A Measure of the Goodness of Fit in Unbinned Likelihood Fits

Rajendran Raja

Fermi National Accelerator laboratory

Batavia, IL 60510

Abstract

Maximum likelihood fits to data can be done using binned data (histograms) and unbinned data. With binned data, one gets not only the fitted parameters but also a measure of the goodness of fit. With unbinned data, currently, the fitted parameters are obtained but no measure of goodness of fit is available. This remains, to date, an unsolved problem in statistics. Using Bayes theorem and likelihood ratios, we provide a method by which both the fitted quantities and a measure of the goodness of fit are obtained for unbinned likelihood fits, as well as errors in the fitted quantities. We provide an ansatz for determining Bayesian *a priori* probabilities.

1 Introduction

We outline a method by which goodness of fit measures can be calculated in an unbinned likelihood analysis. We are able to also calculate the probability

Email address: `raja@fnal.gov` (Rajendran Raja).

density function of the fitted variables and hence their errors in a rigorous manner. We briefly describe the currently used method of “maximum likelihood”, originally due to R.A.Fisher [1]. Let s denote a set parameters defining our theoretical model used to describe data. Example of s are the mass of the top quark or the lifetime of a particle. The symbol s (for signal) can in general denote a discrete or continuous set of variables. Let c denote a set of observations describing a high energy physics event and there are n events in our dataset. In general, for each event, c can be a vector of dimension d. Let $P(c|s)dc$ describe the probability of observing the configuration c in the d-dimensional phase space volume dc given the theoretical parameter set s. Thus $P(c|s)$ is a probability density function (*pdf*) in the variable c and obeys

$$\int P(c|s)dc = 1 \tag{1}$$

Then one can define a likelihood $\mathcal{L}$ of observing the dataset as

$$\mathcal{L} = \prod_{i=1}^{i=n} P(c_i|s) \tag{2}$$

The maximum likelihood point can be found of observing by minimizing the negative log-likelihood $-log_e\mathcal{L}$ defined as

$$-log_e\mathcal{L} = -\sum_{i=1}^{i=n} log_e(P(c_i|s)) \tag{3}$$

while varying the parameters s either analytically or numerically to obtain the best values s^* of s that fit the data.

At the maximum likelihood point, s^*, the best fit values of s, are obtained. There is however no measure of the goodness of fit, since the likelihood at the optimal value is not normalized to anything. There is strictly no measure for the error on the fitted parameters, since $\mathcal{L}$ is not a probability density function of s, though people have calculated errors by treating the $-log_e\mathcal{L}$ at the minimum as though it were equivalent to $\frac{1}{2}\chi^2$. Such error calculations are not hitherto considered rigorously justifiable.

Unbinned likelihood fits, despite these disadvantages, are extremely useful in finding s^* since one does not have to treat bins with small populations in a special manner as would be the case for binned fits.

In this paper we use Bayes theorem to rectify the above disadvantages. In the process, we obtain a measure for the goodness of fit and also $P(s|c)$, the posterior *pdf* of s, enabling us to calculate the errors of the fitted values in a rigorous way.

2 Bayes Theorem

We derive Bayes theorem here for the sake of completeness and to illustrate the main ideas. In the Bayesian approach [2], the theoretical parameters s can have a probability distribution both *a priori* and *a posteriori*. The *a priori* distribution refers to the knowledge of s before the given set of observations are made. The *a posteriori* probability distribution refers to the distribution of s, given the set of observations c.

2.1 Joint and Conditional Probabilities

We define a joint probability density for the theory parameters s and the observables c as

$$dP_{joint} = P_{joint}(s, c)ds\, dc \tag{4}$$

which is the probability that s occurs in interval s and $s + ds$ and c occurs in a volume element dc centered around c.

We define the conditional probability density

$$dP_{conditional} = P(s|c)ds \tag{5}$$

as the probability density of observing s in the interval s and $s+ds$ *given* that c occurs in a volume element dc centered around c .

Similarly, the conditional probability density

$$dP_{conditional} = P(c|s)dc \tag{6}$$

is defined as the probability density of observing c in a volume element dc centered around c, *given* that s occurs between s and $s + ds$. Then, by the laws of probability , we can write the joint probability

$$dP_{joint} = P_{joint}(s, c)ds\, dc = P(c|s)dc \times P(s)ds \tag{7}$$

Where $P(s)$ is the *a priori* probability of observing s in interval s and $s + ds$, and $P(c|s)dc$ is the probability of observing c given s . One can also obtain the same joint probability, by first observing c with *a priori* probability $P(c)$

and then using the conditional probability $P(s|c)$, i.e.

$$dP_{joint} = P_{joint}(s,c)dsdc = P(s|c)ds \times P(c)dc \tag{8}$$

By equating 7 and 8, one gets the fundamental relation leading up to Bayes Theorem.

$$dP_{joint} = P(c|s)dc \times P(s)ds = P(s|c)ds \times P(c)dc \tag{9}$$

Expressed in terms of densities, dropping ds and dc terms, this yields

$$P(c|s) \times P(s) = P(s|c) \times P(c) \tag{10}$$

One is interested in evaluating $P(s|c)$, the probability of the theory parameters, given a set of observations c. This becomes,

$$P(s|c) = \frac{P(c|s) \times P(s)}{P(c)} \tag{11}$$

The *a priori* probability $P(c)$ is not an independent quantity, given the *a priori* probability $P(s)$ which represents the knowledge of s before the set of observations c . The reason for this is that $P(s|c)$ integrated over s must add up to unity.

2.2 Some Normalization Formulae

Integrating over one of the variables in the joint probability yields, using equation 7, the following relations.

$$P(c) \equiv \int P_{joint}(s,c)ds = \int P(c|s) \times P(s)ds \tag{12}$$

where the $\equiv$ sign is the definition of the *a priori* probability $P(c)$, since one integrates the joint probability $P_{joint}(s,c)$ over all values of s . This then yields

$$P(c) = \int P(c|s) \times P(s) ds \tag{13}$$

also, integrating the joint probability over c, one gets

$$P(s) \equiv \int P_{joint}(s,c) dc = \int P(c|s) \times P(s) dc \tag{14}$$

i.e.

$$P(s) = \int P(c|s) \times P(s) dc \tag{15}$$

$$or \int P(c|s) dc = 1 \tag{16}$$

Similarly, using equation 8, one gets relations similar to the above with c and s interchanged. Summarizing, one gets the following normalization relations.

$$P(c) = \int P(c|s) \times P(s) ds \tag{17}$$

$$P(s) = \int P(s|c) \times P(c) dc \tag{18}$$

$$\int P(s) ds = 1 \tag{19}$$

$$\int P(c) dc = 1 \tag{20}$$

$$\int P(c|s) dc = 1 \tag{21}$$

$$\int P(s|c) ds = 1 \tag{22}$$

Substituting 17 in 11, one gets the derivation of Bayes Theorem.

$$P(s|c) = \frac{P(c|s) \times P(s)}{\int P(c|s) \times P(s) ds} \tag{23}$$

The above equation normalizes to unity as per equation 22. This is the central expression of Bayes' theorem.

