Probabilistic Event Weighting to Separate Signal from Background

GlueX-doc-839

Mike Williams, Matt Bellis and Curtis A. Meyer

Carnegie Mellon University

June 27, 2007

Abstract

A common situation in physics is to have a signal which is mixed in with some background. In this note, we describe a procedure for determining on an event-by-event basis the chance (Q-factor) that a given event comes from the signal distribution. This chance can then be used as an event weight in subsequent analysis procedures, allowing one to more directly access the true spectrum of the signal. In addition to the Q-factor, we describe a method to determine the error on the Q-factor, σ_Q and how these errors are combined to determine the errors on observables.

1 Introduction

In many analysis situations one has background events which cannot be cleanly separated from desired signal events. One typical way to handle such events is with so-called side-band subtraction. Events from outside the signal region are subtracted from those in the signal region to attempt to create a sample that has the background removed. While this can work for relatively simple situations, it becomes problematic if the kinematics of the background are different from the signal, or if the signal is actually a function of several variables such that the binning to produce the subtraction is severely limited by statistics and/or systematic errors. In this note, we describe a procedure which allows us to assign a signal and background *chance* to each event in the sample. Fits can then take advantage of this weight to separate the signal from the background.

As a note, while the quantity that we compute may be called a probability, there is some concern that this is not precise. While we use it as if it was a probability, we want to be careful to distinguish this here. In particular, rather than calling this a probability, we will refer to it as a Q-factor, or simply as Q of a given event.

2 Determination of the Event Weight

2.1 The Method

Consider a data set which can composed of n events e_i which can be described by m *coordinates*, ξ_i, (where m is at least two).

$$e_i(\xi_j) \quad (i = 1, n), (j = 1, m)$$

This data set is made up of some combination of *signal*, S, and *background* , B events. Both the signal and background are functions of the coordinates, $S(\xi_j)$ and $B(\xi_j)$. The coordinates can be masses, energies, angles, *etc.*. For this procedure, we need to know a-priori the functional shape of the signal distribution and the background distributions as a function of one coordinate, ξ_r. We will refer to ξ_r as the reference coordinates. For example, if the reference coordinate, ξ_r, is a mass, the distribution might be a known line shape such as a Gaussian or Breit-Wigner shape and the background is a polynomial in ξ_r.

The aim of this procedure is for a given event, e_i, to be able to explicitly identify the chance that the event is a signal event, Q_i or a background event, $(1 - Q_i)$. In order to proceed, we need to define a *distance measure* that can identify how close together two events are in the non-reference (not ξ_r) coordinates. For a given coordinate, ξ_j, we will assume that the maximum difference between any pair of events is r_j. We can then define the simple distance function given in equation 1.

$$d_{ij} = \sum_{k\neq 2}^{m} \left[\frac{(\xi_k)_i - (\xi_k)_j}{r_k} \right]^2 \tag{1}$$

While we could take the square root of d_{ij} in equation 1, this would not change anything in the procedure. It also saves the computation of a large number of square roots.

We note that the above distance function is a reasonable choice. In a traditional approach to separating the signal and background, one would typically bin the events in the non-reference coordinates and then fit the reference coordinate in each bin. The simple act of binning is a distance function. However, in this procedure we guarantee a certain sample size by having a dynamic bin width.

For each event in the data set, e_i, we compute the distance to all events in the data set, d_{ij} $(i = 1, n)$. We then retain the N_d events which are closest to e_i (this set always includes e_i). Typical for the work that we have done, we choose N_d on the order of 100. Using the N_d retained events, we will fit the distribution composed of the reference variables $(\xi_r)_j$ to the known signal shape plus a background. To avoid binning issues with small samples, this fit is performed using an unbinned maximum likelihood method.

We will take $f_s(\xi_r)$ to be the function describing the signal distribution and assume that the fit will yield a set of fit parameters, $\vec{\eta}_s$. Similarly, we take the background function to be $f_b(\xi_r)$ and the fit yields a set of fit parameters, $\vec{\eta}_b$. We will write the set of all fit parameters as $\vec{\eta}$ and the corresponding covariance matrix from the fit as C_η. For the given event, e_i, we can use the fit results to compute the expected number of signal, $s_i = f_s((\xi_r)_i, \vec{\eta})$, and the background, $b_i = f_b((\xi_r)_i, \vec{\eta}$. Using this, we can define a quality factor, or Q-value for the given event as in equation 2.

$$Q_i = \frac{s_i}{s_i + b_i} \tag{2}$$

This Q-value can be thought of as the chance that the event came from the signal distribution, while $1 - Q$ is the chance that the event came from the background distribution. In further analysis, the Q-factor can be used to weight the event.

