Chapter 8

Fitting

8.1 Optimisation

The process of ***fitting*** a sample of data $D(\underline{x})$ containing N entries involves iterative comparison of data and a theoretical model $\mathcal{P}(\underline{x}, \underline{p})$ assumed to describe that data[1]. The $\underline{x}$ are discriminating variables to be used in the fit taking the value $\underline{x}_i$ for the i^{th} event, and $\underline{p}$ are the parameters of the theoretical model. The parameters are allowed to vary in the fit to data, i.e. each iteration of the optimisation process uses a different value for $\underline{p}$, in order improve the compatibility of data and model. This process requires the definition of a test statistic T that is used to quantify how well the model and data agree and to then vary $\underline{p}$ in such a way as to try and obtain an improved model description of the data. The test statistics described in this section are χ^2 and likelihood based. In both cases we numerically minimise the test statistic summed over the data. So in order to perform a fit to data we perform an optimisation process in the parameter space $\underline{p}$ which involves minimising the sum

$$S = \sum_{i=1}^{N} T[D(\underline{x}_i), \mathcal{P}(\underline{x}, \underline{p})]. \tag{8.1.1}$$

In order to converge on a solution one has to start with an initial estimate of the parameter set to evaluate S. Having done this one then determines a new estimates of $\underline{p}$ following a pre-defined rule, at each step evaluating the corresponding S. After making an initial set of estimates of the parameter set, it is normally possible for the algorithm to determine in which direction a more optimal set lies. Having done this, the algorithm will perform another search starting from a point $\underline{p}'$ that is closer to the assumed minimum than the previous one. This process is repeated until such time as the optimisation algorithm has a sufficiently small step size given by

$$\delta = \left| \underline{p} - \underline{p}' \right|. \tag{8.1.2}$$

When δ is smaller than some minimum step size or distance η in the parameter space, the optimisation is said to have converged on a minimum value $\underline{p}_{\min}$. Having found a minimum $\underline{p}_{\min}$ the final step is for the algorithm to determine the corresponding uncertainty $\delta\underline{p}_{\min}$, which depends on the test statistic that is being minimised.

General issues with numerical optimisation procedures is that they do not always distinguish between local and global minima, and they do not always converge to a minimum. If a minimum is found, such that $\delta < \eta$,

[1] Depending on the method we use, we are able to fit the data on an event-by-event basis, or by binning the data in finite intervals.

the algorithm considers this point in space to be the minimum. Further tests must be made in order to determine if one has found a local or global minimum. The procedure to validate that a minimum is global involves scanning the start parameters for the fit, and repeating the minimisation. It is not always practical to perform such a parameter scan, however there are certain circumstances where it is absolutely necessary, for example fits with large numbers of parameters where there are large correlations between parameters.

One thing to be borne in mind is that the number of computational steps required in order to converge to a local or global minimum will depend on how close the initial values of $\underline{p}$ are to the values corresponding to the minimum and on the step size used to iterate $\underline{p}$. It is possible that a fit may not converge to a minimum if the values of $\underline{p}$ are far from the values at a minimum. There are many ways to minimise a quantity, each have their own pitfalls, and only two examples are described in the following to illustrate the process of optimisation. In practice one normally uses a more sophisticated approach to determine the optimal set (or fitted) parameters $\underline{p}$.

8.1.1 Gradient descent method

Consider an m dimensional parameter set that is to be fit to some model given by $\mathcal{P}$ for some data sample. The sum S used to compare the data to the model is given by Eq. (8.1.1). Starting from a point in the parameter space $\underline{p}$, one can numerically compute an estimate for the gradient

$$g(\underline{p}) = \frac{\partial S}{\partial \underline{p}}, \tag{8.1.3}$$

from one (or more) neighbouring point(s) $\underline{p} + \Delta\underline{p}$. In practice we can only determine an approximation of the gradient

$$g(\underline{p}) \simeq \frac{\Delta S}{\Delta \underline{p}}. \tag{8.1.4}$$

Having determined the gradient at the point $\underline{p}_j$, one can use this to estimate a new point $\underline{p}_{j+1}$ some small distance $\underline{\epsilon}$ from $\underline{p}_j$

$$\underline{p}_{j+1} = \underline{p}_j + \underline{\epsilon} \cdot g(\underline{p}). \tag{8.1.5}$$