2.3 Observation of Many Configurations

Now we come to one of the more beautiful properties of formula 23, namely it is recursive. Let us observe two separate configurations say, c_1 and c_2 . Then equation 23 yields for c_1 ,

$$P(s|c_1) = \frac{P(c_1|s) \times P(s)}{\int P(c_1|s) \times P(s)ds} \tag{24}$$

Now we observe c_2. We wish to compute $P(s|c_1, c_2)$, the probability of s given c_1 and c_2 . We can then replace the *a priori* probability for s, $P(s)$ in equation 23 by the probability of s after observing c_1 (i.e. $P(s|c_1)$) to calculate the probability of s given c_1 and c_2 . This yields,

$$P(s|c_1, c_2) = \frac{P(c_2|s) \times P(s|c_1)}{\int P(c_2|s) \times P(s|c_1)ds} \tag{25}$$

Substituting for $P(s|c_1)$ from equation 24, we get

$$P(s|c_1, c_2) = \frac{P(c_2|s)P(c_1|s)P(s)/\int P(c_1|s)P(s)ds}{\int P(c_2|s)P(c_1|s)P(s)ds/\int P(c_1|s)P(s)ds} \tag{26}$$

$$\text{yielding } P(s|c_1, c_2) = \frac{P(c_2|s)P(c_1|s)P(s)}{\int P(c_2|s)P(c_1|s)P(s)ds} \tag{27}$$

$$\text{generalizing, } P(s|c_1, c_2 ... c_n) = \frac{P(c_n|s)...P(c_2|s)P(c_1|s)P(s)}{\int P(c_n|s)...P(c_2|s)P(c_1|s)P(s)ds} \tag{28}$$

Another way to think about equation 28 is to think of the n configurations as one massive super configuration $\mathbf{c_n}$, which also obeys the Bayes theorem equation 23

$$P(s|\mathbf{c_n}) = \frac{P(\mathbf{c_n}|s) \times P(s)}{\int P(\mathbf{c_n}|s) \times P(s)ds} \tag{29}$$

$$\text{where } P(\mathbf{c_n}|s) = P(c_n|s)...P(c_2|s)P(c_1|s) \tag{30}$$

It should be noted that the probability $P(\mathbf{c_n}|s)$ obeys the normalization condition 21. Equation 30 is just the law of multiplication of independent probabilities. This implies that it is possible to chain probabilities in Bayes theorem as in equation 25 if and only if the configurations are statistically independent. This is certainly true in the case of high energy physics events.

The expression for *a posteriori* probability $P(s|\mathbf{c_n})$ in equation 29 cannot be used as is unless one knows P(s), the *a priori* probability of s. In the "Bayesian approach", people use various guesses for P(s) and a lot of care and energy are expended in arriving at "reasonable" functions for P(s).

2.4 Likelihood Ratios

We now recast the Bayes theorem equation 29 as a set of likelihood ratios $\mathcal{L}_{\mathcal{R}}$.

$$\mathcal{L}_{\mathcal{R}} = \frac{P(s|\mathbf{c_n})}{P(s)} = \frac{P(\mathbf{c_n}|s)}{P(\mathbf{c_n})} \tag{31}$$

where we have substituted the function $P(\mathbf{c_n})$ for the normalizing integral in the denominator using equation 13. The likelihood ratio $\mathcal{L}_{\mathcal{R}}$ has a very important invariant property. It is invariant under the transformations of variable sets $c \to c'$ and $s \to s'$ where c' and s' are functions of the variable sets c and s. It is possible to ask what exact variables one uses to form the vector c. For instance,when a jet is measured experimentally, does one use the energy, pseudo-rapidity and azimuth of the jet or the three components of the energy three vector as components of c? Clearly, the probability density function $P(c|s)$ will depend on the choice of the variable set c since,

$$P(c'|s) = |\frac{dc}{dc'}|P(c|s) \tag{32}$$

where, $|\frac{dc}{dc'}|$ denotes a Jacobian of transformation to go from the set of variables c to c'. However, the same Jacobian ocurs in the denominator of the $\mathcal{L}_{\mathcal{R}}$, hence the likelihood ratio is unaffected by the transformation. The same argument can be made with respect to transformations of the variable set $s \to s'$. These are extremely important properties, so we henceforth work with the likelihood ratio $\mathcal{L}_{\mathcal{R}}$ and not the likelihoods $\mathcal{L}$ which do not possess these properties.

3 The Principle of Maximum Likelihood Ratios

The equation 31 for $\mathcal{L}_{\mathcal{R}}$ can be expanded as follows.

$$\mathcal{L}_{\mathcal{R}} = \frac{P(\mathbf{c_n}|s)}{P(\mathbf{c_n})} = \frac{P(c_1|s)}{P(c_1)} \times \frac{P(c_2|s)}{P(c_2)} \ldots \times \frac{P(c_n|s)}{P(c_n)} \tag{33}$$

where we have used the independence of *a priori* probabilities for $P(c_i), i = 1, n$. Similarly, one gets expressions,

$$\mathcal{L}_{\mathcal{R}} = \frac{P(s|\mathbf{c_n})}{P(s)} = \frac{P(s|c_1)}{P(s)} \times \frac{P(s|c_2)}{P(s)} \ldots \times \frac{P(s|c_n)}{P(s)} \tag{34}$$

where we have derived equation 34 from equation 33 by applying equation 31 to the likelihood ratios of the individual events in the product. In order to find the optimal set of parameters s, we maximize the likelihood ratio $\mathcal{L}_{\mathcal{R}}$ in equation 33 with respect to s. This is equivalent to minimizing the negative log likelihood ratio $log_e\mathcal{L}_{\mathcal{R}}$.

$$-\frac{\partial log_e \mathcal{L}_{\mathcal{R}}}{\partial s} = -\sum_{i=1}^{i=n} \frac{\partial log_e P(c_i|s)}{\partial s} = 0 \tag{35}$$

Notice that this is the same set of equations that one gets when maximizing the likelihood as in equation 3, since the *a priori* probabilities $P(c_i)$ are constant with respect to variations in s. So one gets the same set of optimal variables s^* whether one maximizes the likelihood $\mathcal{L}$ or the likelihood ratio $\mathcal{L}_{\mathcal{R}}$. However, at the optimum, the likelihood ratio can be used to obtain a goodness of fit parameter as we show below, whereas the likelihood method would be unable to provide this information. One can now ask what the minimum value of $\mathcal{L}_{\mathcal{R}}$ is with respect to variations in the event configuration, for a fixed value of theory; i.e. what event configurations produce the minimum value of the negative log likelihood? Differentiating equation 33 with respect to c_i, one gets,

$$-\frac{\partial log_e \mathcal{L}_{\mathcal{R}}}{\partial c_i} = -\frac{\partial log_e P(c_i|s)}{\partial c_i} + \frac{\partial log_e P(c_i)}{\partial c_i} = 0 \tag{36}$$

i.e

$$\frac{\partial log_e P(c_i|s)}{\partial c_i} = \frac{\partial log_e P(c_i)}{\partial c_i} \tag{37}$$

$$P(c_i|s) = P(c_i) \tag{38}$$

The equation 38 implies that the lowest value of the likelihood ratio occurs when the experimental probability density $P(c)$ and the theory probability density $P(c|s)$ are the same at the observed events. The negative log likelihood is zero at this point, yielding the best possible fit.

4 Evaluating the Function $P(\mathbf{c})$ and the Goodness of Fit

The key point to note is that just as $P(s)$ is the *a priori* probability of the theoretical parameter s, $P(c)$ is the *a priori* probability of the data. In order to evaluate the likelihood ratio $\mathcal{L}_{\mathcal{R}}$ at the maximum likelihood point, one needs to evaluate the function $P(c)$ at the observed event configurations $c_1, c_2 \ldots c_n$.

So the problem to solve is this: given the event configurations $c_1, c_2 \ldots c_n$, what is their probability density? Well known methods exist to estimate the $pdf's$ given discrete event distributions. These are collectively titled probability density estimators (PDE), which have recently found application in high energy physics analyses [3].

