C. Blocker
April 9, 2002
Version 3.0

Maximum Likelihood Primer

Abstract

This note offers a brief introduction to maximum likelihood methods. It describes the basics of parameter estimation using the likelihood function and examines several potential pitfalls that both the neophyte and the experienced user should avoid. The Run I $\sin 2\beta$ analysis is used as an example of how to construct a complex and realistic likelihood function for an actual High Energy Physics analysis.

1 Introduction

We often wish to infer the value of a constant of nature from a set of measurements. For example, we may measure the invariant mass of the decay products of a particle, such as $B^0 \to J/\psi K_s$, and wish to extract a measurement of the B^0 mass. Or perhaps we determine an efficiency as a function of transverse momentum and wish to determine the parameters of a functional form that well describes that dependence.

Determining an estimate of a parameter from measured data is known as parameter estimation. There are many methods of parameter estimation, such as χ^2 fits, binned likelihood fits, unbinned likelihood fits, average calculation, and linear regression. In general, an unbinned likelihood fit (also known as the maximum likelihood method) is the most powerful (that is, does the "best" parameter estimation, as explained in section 3) for a given set of data. This note discusses the basics of the maximum likelihood method. It is meant as a primer for those just learning about such things but hopefully includes useful information for experienced data fitters. For a more technical discussion of properties of likelihood functions and fits, see reference [1]

2 Maximum Likelihood Basics

To apply the maximum likelihood method, a likelihood function must first be constructed. Let $p(X|\alpha)$ be the the probability of getting a measurement X on a given event. Note that X may represent more than one observable. For example, it could be the mass and proper lifetime of a $B \to J/\psi K_s$ in a single event. This probability is assumed to depend on a set of parameters labeled as α, which could be one parameter or more than one.

The likelihood function is the product over N measurements of the probability p, that is,

$$L = \prod_{i=1}^{N} p(X_i|\alpha). \tag{1}$$

The log–likelihood function is the natural logarithm of the likelihood function, that is,

$$\ln L = \sum_{i=1}^{N} \ln(p_i). \tag{2}$$

We often want to infer the true value of the unknown parameter(s) α from the measurements X_i. The maximum likelihood estimator (MLE or mle) for the parameter α is the value that maximizes the likelihood function (hence the name of the method). Since the natural logarithm function is monotonic, the value of α that maximizes L also maximizes $\ln L$.

If there is one parameter α and it is a continuous variable, then the value of α that satisfies

$$\frac{d(\ln L)}{d\alpha} = 0 \tag{3}$$

is the MLE. Solutions to this equation could also be minima or inflection points, which should be checked, but this is usually not a problem in practice. If there is more than one parameter (call them α_j), then we have to solve the following set of simultaneous equations

$$\frac{\partial(\ln L)}{\partial\alpha_j} = 0. \tag{4}$$

In most realistic problems, analytic solutions to these equations are difficult or impossible to obtain. In that case, numeric methods are used (see Section 10 for a brief discussion).

In order to keep things straight, we will use the following notation: α for a parameter, $\hat{\alpha}$ for its MLE, and α_0 for its true value.

As a concrete example, consider the measurement of the mass of $B \to J/\psi K_s$ candidates in the absence of background. Assume the mass resolution is Gaussian and given by

$$p(M|M_B,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(M-M_B)^2}{2\sigma^2}}, \tag{5}$$

where the actual B^0 mass M_B and the mass resolution σ are the parameters. The likelihood and log–likelihood functions are

$$\begin{aligned} L &= \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(M_i-M_B)^2}{2\sigma^2}} && (6) \\ \ln L &= -N\ln(\sqrt{2\pi\sigma^2}) - \frac{1}{2}\sum_i \frac{(M_i-M_B)^2}{\sigma^2}, && (7) \end{aligned}$$

where N is the number of measurements. The parameter values that maximize the log–likelihood are

$$\hat{M}_B = \frac{1}{N}\sum_i M_i \tag{8}$$

$$\hat{\sigma}^2 = \frac{\sum_i (M_i - \hat{M}_B)^2}{N}. \tag{9}$$

These formulae probably look familiar as estimators for the mean and variance of Gaussian distributed measurements. You may prefer a factor of $N-1$ instead of N in the variance formula – this is discussed in the next section.

