq. (2.6) over the models K that differ only by the value of the assumed scale. That is, for a given top quark mass, we

marginalize $P(p, K|D)$ with respect to the renormalization scale. To take into account the uncertainty due to all models considered, we marginalize with respect to all models K:

$$P(p|D) = \sum_K P(p, K|D). \tag{4.1}$$

Thus can we account for all uncertainties irrespective of how we label them: statistical, systematic or theoretical.

5. Estimating the top quark mass

To illustrate the method we apply it to the problem of estimating the top quark mass (m_t), assuming a set of signal plus background models. For this special case Eq. (2.7) can be written as

$$l(D|p_1, p_2, K) = \prod_{i=1}^{M} \sum_{k=0}^{D_i} C_{k,1}\, C_{D_i-k,2}, \tag{5.1}$$

where the terms C may be calculated using the recursion formula

$$C_{0,j} = (1 + p_j)^{-(A_{ji}+1)}, \tag{5.2}$$

$$C_{r,j} = \left(\frac{p_j}{1+p_j}\right)\left(\frac{A_{ji}+r}{r}\right) C_{r-1,j},$$

$$(r = 1 \ldots D_i,\ j = 1, 2),$$

which is convenient for numerical calculations.

The top quark was discovered recently in proton anti-proton collisions at the Fermilab Tevatron by the CDF and DØ collaborations [6]. At present, the most accurate measurement of the top quark mass comes from the analysis of the decay mode $t\bar{t} \to W^+bW^-\bar{b} \to l\nu bq\bar{q}\bar{b}$ where one W boson decays into a lepton (either a muon or an electron) and a neutrino, and the other W boson decays into a quark anti-quark pair. The dominant background in this decay mode comes from the quantum chromodynamic (QCD) production of a W boson in association with multiple jets (W+jets).

To obtain an estimate of the top quark mass one can use any kinematic quantity in the event that depends on the mass. For simplicity, we have generated hypothetical distributions of fitted masses for signal and background models to roughly simulate the data sets obtained by the CDF and DØ experiments. The CDF and DØ fitted mass data sets are derived by fitting each observed event to the top quark decay hypothesis. For the signal, our hypothetical fitted mass distributions are taken to be Gaussian with mean at the top quark mass and a standard deviation of 30 GeV/c^2. We have generated 25 such distributions for top quark masses in the range 110 GeV/c^2 to 230 GeV/c^2, in steps of 5 GeV/c^2. To model the background we superpose, in the ratio of 10 to 3, two Gaussian distributions centered at 110 GeV/c^2 and 140 GeV/c^2, and with standard deviations of 15 GeV/c^2 and 25 GeV/c^2, respectively. The simulated data are binned in forty uniform bins in the mass range of 80 GeV/c^2 to 280 GeV/c^2. The simulated distributions of signal (for m_t = 170 GeV/c^2) and background are shown in Figs. 1(a)–(b).

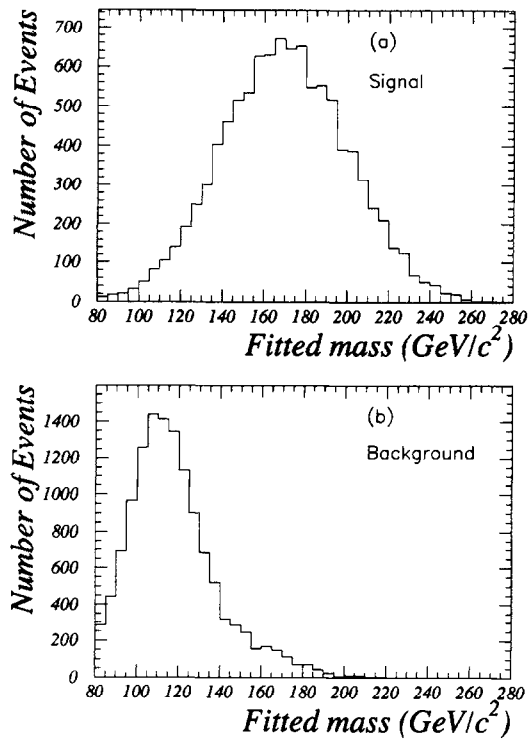

Fig. 1. The distributions of fitted mass for simulated (a) top quark events with mass of 170 GeV/c^2 and (b) background.