As noted above, the probability density function $P(c)$ is the *a priori pdf* of the data. In previous applications, to the author's best knowledge, the function $P(c)$ was subsumed into the equation 13 and expressed in terms of an unknown $P(s)$. This resulted in the theory *pdf* $P(c|s)$ being evaluated at the data points $c_1, c_2 \ldots c_n$, but not the data *pdf*! It is precisely this failure to evaluate $P(c)$ given c that has led to the absence of goodness of fit criteria in unbinned likelihood fits.

In binned likelihood fits, one fits a theoretical curve to a binned set of data points. Two distributions, those of theory and data, are involved in providing a goodness of fit measure such as χ^2 in the binned approach. In the unbinned method, however, one finds the maximum likelihood point by evaluating the theoretical function $P(c|s)$ at the data points $c_i, i = 1, n$. There is only one distribution involved, namely theory! One has hitherto ignored $P(c)$, by subsuming it into a normalization constant. We rectify this lapse here and restore $P(c)$ to its proper role, namely, the *pdf* of the data.

4.1 Probability Density Estimators

Let c_i^α, $\alpha = 1, d$ denote the components of the d-dimensional vector c for the ith event. Then we can define the mean vector $< c^\alpha >$ as

$$< c^\alpha >= \frac{1}{n} \sum_{i=1}^{i=n} c_i^\alpha \tag{39}$$

The covariance (or error) matrix E of c is defined as

$$E^{\alpha,\beta} =< c^\alpha c^\beta > - < c^\alpha >< c^\beta > \tag{40}$$

where the $<>$ implies average over the n events. The Hessian matrix H is defined as the inverse of E. One can define a multivariate Gaussian Kernel $\mathcal{G}(c)$ as

$$\mathcal{G}(c) = \frac{1}{(\sqrt{2\pi}h)^d \sqrt{(det(H))}} exp(\frac{-H^{\alpha\beta} c^\alpha c^\beta}{2h^2}) \tag{41}$$

where the repeated indices imply summing over and the parameter h is a "smoothing parameter", which has[4] a suggested optimal value $h \approx n^{-1/(d+4)}$. The *pdf* of c is then given by

$$P(c) \approx PDE(c) = \frac{1}{n} \sum_{i=1}^{i=n} \mathcal{G}(c - c_i) \tag{42}$$

Simply put, one takes an arbitray point c in configuration space, calculates the separation from this point to all the measured points and sums up the values (at c) of the Gaussians that are centered at the measured points. This sum is divided by the number of Gaussians, which equals n. Since the Gaussians are all normalized to unity, $P(c)$ obeys equation 20. One can feed in any value of c and the PDE will provide a probability density at that value of

c. It is clear that the PDE method is generalizable to arbitrary dimensions and will provide us with $P(c)$. One should note that the Gaussian Kernel function depends on n, the number of events in the sample. This dependence is through the smoothing parameter, which goes to zero as $n \to \infty$. In this limit, equation 42 becomes

$$P(c) = \int P(c)\mathcal{G}_\infty(c - c_i)dc_i \tag{43}$$

This implies that

$$\mathcal{G}_\infty(c - c_i) \equiv \lim_{n\to\infty} \mathcal{G}(c - c_i) = \delta(c - c_i) \tag{44}$$

There exist generalizations [5] of the above scheme where the covariance matrix is made locally variable that can estimate $pdf's$ with greater complexity albeit at a cost in computing speed. In what follows, we introduce a method by which the smoothing factor can be made a function of the configuration variables c in an iterative fashion, which may be equivalent to varying the covariance matrix locally.

5 An Illustrative Example

We illustrate the ideas discussed above with a simple one-dimensional example of events in which the observable c consists of decay times distributed exponentially with a decay constant s=1.0 in arbitrary units. The conditional probability $P(c|s)$ defines our theoretical model and is given by

$$P(c|s) = \frac{1}{s}\exp(-\frac{c}{s}) \tag{45}$$

The PDE one dimensional Gaussian kernel for this simple case would be given by

$$\mathcal{G}(c) = \frac{1}{(\sqrt{2\pi}sh)} \exp(-\frac{c^2}{2s^2h^2}) \tag{46}$$

We generate a thousand events for which the smoothing parameter h is calculated to be 0.125 as per equation[6] $h = 0.5n^{-1/(d+4)}$. Figure 1 shows the generated events, the theoretical curve $P(c|s)$ and the PDE curve $P(c)$ normalized to the number of events.

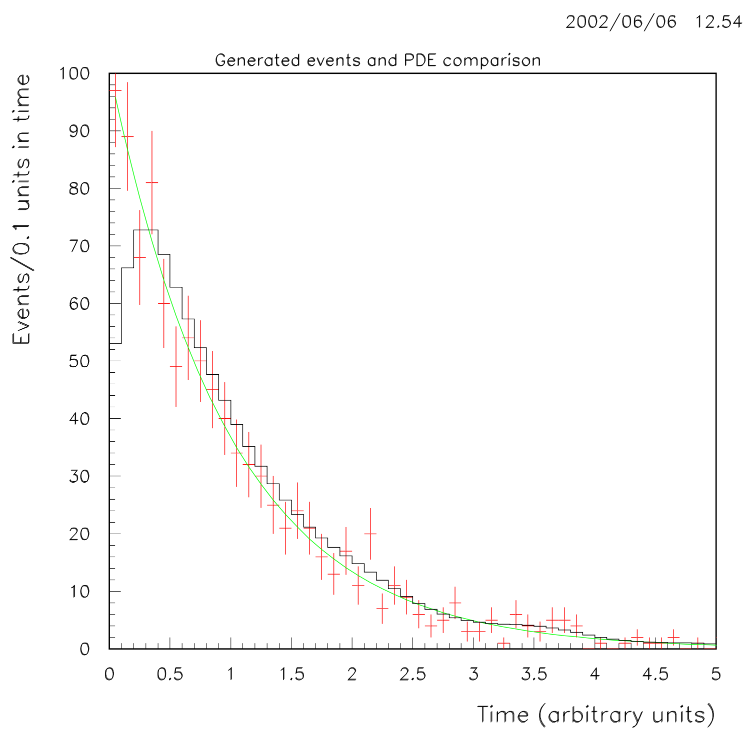

Fig. 1. Figure shows the histogram (with errors) of generated events. Superimposed is the theoretical curve $P(c|s)$ and the PDE estimator (solid) histogram with no errors.

The PDE curve is a poor estimator of the data near the cutoff at c=0. This is because the Gaussians centered at values of negative c would have contributed to the curve near c=0. Also, for large values of c, data are sparse and the

Gaussian approximation with constant smoothing factor h finds it difficult to approximate the data. We choose to restrict our fitting to a fiducial interval in time $t_1 < \ c \ < t_2 = 1 < \ c \ < 5$. Both the theoretical model $P(c|s)$ and the PDE likelihood curves are renormalized so that they integrate to unity in the fiducial interval.

5.1 Iterative Determination of the Smoothing Factor

The expression $h \approx n^{-1/(d+4)}$ clearly is meant to give a smoothing factor that decreases slowly with increased statistics n. It is expected to be true on average over the whole distribution. However, the exponential distribution under consideration has event densities that vary by orders of magnitude as a function of the time variable c. In order to obtain a function $h(c)$ that takes into account this variation, we first work out a PDE with constant h and then use the number densities obtained thus [7] to obtain $h(c)$ as per the equation

$$h(c) = \left(\frac{n\,PDE(c)}{(t_2 - t_1)}\right)^{-0.6} \tag{47}$$

The equation is motivated by the consideration that a uniform distribution of events between t_1 and t_2 has a *pdf* $= 1/(t_2 - t_1)$ whereas the real *pdf* is approximated by PDE. The function $h(c)$ thus obtained is used to work out a better $PDE(c)$. This process is iterated three times to give the best smoothing function.