In general, one expects $\underline{p}_{j+1}$ to be a better estimate of the true minimum than $\underline{p}_j$. One can continue iterating on this process, each time estimating a new parameter set until such time as the estimated gradient is close enough to zero to be considered the minimum. In general the condition for having determined the minimum is

$$\frac{\Delta S}{\Delta \underline{p}} = 0, \tag{8.1.6}$$

and the value of S either side of the stationary point found in the parameter space can be used to distinguish between possible maxima, minima, and points of inflection. As this search is numerical, sometimes the new estimate $\underline{p}_{j+1}$ can have a value that is further from the minimum than $\underline{p}_j$. This can occur as the step size $\underline{\epsilon}$ is a pre-determined parameter for the search and may simply be too large. If subsequent iterations fail to re-converge on a minimum, this optimisation procedure can fail to converge. For complicated models, failure to converge may not be uncommon, and as such one should take care to choose a reasonable starting parameter set $\underline{p}_0$.

8.1.2 Parameter scan

In certain situations it may not be possible to optimise or fit the values of all parameters that a model depends upon, or one may want to understand the change in S as a function of a particular parameter without resorting to an optimisation algorithm. In such scenarios one can perform a parameter scan. This involves stepping through the values of a parameter of interest p from some minimum value through to some maximum value. The minimum and maximum are chosen such that the best fit value that one is attempting to determine is bound by these limits. At each point between p_{min} and p_{max} the sum S is computed. Hence one can plot S as a function of p in order to determine the optimal value of the parameter via such a *scan*. If the model depends on more than one parameter then the ancillary parameters should be optimised at each value of p used in the scan. An example of this approach is given in section 8.2.1.

8.2 The least squares or χ^2 fit

In order to perform a χ^2 fit, one typically bins data such that there are at least 5 to 10 events that contribute to each bin. As a result the data are generally not binned in samples that are of equal size in the discriminating variable space. The test statistic used for this type of fit is a χ^2 constructed between the data D_i and theoretical model describing the data $\mathcal{P}_i$ i.e.

$$\chi^2 = \sum_{i=1}^{N} \left(\frac{D_i - \mathcal{P}_i}{\sigma(D_i)} \right)^2, \tag{8.2.1}$$

[see Eq. (4.5.2)] where the sum is over the number of bins. The parameter ($\underline{p}$) and discriminating variable ($\underline{x}$) dependence have been suppressed here, and are implied. The quantity $\sigma(D_i)$ is the uncertainty on the datum D_i, which in the case of an event yield is given by the corresponding Poisson error on that yield. As by definition the χ^2 distribution is normalised by the uncertainty from data, the 1σ error on a result is given when the χ^2 changes by one unit from the minimum value $\chi^2_{\rm min}$. Once you have a data sample to fit to the only remaining issue is the choice of model $\mathcal{P}$. Some commonly used PDFs are discussed in appendix B, in addition to those already encountered earlier in the book. The test statistic in Eq. (8.2.1) is sometimes referred to as a least squares statistic. In this book the terminology χ^2 fit is generally used in the context of an arbitrary model $\mathcal{P}$ describing the data, where a numerical minimisation is performed to obtain the optimal set of fit parameters $\underline{p}$, and least squares optimisation is used in the context of problems that can be solved by analytic means (see section 8.3). This distinction is artificial and introduced here to distinguish between the two ways of solving a problem using this type of test statistic.

8.2.1 Example: determining the average of a set of measurements

Consider the situation where one has several measurements of the value of some quantity S as shown in Table 8.1. This quantity is the measure of difference between matter and anti-matter decaying from an initial state called a B meson into a final state involving so-called charmonium particles and a strange particle (a kaon). The data are taken from a journal article written by collaborators working on a high-energy physics experiment called *BABAR* (Aubert *et al.*, 2009). The parameter S has to be zero for matter and antimatter to behave in the same way for these measurements.