If we produce a histogram of one of the variables, ξ_j, weighted by the Q_j, the resulting distribution would be the signal distribution in that variable. Similarly, a two-dimensional histogram of two ξ_js could also be weighted by Q_j. As an example, one could imagine a Dalitz plot in two invariant masses (squared).

In addition to the Q-value for each event, we have a functional description for the signal and background for the events close to the given event. If we integrate these distributions, we can define a variable that represents the chance of finding a signal event near the given event. We will call this the R-factor. For the event e_i, these integrals are given in equation 3 and 4.

$$S_i^{total} = \int f_s(\xi, \vec{\eta}) d\xi \tag{3}$$

$$B_i^{total} = \int f_b(\xi, \vec{\eta}) d\xi \tag{4}$$

$$\tag{5}$$

The R-factor for event e_i is then given in equation 6.

$$R_i = \frac{S_i^{total}}{S_i^{total} + B_i^{total}} \tag{6}$$

2.2 Estimation of Errors

While the Q-factor as defined in equation 2 is a very useful tool in analysis, it is important to be able to assign a reasonable error to a given Q_i, σ_{Q_i}. We would also like to understand how the individual σ_Qs are combined to yield the error on some measurable quantity. To compute the error on Q, we will use the functional forms for the signal and background, $f_s(\vec{\eta})$ and $f_b(\vec{\eta})$, and the covariance matrix of the fit parameters, C_η. For Q, we have

$$Q(\xi_r) = \frac{f_s(\xi_r,\eta)}{f_s(\xi_r,\eta)+f_b(\xi_r,\eta)} \,. \tag{7}$$

The error in Q can be written as

$$\sigma_Q^2 = \sum_{i,j} \frac{\partial Q}{\partial \eta_i} (C_\eta)^{-1}_{ij} \frac{\partial Q}{\partial \eta_i} \,. \tag{8}$$

In combining errors to yield a total error, we recall that there are two limiting cases. If the errors are all completely uncorrelated, then the errors will add in quadrature to give

$$\sigma^2_{total} = \sum_{i=1}^{n} \sigma_i^2 \,.$$

In the case where the events are 100% correlated, the errors add as a simple sum

$$\sigma^2_{total} = \left(\sum_{i=1}^{n} \sigma_i\right)^2 \,.$$

If we consider how the Q values are determined, we note that for a given event, there is very large correlations with the N_d closest events. Probably something approaching 100%. This correlation propogates throughout the entire data set, so in the end, there is a very large correlation inherent in the procedure between any pair of Qs. This means that we expect that the errors will add as if the events are correlated, rather than uncorrelated. This correlated error is that which is associated with the procedure to determine Q. In addition to this, there will also be an error associated purely with the statistics of the quantity being measured.

If we are measuring the signal as a function of one of the observables, ξ_j, then we would bin the n events in $(\xi_j)_i$. We define a bin function $U(\xi_j, l_j, h_j)$, where l is the lower and h is the higher edge of the bin. The function U is 0 when ξ_j is outside of the bin and 1 when it is inside the bin. The total number of events in a particular bin is then given in equation 9.

$$N = \sum_{i=1}^{n} Q_i \cdot U((\xi_j)_i, l_j, h_j) \tag{9}$$

The fit error on this number is given as

$$\sigma_N = \sum_{i=1}^{n} \sigma_{Q_i} \cdot U((\xi_j)_i, l_j, h_j) \tag{10}$$

and the statistical error on the number of events is given by Poisson statistics. For large vales of N, this is $\sqrt{N}$, however for small N, the error is estimated as shown in table 1. The total error on N is then given as the statistical and fit errors added in quadrature.

Count k	Normal Error $\sqrt{k}$	Lower Error $k - a_l$	Upper Error $a_h - k$	Approx Upper $1 + \sqrt{k + 0.75}$
0	0	0	1.84	1.87
1	1	0.83	2.30	2.32
2	1.41	1.29	2.64	2.66
3	1.73	1.63	2.92	2.94
4	2.00	1.91	3.16	3.17
5	2.24	2.16	3.37	3.40
6	2.45	2.38	3.58	3.60
7	2.65	2.58	3.77	3.79
8	2.83	2.77	3.94	3.96
9	3.00	2.94	4.11	4.12
10	3.16	3.11	4.26	4.28
11	3.32	3.26	4.41	4.43
12	3.46	3.41	4.55	4.57
13	3.60	3.56	4.69	4.71
14	3.74	3.69	4.82	4.84
15	3.87	3.83	4.95	4.97
16	4.00	3.95	5.07	5.09
17	4.12	4.08	5.19	5.21
18	4.24	4.20	5.31	5.33
19	4.36	4.31	5.42	5.44
20	4.47	4.47	5.54	5.56
50	7.07	7.05	8.11	8.12
100	10.00	9.97	11.00	11.04
...	...	...	...	...
k	$\sqrt{k}$	$\sqrt{k}$	$\sqrt{k}$	$1 + \sqrt{k + 0.75}$

Table 1: Upper and lower limit errors for small numbers of counts.