3 Bias, Consistency, and Efficiency

It may seem reasonable that a maximum likelihood estimator is a good estimator of a parameter because it is the one that maximizes the probability of getting the observed measurements. However, we want to be sure mathematically whether an MLE is a "good" estimator, or even whether it is the "best" estimator, for the parameter. To define "good" and "best", we define some mathematical properties that we wish our estimators to have.

To do this, you must first realize that an estimator is itself a random variable. If we repeat an experiment, we will get a different set of measured values. Since an estimator is a function of the measured values, we will get a different estimator. Thus, when averaged over many experiments (that is, sets of measurements), an estimator has an average value and a variance (and all other statistical measures).

Statisticians define an unbiased estimator as one whose expectation value is equal to the true value of a parameter. This is obviously something we would like to require of our estimators. Unfortunately, maximum likelihood estimators do not always have this property. Sometimes they do, such as for the mean of a Gaussian in equation 8, as can be seen from

$$\overline{\hat{M}} = \overline{\frac{1}{N}\sum_i M_i} = \frac{1}{N}\sum_i \overline{M_i} = \frac{1}{N}\sum_i M_0 = M_0, \tag{10}$$

where the overline indicates average or expectation value over many experiments. The variance of a Gaussian in equation 9 provides an example of a biased MLE, since a careful calculation shows that

$$\overline{\hat{\sigma}^2} = \frac{N}{N-1}\sigma_0^2, \tag{11}$$

that is, the average estimated variance and the true variance differ by a factor of $N/(N-1)$. It can easily be seen that replacing N with $N-1$ in equation 9 does give an unbiased estimator.

Note that the MLE estimator for the variance of a Gaussian does approach the true variance for large N. An estimator with this property is called consistent by statisticians. Under fairly general conditions, MLE's are consistent.

Also note that bias depends on the functional form used for the fit. For example, in a lifetime fit to an exponential of the form $\frac{1}{\tau}e^{-t/\tau}$, the fit parameter can be either the lifetime τ or the decay rate $\Gamma = 1/\tau$. The fit for τ is unbiased, but the fit for Γ is biased.

Another desirable property of an estimator is that it have the smallest possible variance. (Don't confuse this variance with the Gaussian example above. The variance here is the variance of the estimator about its mean over many experiments.) An estimator with the smallest possible variance is known as an efficient estimator. Maximum likelihood estimators are efficient for large N but aren't necessarily so for small N.

The "best" estimator is an unbiased, efficient one. For large N, maximum likelihood estimators are unbiased and efficient, which is why maximum likelihood methods are popular. For small N, MLE's are not necessarily either unbiased or efficient. However, there does not exist a general method for finding an unbiased, efficient estimator, so people use maximum likelihood.

Since maximum likelihood estimators may be biased, it is essential to check this feature for a given measurement. This is usually done with Monte Carlo methods as described in section 8. If there is a bias, the result must be corrected so that published numbers are unbiased.

4 Uncertainties on Parameters

As important as getting a correct estimate of a parameter is knowing how precise this estimate is, that is, how likely is it to be close to the true value. This is usually expressed as the standard deviation of the estimator about its true value or as a confidence interval. A confidence interval at CL confidence level has a $1 - CL$ probability that the interval contains the true value. The probability that the confidence interval contains the true value is also known as the coverage. When someones gives a one standard deviation uncertainty on a result, they imply that the estimated value plus or minus one standard deviation is a 68% confidence interval (68% is the area under a Gaussian from the mean less σ to the mean plus σ). In this note, $\Delta\alpha$ will represent one standard deviation on the parameter α.