Based on this information *what is the average value of S*? While it is possible to compute a weighted average using the formalism outlined in section 5.4.1, it is also valid to consider using a fit or scan to determine the average value of a set of measurements. The advantage of using a scan-based approach to compute the average is the retention of more information concerning the observable we are trying to measure. Instead of a single number to represent the uncertainty, one has a curve that can be used to determine confidence levels of arbitrary significance.

Table 8.1: Individual measurements of a parameter S as described in the text. The uncertainties σ_S considered are statistical only.

S	0.662	0.625	0.897	0.614	0.925	0.694	0.601
σ_S	0.039	0.091	0.100	0.160	0.160	0.061	0.239

Each column in the table corresponds to an event i that contributes to a χ^2 sum. As we are trying to determine the most optimal value of S, the model is given by the assumed value of that parameter, in other words we perform a parameter scan of S, from S_{min} to S_{max}, and for each point in this range we can compute the χ^2 sum

$$\chi^2 = \sum_{i=1}^{7} \left(\frac{S_i - S}{\sigma_{S_i}} \right)^2 , \tag{8.2.2}$$

where S_i and σ_{S_i} are the central values and uncertainties for the i^{th} event. This is shown in Figure 8.1. The minimum value of the χ^2 is at $S = 0.690$, and the corresponding error (how far one moves away from the minimum value of S in order to obtain a change in χ^2 of one from the minimum value) is 0.028; thus the average value of the data in the table using this method is $S = 0.690 \pm 0.028$. For comparison, the average value obtained in the original reference is $S = 0.687 \pm 0.028$. Note that the method used for the average computed here is not as sophisticated as that in Aubert *et al.* (2009), which explains the small difference obtained between these two results. The results of the method outlined here and the one used in the original reference give essentially the same average given the precision of this set of measurements.

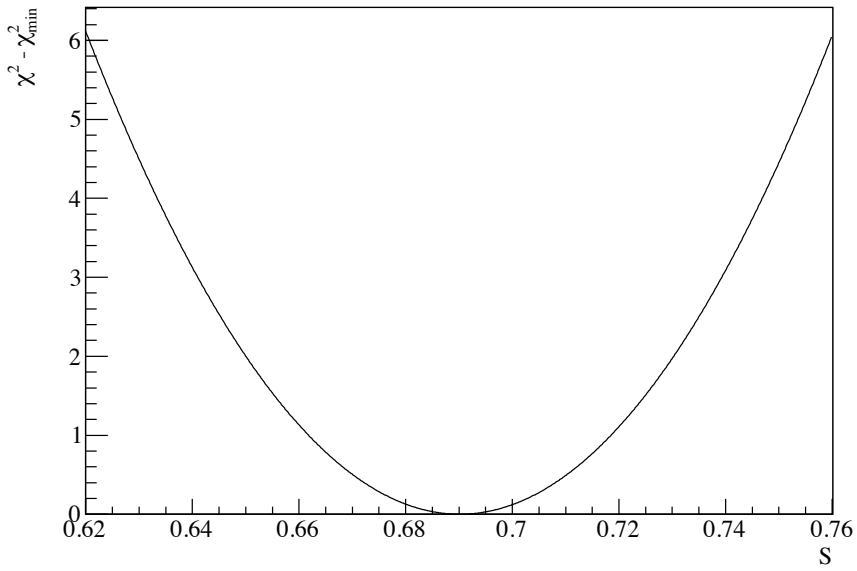

Figure 8.1: The χ^2 parameter scan in terms of S for the data given in Table 8.1.

It turns out that the parameter S is a function of a more fundamental quantity, an angle β. Given that $S = \sin(2\beta)$ one can repeat the parameter scan in terms of the angle β by replacing S in Eq. (8.2.2) by $\sin(2\beta)$. For each assumed value of β in $[0, 360]^\circ$ one can compute $\sin(2\beta)$ and hence determine the χ^2 sum. The corresponding average value obtained for β in the first quadrant is $(21.8 \pm 1.1)^\circ$.