3 Examples

3.1 Resonance Shapes

In the following sections we will examine several toy models. In these, we will assume that the reference variable is a mass whose signal distribution is a peak on top of a smooth background. The background will be parametrized as a polynomial of the mass, while the signal shape can be Gaussian shape, a Breit-Wigner shape, or a convolution of the two—known as a Voigtian function. In equation 11 is the normalized form of the Gaussian. The peak is centered around the mass m_0 with a width of σ. The integral over the mass, m, from $-\infty$ to ∞ is one. Similarly, the normalized Breit-Wigner function is given in equation 12. The function is centered at m_0 and has a width Γ. For the Voigtian function, the normalized form is given in equation 13. The function is centered at m_0 and has a Gaussian width of σ and a Breit-Wigner width of Γ. The function $w(z)$ is the complex error function.

$$
\begin{aligned}
g(m, m_0, \sigma) &= \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{m-m_0}{\sigma}\right)^2} && (11) \\
bw(m, m_0, \Gamma) &= \frac{1}{\pi} \frac{\frac{\Gamma}{2}}{(m-m_0)^2 + \left(\frac{\Gamma}{2}\right)^2} && (12) \\
v(m, m_0, \sigma, \Gamma) &= \frac{1}{\sqrt{2\pi}\sigma} Real \left[w\left(\frac{1}{\sqrt{2}\sigma}(m-m_0) + i\frac{\Gamma}{2}\frac{1}{\sqrt{2}\sigma} \right) \right] && (13)
\end{aligned}
$$

In examining our toy models, we will also consider examples where the background and signal are separable using the R parameter (equation 6, and cases where they are not. In the separable cases, we have large regions of phase space where R is either very close to zero (all background) or very close to one (all signal).

3.2 A Gaussian Signal

In this example we have generated a 3-dimensional data set in mass, and two angles. The mass distribution is assumed to follow a Gaussian shape of known centroid and approximately known σ. The background is taken as a polynomial in mass. The data and background distributions for the mass are shown in Figures 1. In addition to the mass, the events are distributed if $\cos\theta$ and ϕ angles as shown in Figure 2. The exact functional forms that we use are as follows.

$$
\begin{aligned}
S(m, \cos\theta, \phi) &= \frac{1}{\sqrt{2\pi}\sigma} \cdot e^{-\frac{(m-m_0)^2}{2\sigma^2}} \left[(\sin(2\phi) + \cos(\phi)) \cdot (\sqrt{\frac{5}{4\pi}}) \cdot (\frac{5}{2}\cos^2\theta - \frac{3}{2}) \right]^2 && (14) \\
B(m, \cos\theta, \phi) &= [2.0 + 10 \cdot (m - m_0)] \cdot \left[\sqrt{\frac{3}{4\pi}} \cos\theta \sin 2\phi \right]^2 && (15)
\end{aligned}
$$

The center of the Gaussian, m_0, is set equal to 0.750 and the σ of the Gaussian is set equal to 0.005. The data are generated with masses m in the range 0.70 to 0.80 with 6000 events in the signal and 9000 events in the background.

We define the distance function as in equation 1

$$
d_{ij} = \left[\frac{1}{2} (\cos\theta_i - \cos\theta_j) \right]^2 + \left[\frac{1}{2\pi} (\phi_i - \phi_j) \right]^2 ,
$$

where Figure 3 shows the distribution of the distance between every pair of points in the 15000 event data set.

$$
\begin{aligned}
s(m) &= A \cdot e^{-\frac{(m-0.75)^2}{2\sigma^2}} && (16) \\
b(m) &= B + C \cdot m && (17)
\end{aligned}
$$