For one parameter and for large N, it can be shown (basically from the Central Limit Theorem) that the likelihood function is a Gaussian function of the parameter. This implies that $\ln L$ is a parabolic function of α, that is,

$$\ln L(\alpha) = \ln L(\hat{\alpha}) + \frac{1}{2}\frac{\partial^2 \ln L}{\partial \alpha^2}\bigg|_{\alpha=\hat{\alpha}} (\alpha - \hat{\alpha})^2 \,. \tag{12}$$

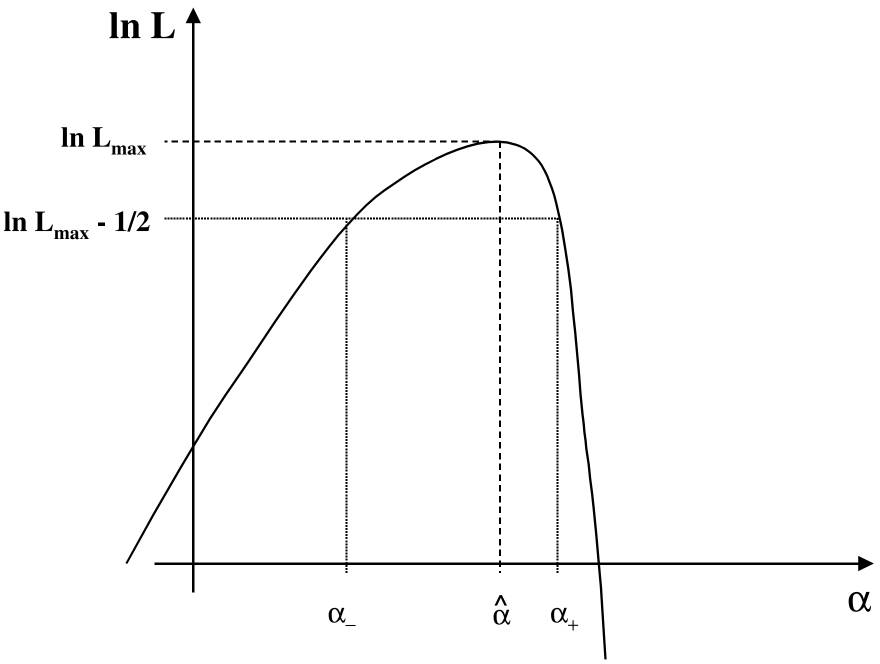

Figure 1: Example of an asymmetric log–likelihood function.

In this case, the rms standard deviation of the likelihood function is a good estimate of the rms standard deviation of the MLE for α. This is equivalent to

$$\Delta\alpha = -\frac{1}{\left.\frac{\partial^2 \ln L}{\partial \alpha^2}\right|_{\alpha=\hat{\alpha}}}. \tag{13}$$

For small N, $\ln L$ may not be parabolic. In fact, it may not even be symmetric about the maximum. In this case, a good estimate for the limits of the 68% confidence interval are those values for which $\ln L$ changes by 1/2 from its maximum value. For the parabolic case, this is clearly equivalent to the definition above. An asymmetric case is illustrated in figure 1, where α_- and α_+ are the limits of the 68% confidence interval. This is usually denoted as $\alpha = \hat{\alpha}^{+\alpha_+}_{-\alpha_-}$. If an n sigma confidence level is desired, then look to where $\ln L$ changes by $n^2/2$.

It is not always the case that the $\Delta \ln L = 1/2$ interval gives a 68% confidence interval. See reference [3] for a counter example involving a triangular probability distribution.