If instead of performing a scan in S or β, we chose to use a minimisation algorithm, we would have arrived at the same results as obtained from Figure 8.1. However in performing just a fit to the data, we would have obtained only a central value, and uncertainty estimate for the average value of the observable. By performing both the fit, and the parameter scan we have an estimate of the optimal value of the observable, and a visual representation of the behaviour of the test statistic in the vicinity of this value. This enables us to verify that the test statistic is smoothly varying (i.e. well behaved) in the vicinity of the minimum.

Having obtained the optimal result from a minimisation process, it is possible to compute the probability of obtaining that result given the number of degrees of freedom, using $P(\chi^2, \nu)$ from Eq. (4.5.1). This is an indicator of how well the model agrees with data and is often referred to as the goodness of fit, or GOF (see section 4.5), and it is something that should be considered when validating a result. For example in the case described here we have obtained the average result $S = 0.690 \pm 0.028$. The sum of χ^2 deviations from the average for this data (often abbreviated as the χ^2) is $\chi^2 = 7.83$. There are $n - 1$ degrees of freedom, as there are $n = 7$ data used in the evaluation of the χ^2, and the total sum is constrained by the number of data. Thus there are 6 degrees of freedom ($\nu = 6$). The χ^2 probability for this situation is given by $P(\chi^2, \nu) = 0.25$, which means that the result has a reasonable outcome.

If we had obtained a result where $\chi^2 \sim 0$, then the uncertainties on each of the individual data would have been overestimated. This would indicate that one has probably inflated or over estimated the errors in order to improve the level of agreement of the individual measurements, or alternatively that the data are highly correlated, and can not be combined without accounting for the correlation. Similarly had we obtained $\chi^2/\nu >> 1$, then we would conclude that the data are not in good agreement, as again one would find that $P(\chi^2, \nu) \simeq 0$. In either of these extremes we have to try to understand if it is meaningful to combine the results (this is best done before attempting any such combination where possible). If there is a particular piece (or several pieces) of data that dominates the χ^2 then studying that input in more detail would be in order. If on inspection it turned out that the data for a given event is suspect, for example the measurement was wrong, then that data point may be ignored. If however that experiment appeared to be reasonable, then it is not appropriate to discard the data as the result may simply be a statistical out lier.

8.3 Linear least-squares fit

The procedure adopted in section 8.2 can be used to study general situations where the theoretical model described by $\mathcal{P}$ is arbitrary as illustrated through the previous example of determining average values of S and β. Often one can take an analytical approach to solve a given problem. One situation that often arises is the case of comparing data to the model $\mathcal{P}_i = ax_i + b$, where the uncertainty on each point $\sigma(y_i)$ is some constant value denoted by σ. In this case we can write Eq. (8.2.1) as

$$\chi^2 = \sum_{i=1}^{N} \left(\frac{y_i - ax_i - b}{\sigma(y_i)} \right)^2 = \frac{1}{\sigma^2} \sum_{i=1}^{N} (y_i - ax_i - b)^2 \,, \tag{8.3.1}$$

where we have replaced D_i by y_i and made the appropriate substitution for $\mathcal{P}_i$. The task at hand now is to minimise χ^2 with respect to both a and b in order to determine the optimal values of the slope and offset of our model. In order to do this we differentiate χ^2 with respect to a and b, and simultaneously solve for values that correspond to the point where both derivatives are zero[2]. In the simplified case where the uncertainties on the y_i are all some constant value, we can remove the constant $1/\sigma^2$ from the problem. If we consider

[2]In general one can use the second derivative or numerical means to establish the nature of the turning point, and ensure one has located a minimum.

the derivative with respect to a first, this is just

$$\sigma^2\frac{\partial\chi^2}{\partial a} = \sum_{i=1}^{N}\frac{\partial}{\partial a}\left(y_i - ax_i - b\right)^2, \tag{8.3.2}$$

$$= \sum_{i=1}^{N} -2x_i(y_i - ax_i - b), \tag{8.3.3}$$

$$= -2\sum_{i=1}^{N} x_iy_i - ax_i^2 - bx_i, \tag{8.3.4}$$

$$= -2N(\overline{xy} - a\overline{x^2} - b\overline{x}). \tag{8.3.5}$$

Similarly for the derivative of χ^2 with respect to b one obtains

$$\sigma^2\frac{\partial\chi^2}{\partial b} = -2N(\overline{y} - a\overline{x} - b). \tag{8.3.6}$$