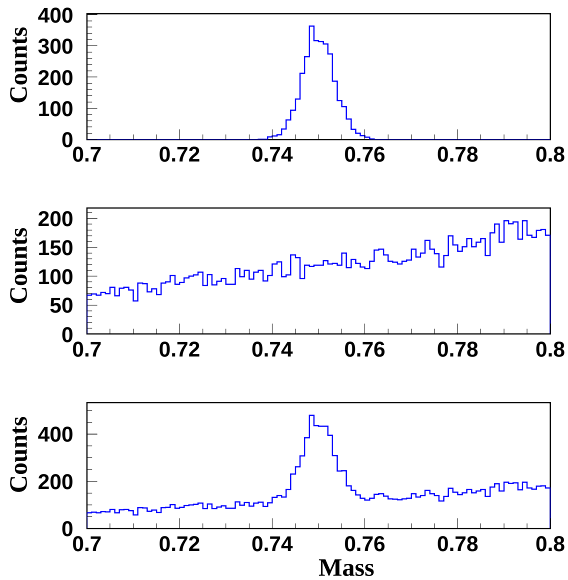

Figure 1: Generated Mass distribution. The upper plot is the signal distribution, the middle plot is the background distribution and the bottom plot is what is observed.

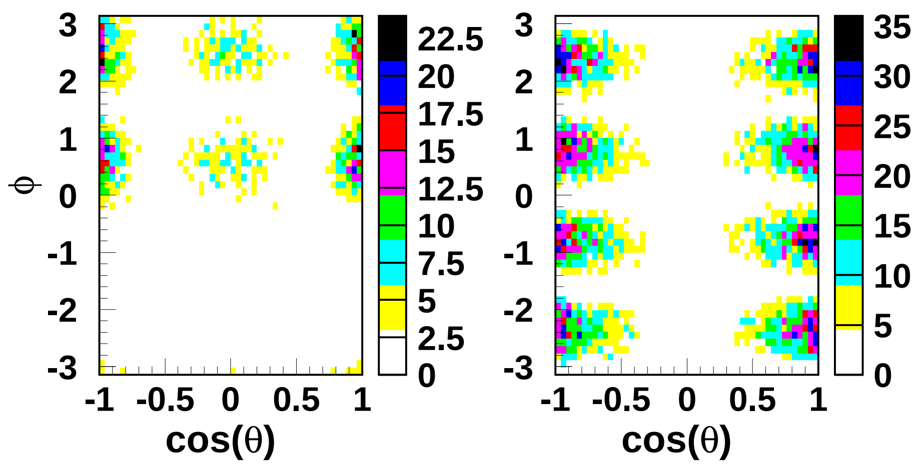

Figure 2: Generate angular plots of ϕ versus $\cos(\theta)$. The left-hand plot is the data distribution while the right-hand plot is the background distribution.

Before proceeding, we have taken taken three events from the data set which we will refer to as (a), (b) and (c). Figure 4 shows the mass distribution for the closest events to each of the three points. Each of these three distributions are then fit by the signal and background functions as given in equation 16 and 17. The mass of each event is indicated

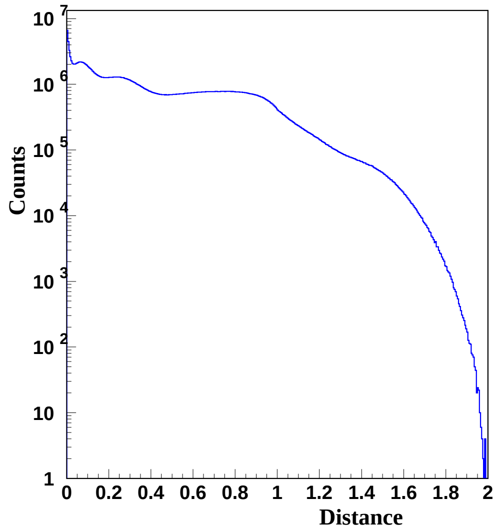

Figure 3: The distance function plotted for the sample data set described in equations 14 and 15.

by the arrows in the three figures. In carrying out the fits, we have held the centroid of the Gaussian constant at $m_0 = 0.75$, but let the σ vary in the fit. The resulting fit parameters are given in Figure 2