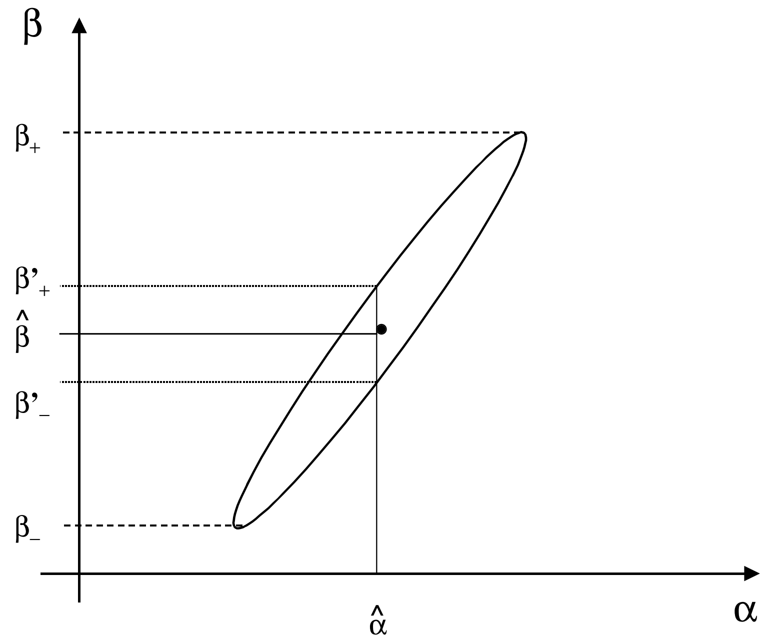

Figure 2: Uncertainties on correlated parameters. Contour is where $\ln L$ changes by 1/2 from its maximum value.

If there are multiple parameters, things are a little more complicated due to possible correlations between the parameters. Consider a case with two parameters (α and β), where the likelihood function is Gaussian in α and β. The contour in the $\alpha\beta$ plane where $\Delta \ln L = 1/2$ is an ellipse. If α and β are correlated, then the ellipse will be tilted, as illustrated in figure 2. Suppose we are interested in the uncertainty on $\hat{\alpha}$. If we hold α fixed, then $\beta'_{\pm}$ are where $\ln L$ changes by 1/2. These do not give the proper uncertainties. Instead, we must take the extreme points on the ellipse, given by $\beta_{\pm}$ in the figure. This is equivalent to finding the values of β where $\ln L$ changes by 1/2 and $\ln L$ has a maximum value with respect to α. For more than two parameters, the prescription is similar, namely, to find the uncertainty on a parameter, find where $\ln L$ changes by 1/2 while maximizing $\ln L$ with respect to all other parameters. In most cases, this will be done numerically (see section 10).

For multiple parameters, the correlations between the parameters are also important. The first order correlations are expressed as the covariance matrix V, which is an $M \times M$

matrix for a problem with M parameters defined by

$$V_{ij} = \overline{(\alpha_i - \overline{\alpha_i})(\alpha_j - \overline{\alpha_j})}, \tag{14}$$

where α_i and α_j are two parameters and the overlines indicate expectation values. The diagonal elements of the covariance matrix give the variance of the corresponding parameter. The off diagonal elements give the covariances. In the parabolic case, the inverse of the covariance matrix $U = V^{-1}$ is given by

$$U_{ij} = -\frac{\partial^2 \ln L}{\partial\alpha_i \partial\alpha_j}. \tag{15}$$

5 Normalization Issues

It is sometimes argued that it is not necessary to properly normalize the probability density function p in the likelihood function, since the normalization factor is a multiplicative factor in the likelihood, and hence an additive factor in the log–likelihood, and doesn't affect where the maximum of $\ln L$ is nor the shape about the maximum.

This argument is correct as long as the normalization doesn't depend on any of the parameters. If it does, then omitting the normalization will give incorrect results. My recommendation is to always normalize.

As an example, consider the Gaussian case given in equations 5 to 9. The normalization factor for the Gaussian is $1/\sqrt{2\pi\sigma^2}$, which doesn't depend on the data or on the mean. If we are only interested in the mean, then we could drop the normalization, and maximizing the likelihood function would give the same result. However, if we are interested in determining the variance, then the reader can easily verify that dropping the normalization and maximizing the likelihood function will give the nonsense result of $\hat{\sigma}^2 = \infty$.