As the optimal solution exists for

$$\frac{\partial\chi^2}{\partial a} = 0, \text{ and } \frac{\partial\chi^2}{\partial b} = 0, \tag{8.3.7}$$

we need to simultaneously solve (ignoring constant multipliers)

$$\overline{xy} - a\overline{x^2} - b\overline{x} = 0, \tag{8.3.8}$$

$$\overline{y} - a\overline{x} - b = 0, \tag{8.3.9}$$

for a and b. The results of this are

$$a = \frac{\overline{xy} - b\overline{x}}{\overline{x^2}} = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - \overline{x}^2}, \tag{8.3.10}$$

$$b = \overline{y} - a\overline{x}. \tag{8.3.11}$$

Hence, for the situation where we want to determine the coefficients of a straight line fit to data, where the data y_i have equal uncertainties, and the abscissa values x_i are precisely known, we can analytically solve for the slope and intercept parameters without having to perform a numerical optimisation.

Using the combination of errors procedure outlined in chapter 5 on Eqns (8.3.10) and (8.3.11) it can be shown, for example see Barlow (1989), that

$$\sigma^2(a) = \frac{\sigma^2}{N(\overline{x^2} - \overline{x}^2)} \tag{8.3.12}$$

$$\sigma^2(b) = \frac{\sigma^2\overline{x^2}}{N(\overline{x^2} - \overline{x}^2)}. \tag{8.3.13}$$

In general the least squares method can be written in matrix form as

$$\chi^2 = \Delta^T V^{-1}\Delta, \tag{8.3.14}$$

where V is the covariance matrix and Δ is a column matrix of difference terms of the form $x_i - f(x_i, \underline{p})$, and $\underline{p}$ are parameters of the model f. From this general form it is possible to derive the weighted averaging procedure introduced in chapter 5. The use of least squares regression is discussed in more detail for example in the books by Barlow (1989); Cowan (1998); Davidson (2003); James (2007).

8.4 Maximum-likelihood fit

It is possible to define a likelihood function $\mathcal{L}$ that uses the PDFs $\mathcal{P}$ to describe the distribution of discriminating variables $\underline{x}$ in data. As the functions $\mathcal{P}$ are normalised so that the total probability of an event is unity, we can write the likelihood for an event as

$$\mathcal{L}_i = \mathcal{P}(\underline{x}_i). \tag{8.4.1}$$

The quantity $\mathcal{P}(\underline{x}_i)$ is the probability assigned by the model for the i^{th} event. If the event $\underline{x}_i$ is certainly signal, then the part of $\mathcal{P}(\underline{x}_i)$ corresponding to signal will be one. Similarly if the event is definitely not signal then the part of $\mathcal{P}(\underline{x}_i)$ corresponding to signal will be zero. One can make similar statements for background, or indeed any other type or class of event described by the model $\mathcal{P}$. Usually the probability assigned to an event for a given component j is between these two extreme values, so in general $0 \leq \mathcal{P}_j(\underline{x}_i) \leq 1$. In order to express these possibilities mathematically we need to consider the situation when there are distinct components in the model. For example, if there are m components in a model, then the likelihood for an event i is given by a sum over these m components

$$\mathcal{L}_i = \sum_{j=1}^{m} f_j \mathcal{P}_j(\underline{x}_j), \tag{8.4.2}$$

where f_j are the fractions of the different components. For a single event the f_j are interpreted as the probability that an event is of type j. Thus in order to conserve probability we require[3]

$$\sum_{j=1}^{m} f_j = 1. \tag{8.4.3}$$

So far we have only considered a single event that may be one of m different possible types. In reality there is a limited amount of information that can be gleamed from a single event and we are usually faced with interpreting results from data samples of many events. As each event is independent, the likelihood for a data set containing N events is the product of the likelihoods for the individual events

$$\mathcal{L} = \prod_{i=1}^{N} \mathcal{L}_i, \tag{8.4.4}$$

$$= \prod_{i=1}^{N} \sum_{j=1}^{m} f_j \mathcal{P}_j(\underline{x}_i). \tag{8.4.5}$$