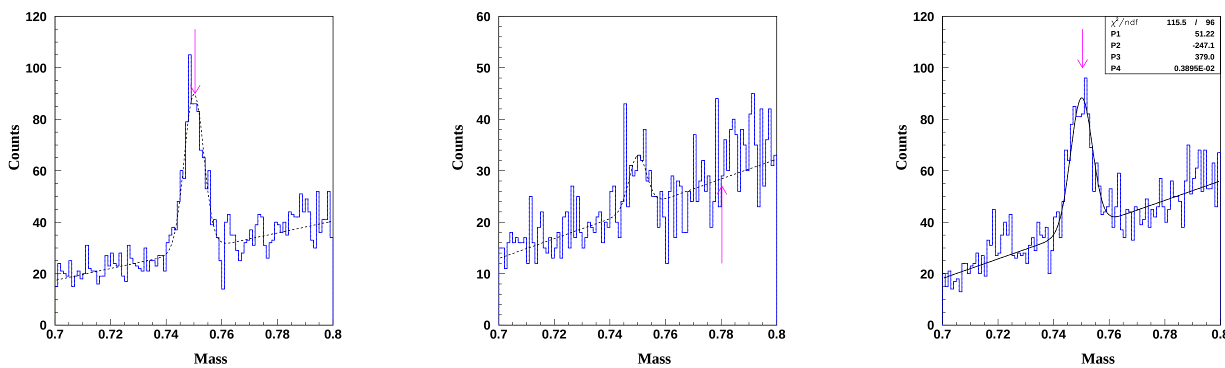

Figure 4: The left-hand figure (a) is at $\cos\theta = 0.970$ and $\phi = 0.0674$. The center figure (b) is at $\cos\theta = 0.645$ and $\phi = -2.48$. The right-hand figure (c) is at $\cos\theta = -0.866$ and $\phi = -0.776$.

Bin	A	σ	B	C	R	m	Q
(a)	60.8	0.0050	-143.7	230.08	0.209	0.75	0.676
(b)	10.4	0.0031	-122.2	193.09	0.034	0.78	0.000
(c)	51.2	0.0039	-247.1	378.98	0.119	0.75	0.571

Table 2: The fit parameters for the three distributions shown in Figure 4. The centroid has been held fixed at $m_0 = 0.75$ in each case. Using the fit parameters, we compute the R-factor (equation 6 and the Q-factor (equation 2 for each event.

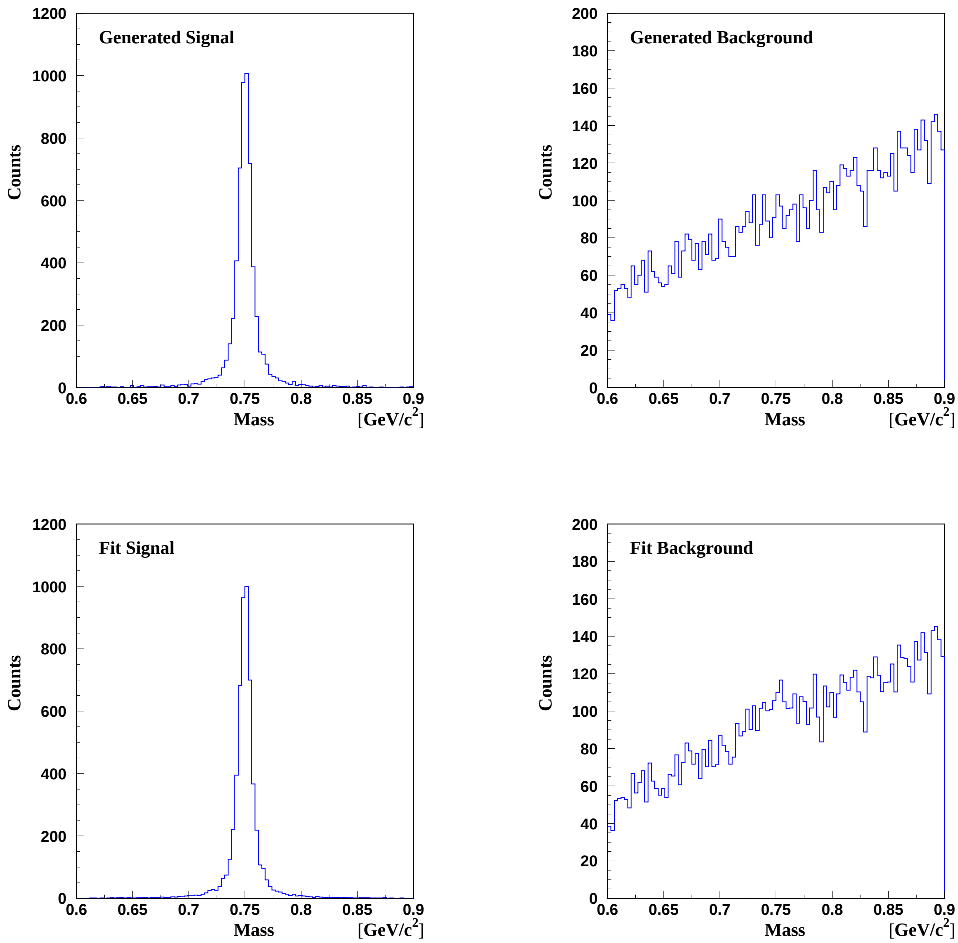

Figure 5: The upper row shows the generated signal and background. There are 6000 signal events and 9000 background events. The lower row shows the extracted signal and background using the nominal fit probabilities.