We generate n=1000 events in the fiducial interval. If now we were to superimpose a Gaussian with 500 events centered at c=2.0 and width=0.2 on the data, the PDE estimator will follow the data as shown in Figure 2. This shows that the PDE estimator we have is adequate to reproduce the data, once the smoothing parameter is made to vary with the number density appropriately.

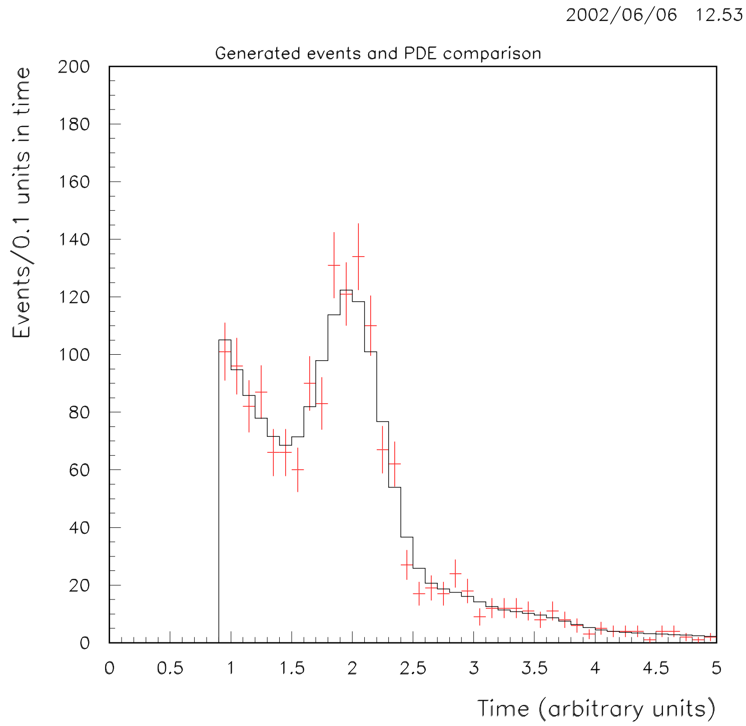

Fig. 2. Figure shows the histogram (with errors) of 1000 events in the fiducial interval $1.0 < c < 5.0$ generated as an exponential with decay constant s=1.0. with a superimposed Gaussian of 500 events centered at c=2.0 and width=0.2. The PDE estimator is the (solid) histogram with no errors.

The smoothing function $h(c)$ for the events in Figure 2 is shown in Figure 3. It can be seen that the value of h increases for regions of low statistics and decreases for regions of high statistics. Superimposed is the constant smoothing factor obtained by the equation $h \approx 0.5n^{-1/(d+4)} = 0.5n^{-0.2}$, with n being the total number of events generated, including those outside the fiducial volume.

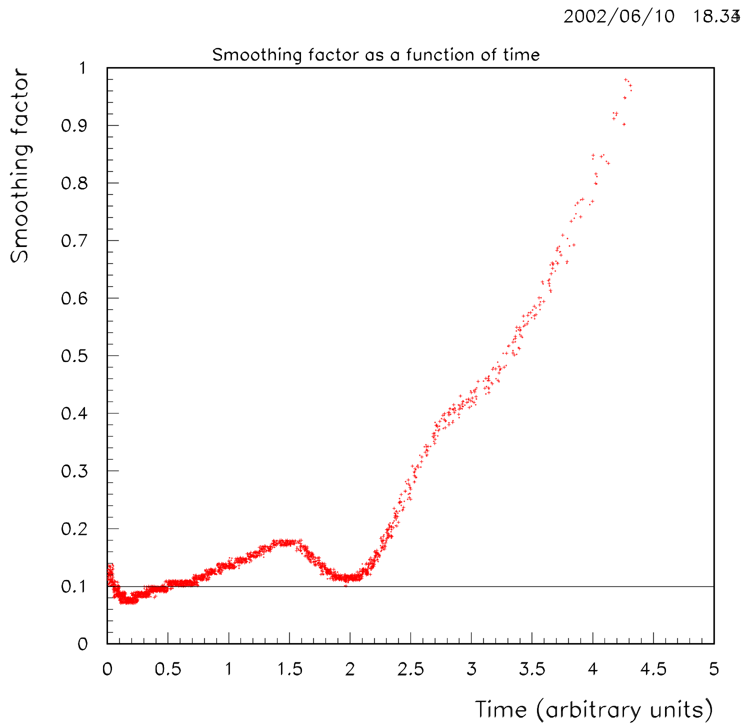

Fig. 3. The variation of h as a function of c for the example shown in Figure 2. The variation of the smoothing parameter is obtained iteratively as explained in the text. The flat curve is a smoothing factor resulting from the formula $h \approx 0.5n^{-1/(d+4)}$.

5.2 An Empirical Measure of the Goodness of Fit

The negative log-likelihood ratio $\mathcal{NLLR} \equiv -log_e\mathcal{L}_\mathcal{R}$ at the maximum likelihood point now provides a measure of the goodness of fit. In order to use this effectively, one needs an analytic theory of the sampling distribution of this ratio. This is difficult to arrive at, since this distribution is sensitive to the smoothing function used. If adequate smoothing is absent in the tail of the exponential, larger and broader sampling distributions of $\mathcal{NLLR}$ will result. One can however determine the distribution of $\mathcal{NLLR}$ empirically, by generating the events distributed according to the theoretical model many times and determining $\mathcal{NLLR}$ at the maximum likelihood point for each such distribution. The solid histogram in figure 4 shows the distribution of $\mathcal{NLLR}$ for 500 such fits. This has a mean of 2.8 and an rms of 1.8. The dotted histogram shows the corresponding value of $\mathcal{NLLR}$ for the constant value of smoothing factor shown in figure 3. This distribution is clearly broader (rms=2.63) with a higher mean(=9.1) and thus has less discrimination power in judging the goodness of fit than the solid curve. We now fit the same exponential back-

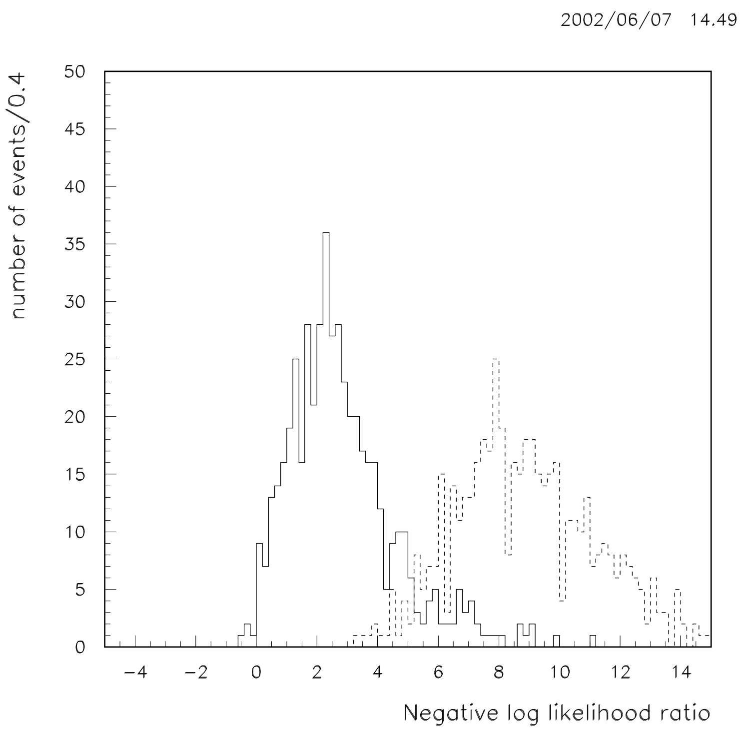

Fig. 4. The solid curve shows the distribution of the negative log likelihood ratio $\mathcal{NLLR}$ at the maximum likelihood point for 500 distributions, using the iterative smoothing function mechanism. The dashed curve shows the corresponding distribution in the case of a constant smoothing function.

ground distribution with different numbers of Gaussian events superimposed on an exponential background. Table 1 shows the results of the fit. When a Gaussian of 500 events with width 0.2 and mean 2.0 is superimposed on the exponential distribution of 1000 events, a value of $\mathcal{NLLR}$=189 is obtained while trying to fit for the exponential using the unbinned maximum likelihood method. This is 103σ away from the mean of the $\mathcal{NLLR}$ distribution shown in figure 4 with the iterated smoothing function. A 3σ effect is observed when the number of events in the Gaussian is 85. Figure 5 shows the generated events, the PDE and the fitted curve for this case.