It can be troublesome to numerically compute $\mathcal{L}$ for large samples of data. The test statistic that is usually optimised is

$$-\ln \mathcal{L} = -\ln \prod_{i=1}^{N} \mathcal{L}_i, \tag{8.4.6}$$

$$= -\ln \prod_{i=1}^{N} \sum_{j=1}^{m} f_j \mathcal{P}_j(\underline{x}_i), \tag{8.4.7}$$

$$= -\sum_{i=1}^{N} \ln \sum_{j=1}^{m} f_j \mathcal{P}_j(\underline{x}_i), \tag{8.4.8}$$

which is easier to compute numerically and follows the form of Eq. (8.1.1). All that remains is to define the fit model by choosing the $\mathcal{P}_j$ distributions. A number of PDFs that can be used to construct models are discussed in chapter 4 and appendix B.

[3] It is possible to use the normalisation constraint of Eq. (8.4.3) to reduce the total number of parameters by one, resulting in the final component fraction $f_m = 1 - \sum_{i=1}^{m-1} f_j$.

8.4.1 Extended maximum-likelihood fits

If one is interested in determining event yields n_j of a set of categories, instead of fractions, then the Poisson nature of those yields needs to be taken into account. In such circumstances the likelihood function given in Eq. (8.4.5) is modified by a Poisson term that depends on the total fitted event yield $n' = \sum n_j$, where n_j is the number of fitted events of type j, to give (Barlow, 1990)

$$\mathcal{L} = \frac{e^{-n'}}{N!} \prod_{i=1}^{N} \sum_{j=1}^{m} n_j \mathcal{P}_j(\underline{x}_i), \tag{8.4.9}$$

The total number of fitted events n' does not have to exactly match the total number of events. The fit has the ability to converge on some minimum value close to, but not necessarily satisfying $n' = N$. More details on the usage of this form of likelihood fit can be found in Refs (Barlow, 1990; Cowan, 1998). If the event yields n_j are observables of primary interest then an extended maximum-likelihood fit provides a more convenient approach than the original form.

8.4.2 Interpreting the result of a likelihood fit

In general, as with the case of least squares, one can analytically solve for the optimal value of a parameter for a given problem by requiring

$$\frac{\partial \mathcal{L}}{\partial p} = 0, \tag{8.4.10}$$

and subsequently using the second derivative or numerical means to ensure that the stationary point is a maximum. The maximum will provide the most likely value of p which is denoted here by p_0. In general the variance on p is given by the integral equation

$$(\Delta p)^2 = \frac{\int (p - p_0)^2 \mathcal{L} dp}{\int \mathcal{L} dp}, \tag{8.4.11}$$

which follows directly from Eq. (4.1.2).

In practice it is often cumbersome to analytically compute the optimal value and variance for a parameter for a complicated fit model. In such a scenario one must resort to numerical means of evaluating the best fit value. On computing the value of $-\ln \mathcal{L}$ for a data set with a given assumed parameter set we are able to compute a number that is related to the probability of the agreement of the data to the assumed model with the assumed parameters. After completing this initial step, one performs an optimisation of $-\ln \mathcal{L}$ as described in section 8.1. As a natural part of the optimisation process the fit will converge on some optimal result denoted by $-\ln \mathcal{L}_0$, that corresponds to a minimum minimum value of $-\ln \mathcal{L}$. This is the most probable result that our algorithm has been able to identify. This is a single point in parameter space, and tells us nothing about the uncertainty on that point in space in terms of the parameters. Furthermore, if there is more than one parameter in the parameter set, then one has to worry about how correlated those parameters might be with each other, and that in turn can complicate the determination of the uncertainty on a parameter. For such a scenario the problem has to be solved in a multidimensional space.