6 Uncertainty on the Number of Events

It is common in High Energy Physics measurements to have multiple sources for a particular type of event. For example, suppose we wanted to determine the number of $B^0 \to J/\psi K_s$ events in a data sample by calculating the invariant mass of the decay products and looking at the number of events in the mass peak. Almost always, there would be additional events where the decay products were not all from a B^0, usually known as background events.

As an example, suppose the probability distribution of the background is flat over a mass range ΔM that we fit (it could be anything, but often a low order polynomial is sufficient to accurately describe the background). Also, let f be the probability that an event in our sample is a signal event (that is, one that actually came from a $B^0 \to J/\psi K_s$ decay). The probability density function in this case is

$$p(M|f, M_0, \sigma) = f \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(M-M_0)^2}{2\sigma^2}} + (1-f)\frac{1}{\Delta M}. \tag{16}$$

Calculating the likelihood function and maximizing with respect to f will give an estimate for f and an uncertainty Δf.

It is correct to take $N_s = fN$ as the number of signal events. It is also tempting to take the uncertainty as $\Delta N_s = \Delta f N$. We must be very careful in interpreting this uncertainty. It is actually the uncertainty of the number of signal events in our sample of exactly N events, that is, it is binomially distributed. If we are interested in determining a cross section or branching ratio, we need to know the standard deviation in the number of signal events over many experiments, which is larger due to fluctuations in N.

One solution is to take $N_s = fN$ with an uncertainty given by folding the uncertainties in f and N ($\Delta N = \sqrt{N}$) together in quadrature.

A second method is to add a term to the probability that is a Poisson distribution in the number of events N with mean μ. Then f is expressed as N_s/N, and N_s is one of the parameters. This method is called by some an extended likelihood method [4] and also yields a proper uncertainty on N_s. If you are not directly interested in the uncertainty on the number of events, then doing an extended likelihood fit adds an unnecessary complication.

7 Constrained Parameters

Sometimes a parameter in a fit may be known with some uncertainty from another source. For example, suppose we wished to determine the mass resolution for a certain decay mode by fitting the invariant mass spectrum. Also suppose that the mass is given in the Particle Data book as $M_0 \pm \sigma_M$. We can in principle improve our determination of the resolution by including a Gaussian term that expresses the probability of getting the measured value M in our set of data. The likelihood and log–likelihood functions become

$$L(\sigma) = \frac{1}{\sqrt{2\pi\sigma_M^2}} e^{-\frac{(M-M_0)^2}{2\sigma_M^2}} \prod_i^N \left[\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(m_i-M)^2}{2\sigma^2}}\right] \tag{17}$$

$$\ln L = -\ln(\sqrt{2\pi}\sigma_M) - \frac{(M-M_0)^2}{2\sigma_M^2} - N\ln(\sqrt{2\pi\sigma^2}) - \sum_i^N \frac{(m_i-M)^2}{2\sigma^2}. \tag{18}$$

This is known as a constrained fit.

If the parameter being constrained is known much, much better from other data than it can determined from the fit, then the constraint will simply move the parameter to the known value and is equivalent to a fit without the constraint (fixing the parameter to the known value). For example, you never see Planck's constant constrained to its measured value in High Energy Physics fits.

If the parameter being constrained is very poorly measured compared to how well the fit can determine it, then the constraint also will have little effect, since the optimal value with or without the constraint will be the same.

Constrained fits have their greatest power when the uncertainty on the measured value is similar to the uncertainty that would result from the fit without the constraint.

8 Simple Monte Carlo Tests

It is important, particularly for complex problems, to make sure that the likelihood function is behaving properly and to test whether there is any bias in the parameter estimation. A common method to do this is to write a short program that generates data according to the model assumed in the likelihood function. This is a powerful method to understand likelihood functions and other statistical questions. Many people call such programs "toy" Monte Carlos. Since these programs can be extremely useful in understanding subtleties in likelihood fits and statistical methods in general, I prefer to call them simple Monte Carlos (as distinguished from Monte Carlo programs that do detailed event generation and simulation).