3.3 Voigtian Signals

In this next example, we again consider a 3-dimensional data set where the reference variable is a mass. Here, the signal distribution is described by a voigtian function of mean $m_0 =$

0.750, $\sigma = 0.005$ and $\Gamma = 0.009$. The background's mass distributions is described by a linear polynomial

$$b(m) = 1. + 3. \cdot (m - m_0)$$

We then fold these mass distributions with two different angular distributions as given in equations 18,19, 20 and 21. The two different angular distributions lead to quite different R-distributions as seen in Figure 6. We refer to the first set of these as *separated* in that there are regions of phase space where we have nearly all signal or all background. The second set is *non-separated* as there is no region in phase space where we have all signal or all data.

$$S_s(m, \cos\theta, \phi, \sigma, \Gamma, m_0) = v(m, m_0, \sigma, \Gamma) \cdot \left[(\sqrt{\frac{5}{4\pi}} (\frac{3}{2} \cos^2\theta - \frac{1}{2})(\sin 2\phi + \cos\phi) \right]^2 \quad (18)$$

$$B_s(m, \cos\theta, \phi, m_0) = [1.0 + 3.0 \cdot (m - m_0)] \cdot \left[\sqrt{\frac{3}{4\pi}} \cos\theta \sin 2\phi \right]^2 \quad (19)$$

$$S_n(m, \cos\theta, \phi, \sigma, \Gamma, m_0) = v(m, m_0, \sigma, \Gamma) \cdot \left[(\frac{1}{\sqrt{4\pi}} + 0.2 \cdot \sqrt{\frac{3}{4\pi}} \cos\theta \sin(2.2\phi) \right]^2 \quad (20)$$

$$B_n(m, \cos\theta, \phi, m_0) = [1.0 + 3.0 \cdot (m - m_0)] \cdot \left[\frac{1}{\sqrt{4\pi}} + 0.1 \cdot \sqrt{\frac{3}{4\pi}} \cos\theta \sin 2\phi \right]^2 \quad (21)$$

In studying these two distributions, we have carried out our analysis using the 100 closest events. This yields our Q- and R-factors for each event. Knowing the true weight functions, we are able to compute a true Q-factor for each event in the data set. It is then possible to compare the true Q-factor to the fit one. This is shown in Figure 8 for the two data sets. The only striking discrepancy occurs for events whose true Q-factor is quite close to one. For some fraction of these, the fit Q-factor comes out somewhat less than one. These events appear to arise in regions where the expected number of events are quite small. While the true Q-factor of a given event can be one, when things are averaged over the hypersphere, the fit probability is somewhat reduced.

In Figures 9 and 10 are shown the mass distributions for the separated and non separated samples. These show the total distribution and then the mass weighted by both Q and $1-Q$. There is very clean separation of the events using the Q-factor.

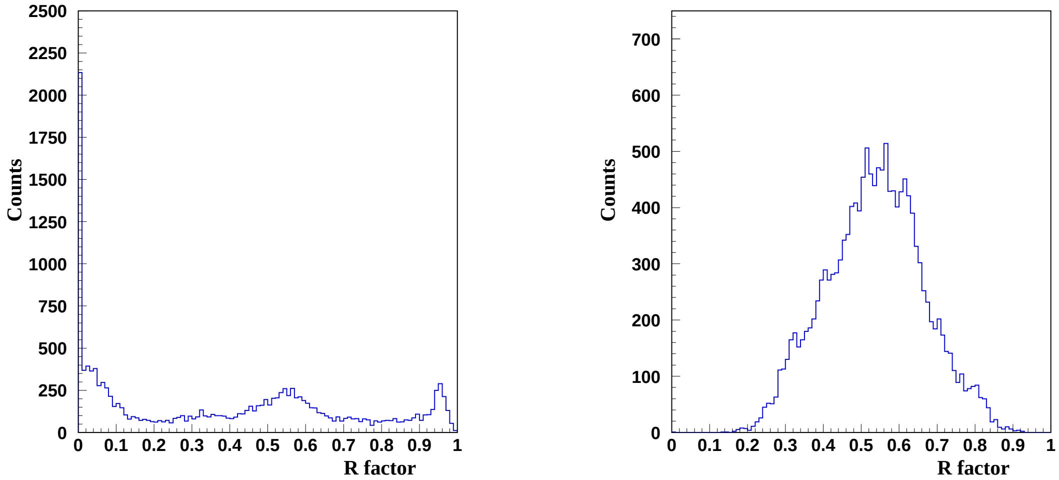

Figure 6: The left-hand plot shows the R-factor for the distributions given in equations 18 and 19. The right-hand plot shows the R-factor for the distributions given in equation 20 and 21.