If we consider an observable p, then we expect the uncertainty on p to be distributed according to a Gaussian for large data samples. Hence, for large samples we expect

$$\mathcal{L}(p, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(p-\mu)^2/2\sigma^2}, \tag{8.4.12}$$

with some mean value μ and standard deviation σ providing an estimate of the parameter p and its uncertainty. The negative-log-likelihood is given by

$$-\ln\mathcal{L}(p,\mu,\sigma) = -\ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right) + \frac{(p-\mu)^2}{2\sigma^2}, \tag{8.4.13}$$

where the first term is a constant and the second term is the equation of a parabola. When $p-\mu$ corresponds to a 1σ Gaussian uncertainty the second term in Eq. (8.4.13) reduces to a factor of one half, and when $p=\mu$, the second term vanishes. So a change of one half in $-\ln\mathcal{L}$ from the minimum value obtained corresponds to a change in p of 1σ. In this limit of large statistics uncertainties are Gaussian by virtue of Eq. (8.4.12). This result encapsulates Wilks' theorem (Wilks, 1937) which states in the limit of large statistics the distribution of $-2\ln\mathcal{L}$ equates to a χ^2 distribution with δ degrees of freedom, where δ is the difference in the number of parameters for models used in computing some likelihood ratio. That theorem is only valid for large statistics as a term of order $1/\sqrt{n}$ is neglected in the approximation made by Wilks.

In the limit of small statistics, or where the $-\ln\mathcal{L}$ distribution is not parabolic about the minimum value it is not appropriate to attribute a change of one half to represent a $\pm1\sigma$ boundary and Monte Carlo simulations should be used to determine the interval corresponding to the uncertainty on a parameter. In practice if the $-\ln\mathcal{L}$ distribution is almost parabolic we assume that it is valid to invoke the theorem of Wilks.

8.4.3 Adding constraints

If one is performing a fit to data given some model θ, where external information is available that may help constrain one or more parameters in the model, then it is possible to include a penalty term to the test statistic being optimised. For example in the case of a χ^2 fit, where some parameter p has been previously measured as $p_{meas}\pm\sigma_p$, one could simply add a term to the χ^2 being minimised in the fit. This additional term would take the form

$$\chi^2_{penalty} = \frac{(p-p_{meas})^2}{\sigma_p^2}. \tag{8.4.14}$$

The quantity $\chi^2_{penalty}$ would be computed for each iteration of the minimisation process and added to the usual χ^2 sum in order to incorporate external knowledge. Additional information can be included in a likelihood fit by extending the model to simultaneously fit the external information in an analogous way. It is also possible to implement constraint equations in a fit using Lagrange multipliers, for example see Edwards (1992); Silvia and Skilling (2008).

8.4.4 Fit bias and checking for self consistency

The range of fit models that one may want to apply to data varies from simple, where there are a few parameters to determine from data, to extremely complicated, where there may be tens, hundreds, or even more parameters to determine. In all cases the optimisation process used is not guaranteed to work properly, and one should take care to validate that a fit result obtained is sensible.

If one fits a sample of data, and obtains some parameter $p_{fitted}=p_0\pm\sigma_{p_0}$, it is reasonable to ask if p_{fitted} is a good representation of the true value of the parameter p. In general the test statistic used in the optimisation has an intrinsic bias, dependent on the sample size. In order to test for fit bias one can generate, and fit to, an ensemble of simulated data. The distributions of fitted values p_0 and σ_{p_0} obtained can be used to evaluate if the fit being performed is biased or not.

If the simulation is an accurate representation of the measurement being performed, then one will obtain a set of fitted parameters p_{fitted} for each simulated measurement in the ensemble. These should be distributed

with a mean value corresponding to that used to generate the ensemble, p_{input}, and a spread representative of the uncertainty obtained from the fit. In general one can plot the so-called '*pull*' distribution

$$\mathcal{P} = \frac{p_{fitted} - p_{input}}{\sigma_p}, \tag{8.4.15}$$

which should be centred on zero for an unbiased measurement, with a width of one if the fit correctly extracts the uncertainty on p from data. A common reason why one might obtain a width of the distribution in Eq. (8.4.15) differing from one is if the simulated experiments do not allow for Poisson fluctuations that are inherent in fitting for event yields.