We then generated data sets of increasing sample size by random sampling from the signal (m_t = 170 GeV/c^2) and background fitted mass distributions. We use a signal to background ratio of one and we use binomially distributed counts. The posterior probability

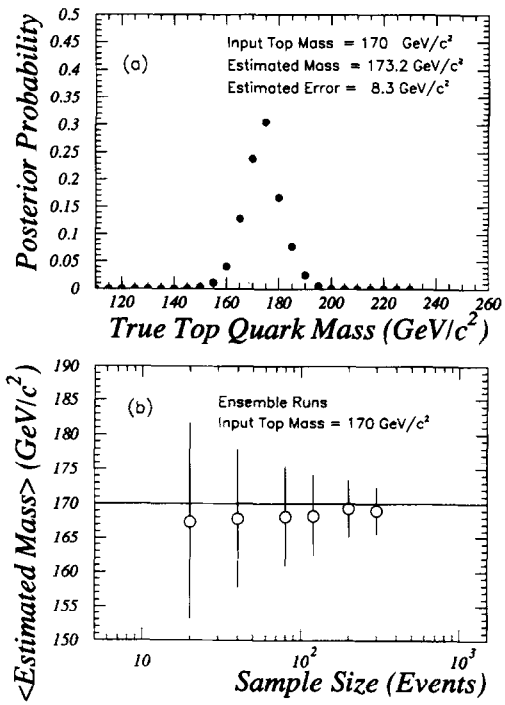

Fig. 2. (a) The posterior probability distribution for one hypothetical experiment with a sample size of 40 events and a signal to background ratio of one. The estimated top quark mass and error are also shown. (b) The estimated top quark mass, as a function of sample size, averaged over ensembles of 200 experiments. The error bars indicate the 68% widths of the distributions of mass estimates.

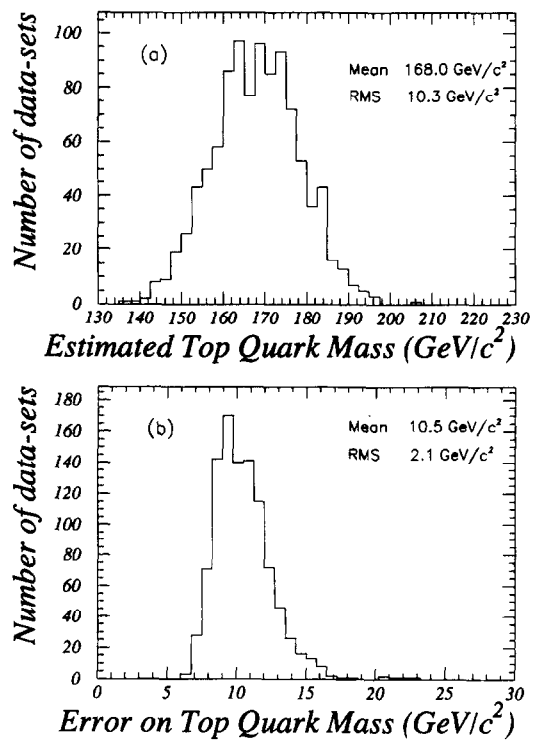

Fig. 3. Distributions of (a) estimated top quark mass and (b) estimated RMS error on the mass from a study of an ensemble of 1000 hypothetical experiments. Note that the most probable error is approximately equal to the standard deviation of the mass distribution.

$P(p, K|D)$, Eq. (2.6), is evaluated for each data set for each signal plus background model using Eq. (5.1) and taking $q(p, K)$ to be uniform. From Eq. (3.1), the posterior probability distribution, $P(K|D)$, is obtained as a function of the assumed top quark mass. We estimate the mass and error using Eq. (3.2). The results are shown in Fig. 2. It can be seen that as the data set grows in size the estimated top quark mass converges to the true top quark mass and the uncertainty in the mass estimate reduces.

To demonstrate that the method produces reliable results on average even for small data sets we have carried out ensemble studies. We generated an ensemble of 1000 data sets (for $m_t = 170\ \mathrm{GeV/c^2}$). The sample size for each data set is 40 events and the signal to background ratio is chosen to be one as before. In Fig. 3 we show the distributions of estimated top quark masses and errors. The estimated top quark mass peaks around the true top quark mass and the most probable error is approximately equal to the standard deviation of the distribution of estimated masses.