Let us note that it is possible to make a cumulative function from the solid histogram in figure 4 and estimate the probability that $\mathcal{NLLR}$ exceeds the observed value, just as we do with χ^2 tests. We have also performed a binned χ^2 fit to an exponential over the same histograms, with the data in the fiducial region binned over 41 bins. The resulting value of χ^2 for 39 degrees of freedom are shown in the last column in table 1. At the 3σ point for the unbinned method, the binned method yields a χ^2 of 42 over 39 degrees of freedom, which may be considered a good fit. This implies that the unbinned method may have more discriminating power against bad fits than the binned one. It is worth noting however that the binned fit is over two parameters (the number of events and the slope) whereas the unbinned fit being considered here is only over a single parameter, namely the slope.

We now can fit the exponential data (with no superimposed Gaussian bumps) and compute the value of the likelihood ratio $\mathcal{L}_{\mathcal{R}} = \frac{P(c|s)}{P(c)}$ as a function of the parameter s about the maximum likelihood point. Figure 6 shows this function, which has the maximum value at $s = 1.019$. Note, however, that $\mathcal{L}_{\mathcal{R}}$, a dimensionless quantity, is not the likelihood distribution of s, which has to have the dimensions of $1/s$.

Table 1

Number of Gaussian events	Unbinned fit $\mathcal{NLLR}$	Unbinned fit $N\sigma$	Binned fit χ^2 39 d.o.f.
500	189.	103	304
250	58.6	31	125
100	11.6	4.9	48
85	8.2	3.0	42
75	6.3	1.9	38
50	2.55	-0.14	30
0	0.44	-1.33	24

5.3 Determination of the a priori Likelihood P(s)

In order to obtain the likelihood distribution of s, $P(s|\mathbf{c_n})$, we need to understand better $P(s)$, the *a priori* distribution of s. We are in a position to do this, since we have identified $P(c)$ to be the distribution of data, and P(s) and P(c) are linked by the equation

$$P(c) = \int P(c|s) \times P(s)ds \tag{48}$$

We are in the process of using the *a posteriori* information contained in the data *pdf* $P(c)$ to infer the *a priori* function $P(s)$. Before we use equation 48 to calculate P(s), let us make the following observations. Using equation 31,

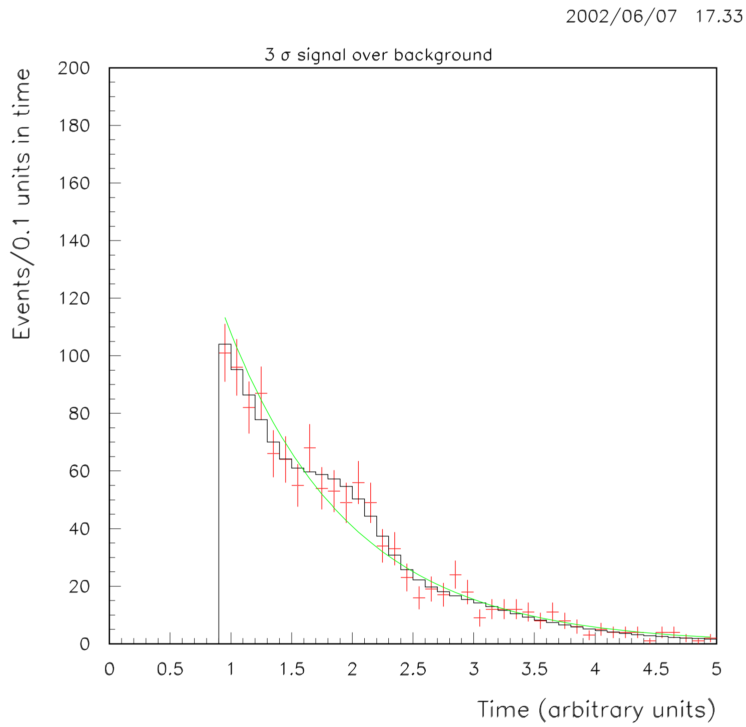

Fig. 5. Figure shows the histogram (with errors) of 1000 events in the fiducial interval $1.0 < c < 5.0$ generated as an exponential with decay constant s=1.0. with a superimposed Gaussian of 85 events centered at c=2.0 and width=0.2. The PDE estimator is the (solid) histogram with no errors. The data are fitted with a goodness of fit that is 3σ away from the average value of $\mathcal{NLLR}$. The continuous curve shows the fit to an exponential.

we can write

$$P(s|\mathbf{c_n}) = P(s) \times \mathcal{L_R} = P(s) \times \frac{P(\mathbf{c_n}|s)}{P(\mathbf{c_n})} \tag{49}$$

As we increase n, the number of events sampled, in the limit $n \rightarrow \infty$ we expect the *a posteriori* probability $P(s|\mathbf{c_n})$ to tend towards the delta function $\delta(s-s^*)$ where s^* is the true value of s. This is because $P(s|\mathbf{c_n})$ is the likelihood distribution of s and we expect to determine the true value of s with infinite precision in this limit. However, the ratio $\frac{P(\mathbf{c_n}|s)}{P(\mathbf{c_n})}$ will tend towards unity in this

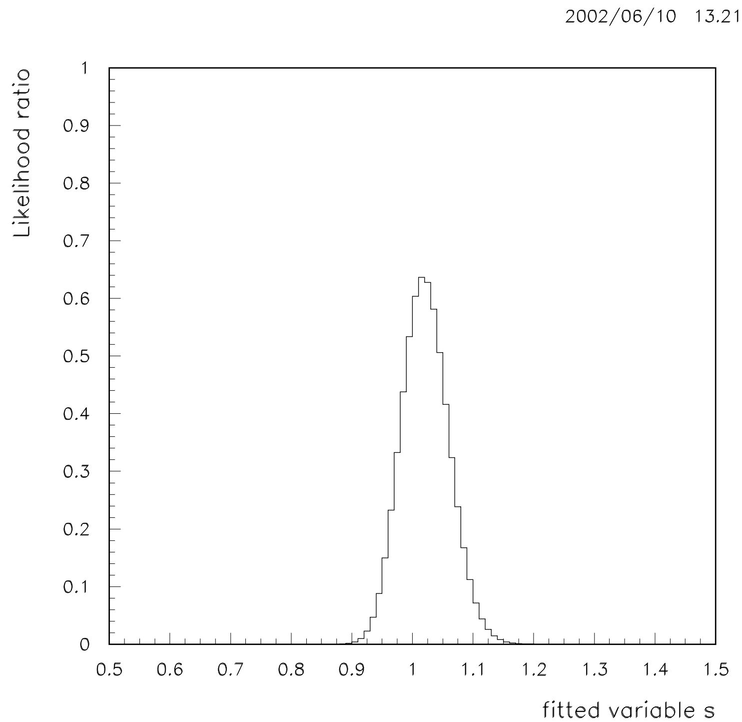

Fig. 6. Figure shows the likelihood ratio $\mathcal{L}_{\mathcal{R}} = \frac{P(c|s)}{P(c)}$ as a function of the fitted parameter s. The maximum likelihood point is at $s = 1.019$.