An additional consistency check that can be made for a maximum likelihood fit is to verify that the value of $-\ln\mathcal{L}$ obtained from fitting data is consistent with the distribution of $-\ln\mathcal{L}$ obtained from an ensemble of simulated measurements. If there is a disagreement, this could indicate that there is a problem with either the fit to the data, or with the simulation used. In general the absolute value of $-\ln\mathcal{L}$ is dominated by the number of events used in the fit to data and corresponds, on average, to a sum over all events of the first term in Eq. (8.4.13) for fits to large samples of data.

8.5 Combination of results

The most transparent way of combining results from several different measurements of a single observable is to construct a fit model that encompasses all necessary parameters to describe all of the data used to extract information on the observable. Having done this, the data are fit using that model. This procedure can be used for situations that vary from different measurements having many parameters in common between them, to ones where only the observable of interest is common. In reality this approach can be complicated, and the fit validations required to understand the fit bias and performance may be impractical. If this is the case, the alternative discussed below may be useful.

Consider the situation when there are several different determinations of some parameter, each with a given likelihood or χ^2 distribution as a function of that parameter. Such a situation is not uncommon with complicated experiments as there can be more than one way to measure and observable. It is possible to combine the likelihood or χ^2 distributions of the different measurements directly. For two χ^2 distributions, the total χ^2, $\chi^2_{\rm TOT}$ is the sum of the individual contributions as:

$$\chi^2_{\rm TOT} = \chi^2_1 + \chi^2_2, \tag{8.5.1}$$

which follows from Eq. (8.2.1). If the estimates of the parameter are correlated with each other then one needs to resort to using the general form of the χ^2 given by Eq. 8.3.14. With regard to combination of two likelihoods $\mathcal{L}_1$ and $\mathcal{L}_2$, as each is a representation of the probability for something to happen, the combined likelihood is just the product $\mathcal{L}_1\mathcal{L}_2$. It follows that, as we usually optimise $-\ln\mathcal{L}$, the optimal value for some parameter derived from a combination of two likelihood functions is given by $-\ln\mathcal{L}_1 - \ln\mathcal{L}_2$. The minimum value of this combination corresponds to the best fit value of the parameter(s) under study, and the $\pm1\sigma$ uncertainties can be obtained as discussed in Section 8.4.2. Figure 8.2 shows the result of combining two likelihood distributions in this way.

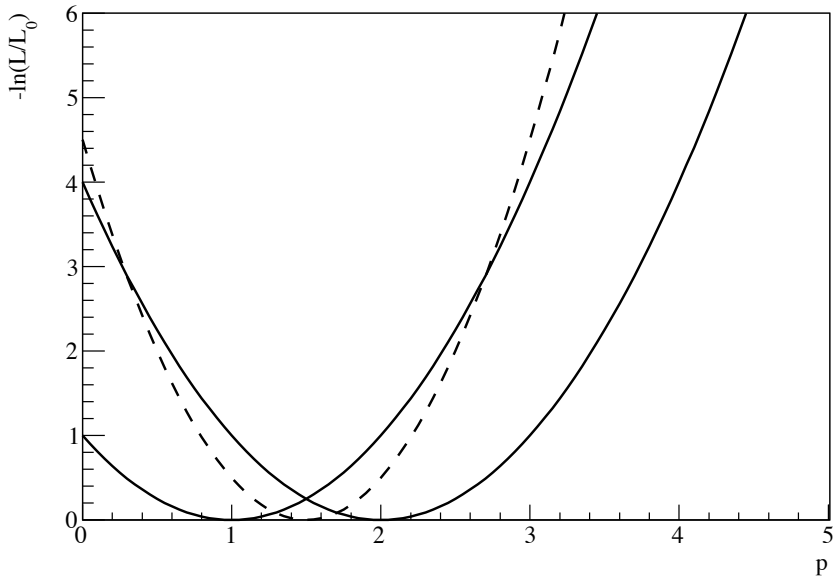

Figure 8.2: The $-\ln(\mathcal{L}/\mathcal{L}_0)$ curves obtained from (solid) two different determinations of a parameter p, with the (dashed) combined distribution shown.