For example, if you think your data is Gaussian distributed and you are interested in knowing how well you can determine the mean in a sample of 100 measurements, you would write a program to generate sets of 100 measurements with a known mean and resolution. Then you minimize the likelihood function for each set of measurements. If the average MLE of the mean is equal to the mean used to generate the data, then the method is unbiased. Also, the rms spread of the MLE's should equal the average of the uncertainties from the likelihood fits. If you do this test for a fit for the variance of a Gaussian, you will discover the bias discussed in section 3. Of course, a simple Gaussian problem can be done analytically and is well understood. However, for problems with many probability distributions with many parameters, the simple Monte Carlo is a very useful check.

Note that using a simple Monte Carlo does not check that your likelihood function properly models the data, which is always a crucial question. For example, if you think the data is Gaussian distributed and use that for the likelihood function, but it really has some other distribution (such as, a Lorentzian), then your fit will be incorrect. A simple Monte Carlo test based on a Gaussian distribution will not reveal the problem.

9 Goodness of Fit

In order to check that the likelihood function properly models the data, it is important to perform some form of goodness of fit check. Often people show binned plots of the data with the function determined from the MLE parameters overlayed. Although information is lost in the binning of the data (particularly in problems with multidimensional data), obviously deficient modelling can be found this way.

Chi-squared (χ^2) fits have the advantage that they automatically provide a goodness of fit estimation, since the estimated parameter should be distributed according to a χ^2 distribution with the appropriate number of degrees of freedom. Sometimes, people do an unbinned likelihood fit and then do a binned χ^2 test for goodness of fit.

Unfortunately, the likelihood often does not give a good measure of goodness of fit. As an example, consider fitting lifetime data to an exponential. The likelihood and log–likelihood functions are

$$L = \prod_i^N \Gamma e^{(-\Gamma t_i)} \tag{19}$$

$$\ln L = N \ln \Gamma - \sum_i^N \Gamma t_i, \tag{20}$$

where Γ is the parameter of interest (the total width in this case), t_i is the proper decay time of the i^{th} event, and N is the number of events. The MLE of Γ is

$$\hat{\Gamma} = \frac{N}{\sum_i^N t_i} \tag{21}$$

and the value of the likelihood function at it minimum is

$$\ln L|_{\Gamma=\hat{\Gamma}} = -N\left(1 + \ln \frac{\sum t_i}{N}\right). \tag{22}$$

Since $\ln L$ depends only on the number of events and the average value of the data, any data sets with similar size and average will give similar log–likelihood values, no matter what the distribution actually is.

For more discussion on goodness of fit tests when using maximum likelihood methods, see reference [2].

10 Numerical Methods

Most maximum likelihood fits are too complex to be solved analytically. Fortunately, there exist numerical methods for maximizing functions. The most common method in experimental high energy physics is to use a software package written at CERN called MINUIT. This program actually minimizes a function of one or more parameters, returns the optimal values of those parameters, and returns the uncertainties on those parameters. Since MINUIT is a minimization program, you provide it with the negative of the log–likelihood function. To find out how to use MINUIT, see its documentation [5]. For a brief discussion of the proper use of MINUIT in likelihood fits, see reference [6].

11 Systematic Errors

A maximum likelihood fit will return the statistical errors on the parameters being estimated. In addition, there may be systematic errors due to uncertainties in the modelling of the data, uncertainty in parameters of the likelihood function that aren't being estimated, and uncertainties in numbers needed to convert the estimated parameter to the desired physics quantity. For a more detailed discussion of systematic errors, including issues relevant to likelihood fits, see reference [7]

12 Sin2β Likelihood Function

As an example of a realistic likelihood function used in High Energy Physics, we will look at the Run I $\sin 2\beta$ analysis. The reader unfamiliar with the physics of this measurement should consult reference [8]. The basic idea is that the decays $B_0 \to J/\psi K_s$ and $\overline{B}^0 \to J/\psi K_s$ have proper decay times that have an oscillating part that depends on the CP–violation parameter $\sin 2\beta$. The goal of the analysis is to extract a measurement of this parameter.