limit (for a good fit), since for each data point c_k, the theoretical *pdf* $P(c_k|s)$ and the data *pdf* $P(c_k)$ will be close to each other. The only way out of this is to allow $P(s)$ to depend discretely on n and let the distribution $P(s) \rightarrow \delta(s-s^*)$ as $n \rightarrow \infty$. We can see the need for this further, using equation 48. In the limit $n \rightarrow \infty$, the data *pdf* P(c) will have the form $P(c|s^*)$, where s^* is the true value of s, if it fits the theoretical model. Then the only solution for equation 48 is again $P(s) = \delta(s - s^*)$.

To repeat, the only way out of this dilemma, is for the *a priori* probability distribution $P(s)$ to be dependent on n and tend towards a delta function as $n \rightarrow \infty$. If we are solving a Bayes theorem problem for n data points, then the *a priori* function $P(s)$ for that problem will be written as $P_n(s)$ indicating that it comes from a discrete familiy of probability distributions that depend on n. Nothing further is known about $P_n(s)$ *a priori* except that it is a *pdf* in s and that the *pdf* depends discretely on n.

5.3.1 Rewriting the Bayes Theorem Equations

The Bayes theorem equations have to be re-written to take into account this change. Equation 31 now becomes

$$\mathcal{L}_{\mathcal{R},n} = \frac{P(s|\mathbf{c_n})}{P_n(s)} = \frac{P(\mathbf{c_n}|s)}{P(\mathbf{c_n})} \tag{50}$$

Equation 33 remains as is and equation 34 becomes

$$\mathcal{L}_{\mathcal{R},n} = \frac{P(s|\mathbf{c_n})}{P_n(s)} = \frac{P(s|c_1)}{P_1(s)} \times \frac{P(s|c_2)}{P_1(s)} \ldots \times \frac{P(s|c_n)}{P_1(s)} \tag{51}$$

where we have also added the subscript n to the likelihood ratio $\mathcal{L}_{\mathcal{R}}$ to indicate its dependence on n. The recursive chain rule can now be rewritten as

$$\mathcal{L}_{\mathcal{R},k} = \frac{P(s|\mathbf{c_k})}{P_k(s)} = \prod_{i=1}^{i=k} \frac{P(c_i|s)}{P(c_i)} \tag{52}$$

$$\mathcal{L}_{\mathcal{R},l} = \frac{P(s|\mathbf{c_l})}{P_l(s)} = \prod_{i=1}^{i=l} \frac{P(c_i|s)}{P(c_i)} \tag{53}$$

$$\mathcal{L}_{\mathcal{R},k+l} = \mathcal{L}_{\mathcal{R},k} \times \mathcal{L}_{\mathcal{R},l} \tag{54}$$

$$= \frac{P(s|\mathbf{c_k})}{P_k(s)} \times \frac{P(s|\mathbf{c_l})}{P_l(s)} = \frac{P(s|\mathbf{c_{k+l}})}{P_{k+l}(s)} = \prod_{i=1}^{i=k+l} \frac{P(c_i|s)}{P(c_i)} \tag{55}$$

where we have two sub-samples of k and l events which are being combined to form a total number of $k+l$ events.

5.3.2 An Ansatz for $P_n(s)$

The expression for $P(c)$ in equation 48 can be thought of as the theoretical prediction for $P(c)$ and the PDE estimator is the experimental measurement of $P(c)$. Then, one can write,

$$\frac{P^{pred}(\mathbf{c_n})}{P^{exp}(\mathbf{c_n})} = \int \frac{P(\mathbf{c_n}|s)}{P^{PDE}(\mathbf{c_n})} \times P_n(s)ds \tag{56}$$

$$= \int \mathcal{L}_{\mathcal{R},n}(s) \times P_n(s)ds = \int P(s|\mathbf{c_n})ds = 1 \tag{57}$$

with the last expression following from Bayes theorem. There are two ways the last equation $\int P(s|\mathbf{c_n})ds = 1$ can be satisfied. Either the likelihood ratio $\mathcal{L}_{\mathcal{R},n}(s) = 1$ or if

$$P_n(s) = \frac{1}{\int \mathcal{L}_{\mathcal{R},n}(s)ds} \equiv \frac{1}{2\lambda} \tag{58}$$

It is very difficult for $\mathcal{L}_{\mathcal{R},n}(s)$ to equal unity even at the maximum likelihood value, since the experimental PDE estimator in the denominator is subject to statistical fluctuations. Equation 58, however, gives us an expression for the *a priori* likelihood $P_n(s)$. $P_n(s)$ is the value of the *a priori* probability distribution at the true value of s. Since $\int P_n(s)ds = 1$, we can satisfy this with a functional form for $P_n(s)$ being a step function $\theta(s|\mu)$ such that

$$\text{with } s_1 = \mu - \lambda \tag{59}$$

$$\text{and } s_2 = \mu + \lambda \tag{60}$$

$$P_n(s) \equiv \theta(s|\mu) = 0 \text{ if } s < s_1 \text{ or } s > s_2 \tag{61}$$

$$P_n(s) \equiv \theta(s|\mu) = \frac{1}{2\lambda} \text{ if } s_1 \leq s \leq s_2 \tag{62}$$

and the value of μ, the mean of the distribution, is totally unknown. *Let us note that it is possible to write down an equation such as equation 58 only due to the fact that we are able to compute a dimensionless quantity such as the likelihood ratio $\mathcal{L}_{\mathcal{R},n}(s)$ and this becomes possible only due to the concept of the PDE estimator for $P(c)$ introduced in this paper.* The integral in equation 58 has thus the dimensions of s giving $P_n(s)$ the dimensions of s^{-1} as required.

As n increases, the value of λ decreases, since the distribution $P(\mathbf{c_n}|s)$ sharpens. This has the effect of narrowing $P_n(s)$ in accordance with the discussion above. It is important to realize that there is only *one* true value of s, and the Bayesian *a priori* probability $P_n(s)$ refers to the value of $P_n(s)$ at that true value, which, according to our ansatz, equals $\frac{1}{2\lambda}$. The data does not result from an admixture of probable values of s, but from a single *true* value of s. So Bayes theorem becomes,

$$P(s|\mathbf{c_n}) \times P(\mathbf{c_n}) = P(\mathbf{c_n}|s) \times P_n(s) = P(\mathbf{c_n}|s) \times \frac{1}{2\lambda} \tag{63}$$

yielding

$$P(s|\mathbf{c_n}) = \frac{P(\mathbf{c_n}|s)}{P(\mathbf{c_n})} \times \frac{1}{2\lambda} = \frac{\mathcal{L}_{\mathcal{R}}(s)}{\int \mathcal{L}_{\mathcal{R}}(s)ds} = \frac{P(\mathbf{c_n}|s)}{\int P(\mathbf{c_n}|s)ds} \tag{64}$$

The last equation in 64 results from the fact that the PDE estimator for $P(c)$ cancels both in the numerator and denominator. Having obtained $P(s|\mathbf{c_n})$, one can proceed to calculate the statistical quantities associated with s, namely the mean, mode, median, variance, errors and limits, in a rigorous fashion. We note here that $P(s|\mathbf{c_n})$ is obtainable only with the use of of Bayes theorem, and our ansatz for the Bayesian *a priori* likelihood $P(s)$. The expression for $P(s|\mathbf{c_n})$, the *a posteriori* likelihood for s does not depend on the PDE estimator of data, but only on the theoretical function $P(\mathbf{c_n}|s)$ evaluated at the data points. The evaluation of $P_n(s)$ and the goodness of fit criteria both require the usage of the PDE estimator for the data *pdf*.