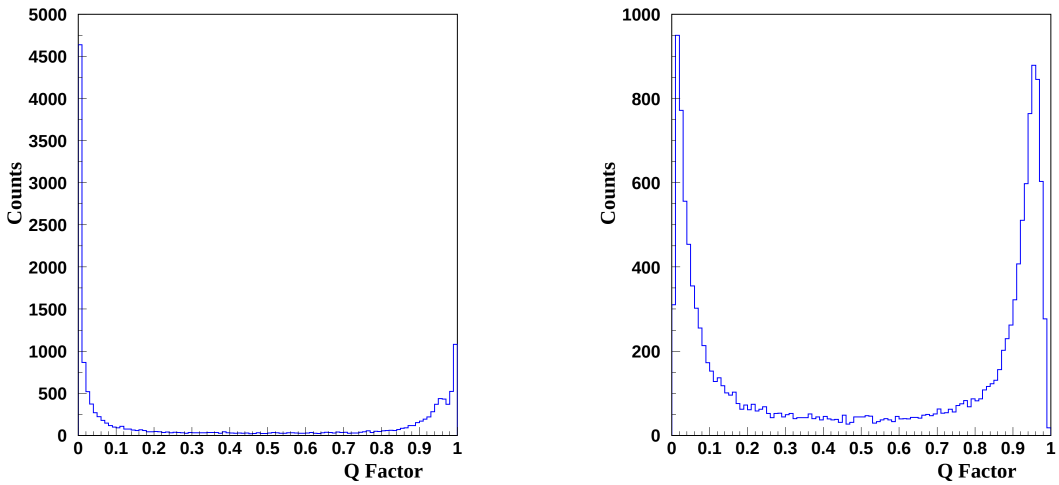

Figure 7: The left-hand plot shows the Q-factor for the distributions given in equations 18 and 19. The right-hand plot shows the Q-factor for the distributions given in equation 20 and 21.

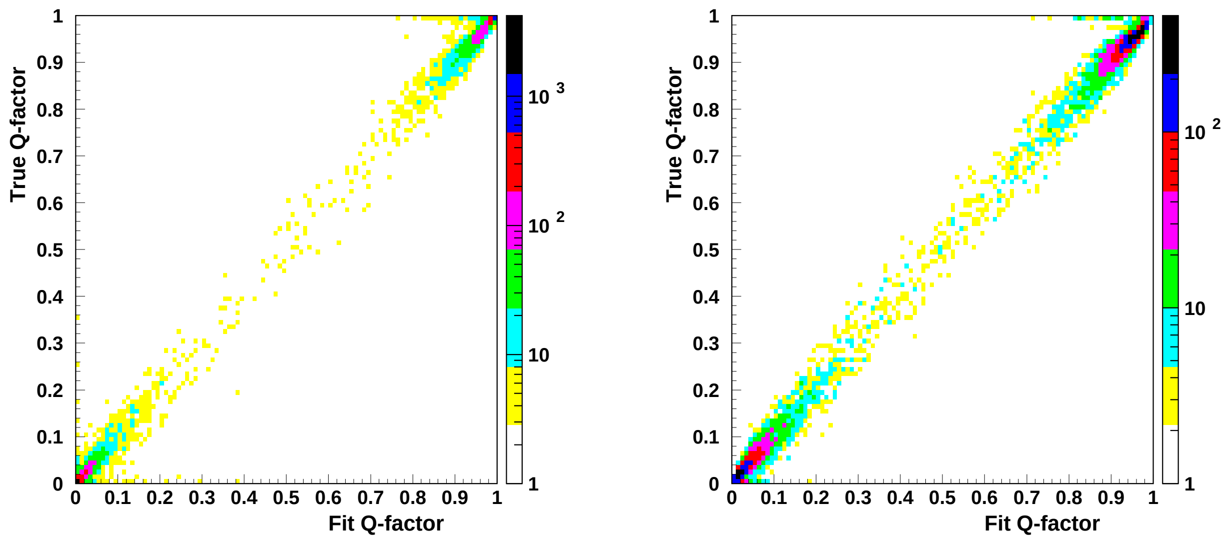

Figure 8: The left-hand plot show the actual Q-factor of each event plotted against the fit Q-factor for the separated data set. The right-hand plot shows the same thing for the non separated data set. In both cases, there is a very strong correlation of the two.