The measured quantities consist of the proper decay time, the invariant mass of the $J/\psi K_s$, and a tag of whether the initial particle was a B^0 or $\overline{B}^0$. The actual fit used in the analysis has over 60 parameters, including mutiple background sources, multiple types of tags, and possible asymmetries of the tags. A full description is beyond the scope of this note, but the interested reader can find the details in reference [8].

A simplied model of the fit is presented here that contains most of the essential features. We assume that there are two souces of events: signal and background. The signal is $J/\psi K_s$ events from B^0 and $\overline{B}^0$ decays. The background events are from tracks that reconstruct to a $J/\psi K_s$ but didn't come from a B decay.

The probability P_S for the signal contains three factors for the proper time, mass, and tagging efficiency, that is,

$$P_S = T_S(t) M_S(m) \mathcal{E}_S(Q), \tag{23}$$

where t is the proper time, m is the invariant mass, and Q is the tag (Q = -1 for a $\overline{B}^0$ tag, 0 for no tag, and +1 for a B^0 tag.

The proper time distribution for the signal is

$$\tilde{h}(t) = \frac{1}{\tau} e^{-t/\tau} \left[1 \pm \sin 2\beta \sin(\Delta m t)\right], \tag{24}$$

where τ is the average B^0 lifetime, Δm is the mass difference of the eigenstates, and the sign depends on whether the initial particle was a B^0 or $\overline{B}^0$.

There is an efficiency ϵ_S for tagging a signal event and a probability f of mistagging an event, that is, tagging an event as a B^0 when it really was a $\overline{B}^0$ and vice versa. Thus, the probability of getting a B^0 tag at time t is (note that the tagging efficiency is included in $\mathcal{E}$)

$$\begin{aligned} h(t) &= \frac{1}{\tau} e^{-t/\tau} \left[1 - \sin 2\beta \sin(\Delta m t)\right](1-f) + \frac{1}{\tau} e^{-t/\tau} \left[1 + \sin 2\beta \sin(\Delta m t)\right] f & (25) \\ &= \frac{1}{\tau} e^{-t/\tau} \left[1 - (1-2f) \sin 2\beta \sin(\Delta m t)\right] & (26) \\ &= \frac{1}{\tau} e^{-t/\tau} \left[1 - D \sin 2\beta \sin(\Delta m t)\right] & (27) \end{aligned}$$

where $D \equiv 1 - 2f$ is known as the dilution. Doing this for the three tagging cases (Q = -1, 0, +1) gives

$$h(t) = \frac{1}{\tau} e^{-t/\tau} \left[1 - DQ \sin 2\beta \sin(\Delta m t)\right] \tag{28}$$

Since we measure the proper decay time with a non-zero resolution, we must fold this lifetime distribution with a Gaussian whose width is the resolution. The resolution σ_t is given by the secondary vertex fit and varies from event to event. Thus, we have $T_S(t) = h * g(t)$, where $h * g$ symbolizes the convolution of $h(t)$ with a Gaussian function g.

The mass distribution is Gaussian with a resolution σ_m that is also given by the secondary vertex fit and varies from event to event, giving

$$M_S(m) = \frac{1}{\sqrt{2\pi\sigma_m^2}} e^{-\frac{(m-M_B)^2}{2\sigma_m^2}} \tag{29}$$

The $\mathcal{E}_S(Q)$ function is the probability of getting a tag Q, where $Q = 1$ for a B^0 tag, -1 for $\overline{B}^0$ tag, and 0 for no tag. We assume that B^0's and $\overline{B}^0$'s are tagged with equal efficiency ϵ_S, giving

$$\mathcal{E}(Q) = \begin{cases} \frac{\epsilon_S}{2} & Q = -1 \\ 1 - \epsilon_S & Q = 0 \\ \frac{\epsilon_S}{2} & Q = +1 \end{cases} \tag{30}$$

The reader should check that P_S is properly normalized when summed over Q and integrated over m and t.