The ansatz for the *a priori* distribution for $P_n(s)$ assumes a flat distribution in $P_n(s)$. This flatness may not be invariant under change of variables and the consequences of this needs further investigation. It is important to stress again that $P_n(s)$ in the Bayes theorem equations is the value of the *a priori* distribution, at the true value of s. The value of the function at the unknown true value of s is known to some statistical precision $(= \frac{1}{2\lambda})$. We then use this to calculate the *a posteriori* distribution $P(s|\mathbf{c_n})$ which gives information about the true value of s to some sttistical precision. Since we use the value of the function $P_n(s)$ at the true value of s only, we may not be sensitive to the shape of $P_n(s)$. Let us also note that we do not use the *a priori* distribution explicitly for any calculations, since the information about the error of s is contained in the normalized $\mathcal{L}_{\mathcal{R}}(s)$. Combining data from different datasets may be done by multiplying likelihood ratios as shown in equation 55, without the use of $P_n(s)$.

We note in passing that the the values of λ are not large enough to span the width of the likelihood distribution. Figure 7 shows the correlation between λ and the ratio $(3\sigma/\lambda)$, where σ is the *rms* of the likelihood ratio distribution, for 500 configurations $\mathbf{c_n}$ of 1000 events per configuration. At no point does the ratio fall below unity, indicating that the likelihood curve is always broader than the step function $\theta(s|\mu)$. We may not blindly use the step function as the *a priori* distribution, centered at the maximum likelihood value and do the integral in equation 57, since it will chop off the likelihood ratio curve in the tails. The step function can only be used after we feed it with a mean value μ. The function in the integral in equation 57 is in fact a constant which equals the value of $P_n(s)$ at the true value of s. This value does *not* change as we change s in the integral in equation 57. It is possible that the true value of s is at the maximum likelihood point. It is also possible that it is at a value 3σ away, albeit with a reduced probability. The key point is that the true value of s is either at the maximum likelihood point *or* at any of the other values at which the likelihood is non-zero. They do not simultaneously have to be true. Hence we can integrate over the whole likelihood distribution with $P_n(s) = \frac{1}{2\lambda}$ without worrying about falling off the edge of the step function. As one varies s, one is testing mutually exclusive hypotheses that the value of s under consideration is the true value of s.

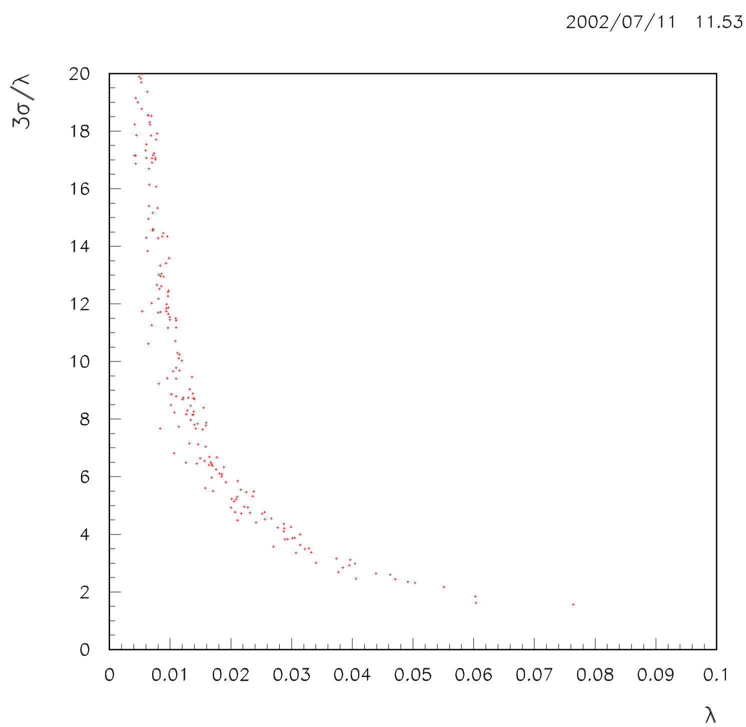

Fig. 7. Figure shows a scatter plot of λ, half the integral under the likelihood curve vs. $3\sigma/\lambda$, where σ is the width of the likelihood distribution for 500 configurations.

It is still instructive to see what happens when one supplies a distribution for the mean value of the step function. The following section deals with the self-consistency of our expressions, when one feeds in the *a posteriori* distribution $P(s|\mathbf{c_n})$ for the mean value μ of the step function.

5.3.3 The Bootstrap

If the mean value μ of the step function distribution has a probability distribution $P(\mu)$, then one can write an expression for the joint probability density of μ and s as

$$P(\mu) \times \theta(s|\mu)d\mu\, ds = \frac{P(\mu)}{2\lambda}d\mu\, ds \tag{65}$$

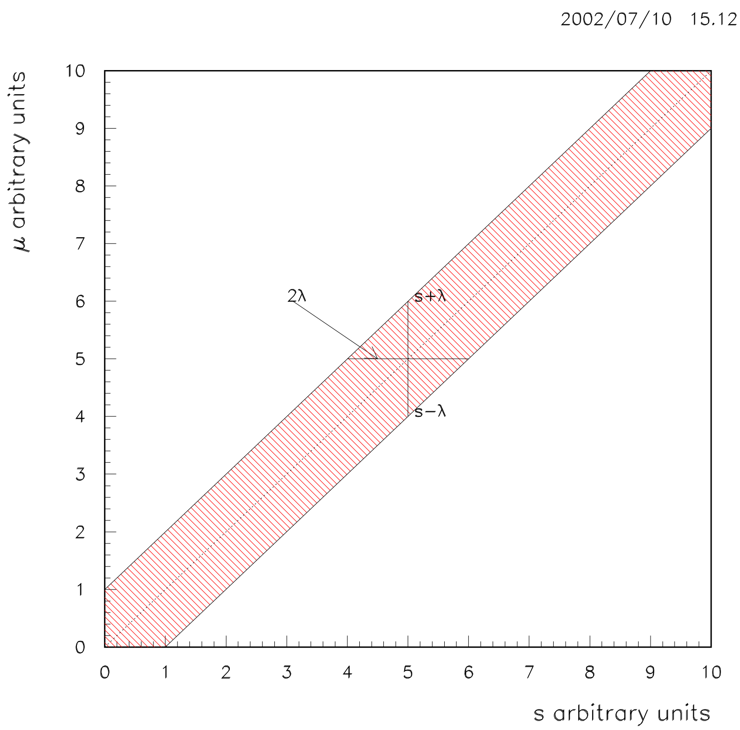

Fig. 8. The abscissa shows the variable s and the ordinate the variable μ, the mean value of the θ function distribution. The hatched region shows the area over which the probability distribution for s is non zero as a function of μ.

inside the shaded region in figure 8 and is zero outside. Integrating the above equation along the s axis first (fixed μ), followed by integration along the μ axis yields

$$\int_{\mu} P(\mu) \times d\mu \int_{s} \theta(s|\mu)ds = 1 \tag{66}$$

We can now reverse the order of integraton, doing the μ integration first, which yields

$$\int_{s} \frac{1}{2\lambda} ds \int_{s-\lambda}^{s+\lambda} P(\mu) d\mu = 1 \tag{67}$$