We noted above that the strength p_j is given as a fraction of the true total source count $\sum_i a_{ji}$. But how should we proceed if we wish to have an estimate of the mean number of events from source j? Let n_j denote that quantity. By definition, $n_j \equiv p_j \sum_i a_{ji}$. Therefore, to get an estimate $\hat{n}_j$ of n_j we need an estimate of $\sum_i a_{ji}$. An obvious estimate is $\sum_i A_{ji}$. Another, less obvious, one – suggested by Eq. (3.3) – is $\sum_i A_{ji} + M$, where M is the number of bins. The two estimates merge when $\sum_i A_{ji} \gg M$, which is the most common situation. It is an open question (which we are currently investigating) whether it is possible to derive a *useful* exact expression for the posterior probability $P(n, K|D)$ rather than $P(p, K|D)$. If so, one would be able to derive another estimate of n_j from the marginal posterior probability $P(n|D)$.

6. Conclusions

From the Bayesian perspective the frequentist method has a simple interpretation. The method is equivalent to 1) choosing a flat prior probability for the parameters p (without, however, restricting the values these parameters might assume), a flat prior for the discrete parameter K, and a gamma prior (as described above) for the nuisance parameters a; and 2) finding the *mode* of the posterior probability $P(a,p,K|D)$. The uncertainty in p, however, is obtained using the sampling distribution of the estimates $\hat{p}$. This is simply one of several different estimates that could be derived from the posterior probability $P(a,p,K|D)$. In our method we have restricted the parameters p to be always positive and we compute the mean, rather than the mode, not of the full posterior probability $P(a,p,K|D)$ but rather of the marginal distributions of p, that is, of $P(p,K|D)$.

Bayesian reasoning leads to a well-founded mathematical procedure to treat *all* uncertainties, to combine results and to compute the conditional probability of a model. We have given a simple, useful and practical application of these ideas.

Acknowledgments

This work was stimulated by the authors' participation in the search for and study of the top quark at the DØ experiment. We thank Herb Greenlee for suggesting useful improvements to the paper. This work was supported in part by the US Department of Energy.

Appendix A

Eq. (2.5) can be written as

$$l(D|p,K) = \prod_{i=1}^{M} \frac{1}{D_i!} \int_0^\infty da_{1i} \frac{\exp[-(1+p_1)a_{1i}]a_{1i}^{A_{1i}}}{A_{1i}!} \cdots \int_0^\infty da_{Ni} \frac{\exp[-(1+p_N)a_{Ni}]a_{Ni}^{A_{Ni}}}{A_{Ni}!} \left(\sum_j^N p_j a_{ji}\right)^{D_i}. \tag{A.1}$$

Expanding the sum over sources gives

$$\left(\sum_{j=1}^{N} p_j a_{ji}\right)^{D_i} = D_i! \sum_{k_1,\ldots,k_N=0}^{D_i} \frac{p_1^{k_1} a_{1i}^{k_1}}{k_1!} \cdots \frac{p_N^{k_N} a_{Ni}^{k_N}}{k_N!}, \tag{A.2}$$

which when inserted into the equation above leads to

$$l(D|p,K) = \prod_{i=1}^{M} \sum_{k_1,\ldots,k_N=0}^{D_i} \prod_{j=1}^{N} \frac{p_j^{k_j}}{A_{ji}!k_j!} \int_0^\infty da_{ji} \times \exp[-(1+p_j)a_{ji}]a_{ji}^{A_{ji}+k_j}, \tag{A.3}$$

where, for each count D_i, the k_j satisfy the multinomial constraint $\sum_{j=1}^{N} k_j = D_i$. The $N \times M$ dimensional integral separates thus into $N \times M$ one-dimensional integrals that are readily evaluated in terms of gamma functions. When this is done we obtain Eq. (2.4).

References

[1] R. Barlow and C. Beeston, Comp. Phys. Comm. 77 (1993) 219.
[2] M. Kendall and A. Stuart, The Advanced Theory of Statistics, C. Griffin & Co. Ltd., London (1973).
[3] E.T. Jaynes, Probability Theory – The Logic Of Science, http://www.math.albany.edu:8008/JaynesBook.html (1995).
[4] D.J.C. Mackay, Hyperparameters: Optimise, or integrate out?, in: Maximum Entropy and Bayesian Methods, Santa Barbara (1993), ed. G. Heidbreder (Kluwer, Dordrecht).
[5] R.D. Cousins, Am. J. Phys. 63 (1995) 398.
[6] F. Abe et al. (CDF Collaboration), Phys. Rev. Lett. 74 (1995) 2626;
S. Abachi et al. (DØ Collaboration), Phys. Rev. Lett. 74 (1995) 2632.