The probability for the background also contains three terms for the time dependence, mass dependence, and tagging efficiency, that is,

$$P_B = T_B(t) M_B(m) \mathcal{E}_B(Q), \tag{31}$$

To simplify things, we assume that the background consists only of prompt events with a linear mass spectrum over the mass range W. The choice of a shape for the mass distribution in a real example would be one that fits the data well. We assume that the background events are tagged with an efficiency ϵ_B, which can differ from the signal tagging efficiency. We also assume that there is an equal probably of tagging a background event as a B^0 or $\overline{B}^0$. This gives

$$T_B(t) = \frac{1}{\sqrt{2\pi\sigma_t}} e^{-\frac{t^2}{2\sigma_t}} \tag{32}$$

$$M_B(m) = \frac{1+bm}{2W} \tag{33}$$

$$\mathcal{E}(Q) = \begin{cases} \frac{\epsilon_B}{2} & Q=-1 \\ 1-\epsilon_B & Q=0 \\ \frac{\epsilon_B}{2} & Q=+1 \end{cases} \tag{34}$$

Note that the dilution D appears in these probabilities only as a product with $\sin 2\beta$, which is the parameter we wish to determine. Thus, it is not possible to fit for both D and $\sin 2\beta$. Thus, the dilution must be determined from other data. The authors included the measured dilution as a constraint. In our case, including it as a constaint or fitting for the product $D \sin 2\beta$ and dividing out the measured value of D will give the same result. However, using the constraint will automatically include the error on D in the error on $\sin 2\beta$.

The authors also chose to include constraints on the B^0 lifetime τ and the mixing frequency Δm. They could have chosen to also constrain the B mass but did not, apparently because they felt that the PDG value was sufficiently more accurate than a determination from the $J/\psi K_s$ data.

Although the number of events was not the main goal of this measurement ($\sin 2\beta$ was), the authors chose to do an extended likelihood fit.

Putting all of this together gives a log–likelihood function of

$$\ln L = N_S + N_B + \sum_i^N \left[N_S P_S^i + N_B P_B^i \right] \tag{35}$$

$$- \ln(\sqrt{2\pi\sigma_D^2}) - \frac{(D-D_m)^2}{2\sigma_D^2} \tag{36}$$

$$-\ln(\sqrt{2\pi\sigma_M^2}) - \frac{(M_B - M_0)^2}{2\sigma_M^2} \tag{37}$$

$$-\ln(\sqrt{2\pi\sigma_\tau^2}) - \frac{(\tau - \tau_0)^2}{2\sigma_\tau}, \tag{38}$$

where the sum over i is over events, D_m is the measured dilution, M_0 is the PDG value of the mass with uncertainty σ_M, and τ_0 is the PDG value of the lifetime with uncertainty σ_τ. Note that the normalizations on the constraint Gaussians do not depend on the fit parameters and could be dropped.

The actual $\sin 2\beta$ analysis includes details such as whether events were measured in the SVX or not, long lived backgrounds, scale errors on uncertainties of measurements of lifetimes and masses, possible tagging asymmetries, and constraints to measurements of efficiencies. This results in a more complicated log–likelihood function with a total of 67 parameters, but the essential form is the same as presented here. Readers interested in the full, gory details can see reference [8].

References

[1] Louis' note on likelihood functions (reference to be specified later).

[2] J. G. Heinrich "Can the Likelihood–Function Value be Used to Measure Goodness of Fit?", Cdf Note 5639 (2001).

[3] F. Porter, Nuclear Inst Meth A368, 763 (1996).

[4] L. Lyons, "Statistics for Nuclear and Particle Physicts," CUP (1986).

[5] F. James, *MINUIT—Function Minimzation and Error Analysis*, Version 94.1, CERN Program Library Long Writeup D506, CERN, Geneva Switzerland, 1994.

[6] Craig's note on MINUIT (to be written).

[7] Craig's note on systematics (reference to be specified).

[8] Joel Heinrich, Nigel Lockyer, and Kevin Pitts, "A Measurement of Sin 2β Using Multiple Flavor Tags", CDF Note 4747 and references therein.