This can be re-written as

$$\int_{s} \frac{1}{2\lambda} g(s) ds = 1 \tag{68}$$

where

$$g(s) = \int_{s-\lambda}^{s+\lambda} P(\mu) d\mu \tag{69}$$

These equations are true for any $P(\mu)$. We need to supply a $P(\mu)$, which is the probability distribution of the mean value of the θ function. The obvious candidate for $P(\mu)$ is clearly $P(s|\mathbf{c_n})$, the *a posteriori* distribution for the true value of s. Since $\int g(s)ds = 2\lambda$, g(s) can be identified with $\mathcal{L}_{\mathcal{R}}(s)$, yielding

$$g(s) \equiv \mathcal{L}_{\mathcal{R}}(s) = \int_{s-\lambda}^{s+\lambda} P(s|\mathbf{c_n}) ds \tag{70}$$

For small λ, we can Taylor expand the above integral yielding,

$$\mathcal{L}_{\mathcal{R}}(s) \approx 2\lambda \times P(s|\mathbf{c_n}) \tag{71}$$

yielding the desired result

$$P(s|\mathbf{c_n}) \approx \frac{\mathcal{L}_{\mathcal{R}}(s)}{2\lambda} = \frac{\mathcal{L}_{\mathcal{R}}(s)}{\int \mathcal{L}_{\mathcal{R}}(s) ds} \tag{72}$$

Also, from equation 68, we can identify $P(s|\mathbf{c_n}) \equiv \frac{g(s)}{2\lambda}$, since $\frac{g(s)}{2\lambda}$ is the projection of the probability density for s in figure 8 along the s axis, computed after knowing $\mathbf{c_n}$. This again yields equation 72 and completes the bootstrap.

To summarize the arguments so far,

- We have made the ansatz that the *a priori* distribution for $P_n(s)$ is $\frac{1}{\int \mathcal{L}_{\mathcal{R}}(s)ds}$. Such an ansatz gives us a distribution $P_n(s)$ whose mean value is unknown but whose width decreases with increased statistics. Both these properties qualify it as a candidate for the *a priori* distribution. This step requires a dimensionless $\mathcal{L}_{\mathcal{R}}$ and is only possible by the use of the experimental $PDE's$ for the goodness of fit test, introduced in this paper.
- We then supply it with a probability distribution for the mean value, which is only known after we have analyzed $\mathbf{c_n}$. The candidate for the probability distribution for the mean μ is $P(s|\mathbf{c_n})$, which is the *a posteriori* distribution for the true value of s, and is the object of our quest. This is then used to calculate the probabilty distribution of s.
- This yields the expression for $P(s|\mathbf{c_n})$ of equation 72 as well as a probability density *a posteriori* for s that is consistent with the same equation.

This results in

$$P(s|\mathbf{c_n}) = \frac{\mathcal{L}_{\mathcal{R},n}(s)}{\int \mathcal{L}_{\mathcal{R},n}(s)ds} = \frac{P(\mathbf{c_n}|s)}{\int P(\mathbf{c_n}|s)ds} \tag{73}$$

for the *a posteriori* likelihood for s.

In multi-dimensional parameter space, with α being the dimension of the parameter vector s. the above equations are generalized as follows

$$P_n(s) = \frac{1}{\int \mathcal{L}_{\mathcal{R}}, n(s)ds} \equiv \frac{1}{(2\lambda)^{\alpha}} \tag{74}$$

with integrals over s being carried out over α dimensions.

6 Towards an Analytic Theory of Unbinned Likelihood Goodness of Fit

Because the likelihood ratio $\mathcal{L}_{\mathcal{R}}(s)$ is invariant under transformation $c \rightarrow c^{'}$, one can use variables $c^{'}$ such that

$$c^{'}(c) = \int_{0}^{c} P(c^{''}|s)dc^{''} \tag{75}$$

This leads to probability distributions $P(c^{'}|s)$ such that

$$P(c^{'}|s) = P(c|s) \times |\frac{dc}{dc^{'}}| = 1 \tag{76}$$

and with the limits of the variable $c^{'}$ being $0 < c^{'} < 1$. These sets of transformations in multi-dimensions is known as the hypercube transformation. The number density is constant in the hypercube which implies that we are not sensitive to systematics associated with the smoothing parameter.

The theoretical curve is a constant =1 in this scheme. The experimental PDE will also be close to being flat. The question to answer is " What is the distribution of the negative log likelihood ratio $\mathcal{NLLR}$ that results from the statistical fluctuation of the PDE in the hypercube"? We leave this question to a subsequent paper.

7 Conclusions

We have introduced a technique for estimating goodness of fit in unbinned likelihood fits by the use of probability density estimators to obtain the *a priori* likelihood distribution of the data. In addition to providing a measure of the goodness of fit in unbinned likelihood fits for the first time, this approach enables us to obtain expressions for the *a priori* likelihood distribution of the theoretical parameters and hence to derive expressions for the *a posteriori* likelihood distributions of the theoretical parameters. We have shown that the *a priori* likelihood of the theoretical parameters depends on the number n of events being employed in the problem. We have emphasized that the *a priori* likelihood is the value of the probability distribution at the true value of s and this does not change as we change s, *a posteriori*, to calculate the likelihood that s is the true value.

The approach outlined in this paper permits the rigorous calculation of errors in the fitted quantites. It makes unnecessary the practice of "guessing" the *a priori* likelihood distributions of parameters, a practice titled "Bayesianism". For the type of problems considered here, the *a priori* likelihood distributions can be computed.

The techniques detailed here are extensible to arbitrary dimensions, even though we have used a one-dimensional problem for illustrative purposes. In the process of using probability density estimators, we have developed an algorithm for iteratively improving the smoothing parameter as a function of local number density.

8 Acknowledgements

The author would like to thank British Rail for a particularly peaceful and appropriately long journey between Durham and London during which the main ideas in this paper were worked out.

9 References

References

[1] R. A. Fisher,"On the mathematical foundations of theoretical statistics", *Philos. Trans. R. Soc. London Ser. A* **222**, 309-368(1922);

R. A. Fisher,"Theory of statistical estimation", *Proc. Cambridge Philos. Soc.* **22**, 700-725 (1925).

[2] "An essay towards solving a problem in the doctrine of chances", Rev. Thomas Bayes, Biometrika,**45** 293-315 (Reprint of 1763) (1958).

[3] B. Knuteson *et al* , FERMILAB-PUB-01-226-E, Physics/0108002, submitted to Comp. Phys. Comm.

L. Holmström, S. R. Sain, and H. E. Miettinen, Comp. Phys. Comm. 88:195 (1995).

[4] D. Scott. *Multivariate Density Estimation.* John Wiley & Sons, 1992.

M. Wand and M. Jones, *Kernel Smoothing.* Chapman & Hall, 1995.

[5] S.Towers, *Kernel Probability Density Estimation Methods*, paper submitted to the Conference on Advanced Statistical Techniques in Particle Physics, Durham, England, March 2002, *http://www.ippp.dur.ac.uk/statistics/papers/towers_pde.ps.gz*

[6] After a modest amount of experimentation with smoothing factors, we settled on the factor of 0.5 as being adequate to reproduce the exponential during this first phase of analysis. Later the smoothing function was introduced to improve the goodness of fit capability of the technique.

[7] Empirically we found that a power of the order of -0.6 is needed to provide sufficently large smoothing factors for large values of time. We can in principle

optimize this smoothing function further, but have not done so.

[8] MINUIT is a general minimization program written by F. James.