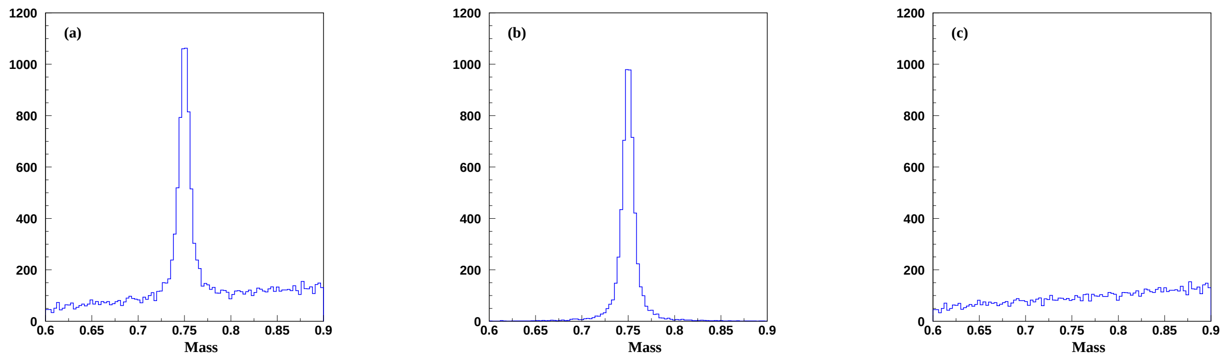

Figure 9: (a) show the mass distribution for the separated data set. (b) shows the mass weighted by the Q-factor and (c) shows the mass weighted by $1 - Q$.

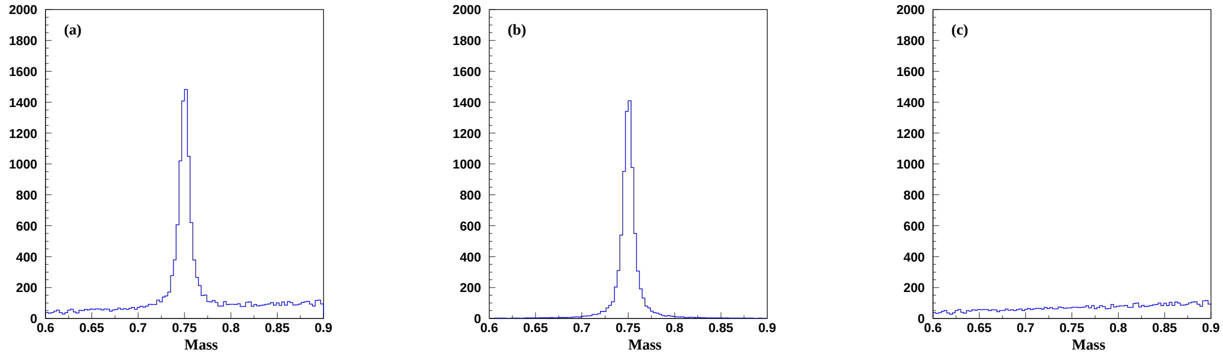

Figure 10: (a) show the mass distribution for the non separated data set. (b) shows the mass weighted by the Q-factor and (c) shows the mass weighted by $1 - Q$.

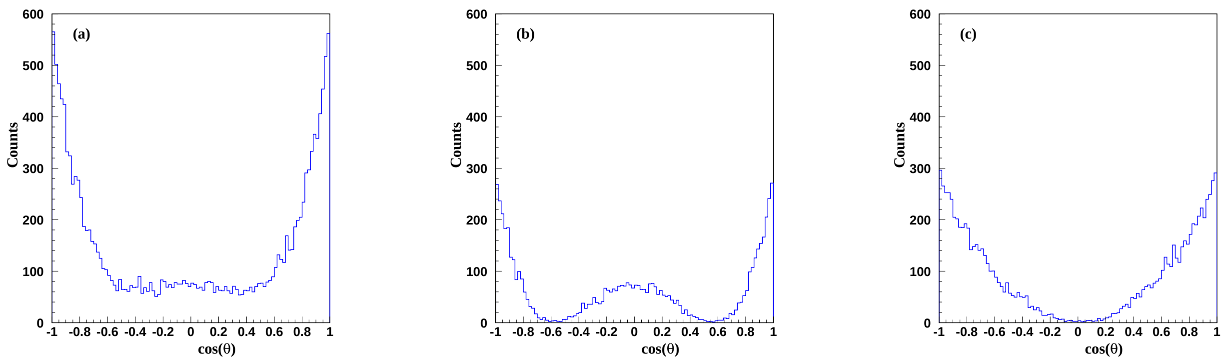

Figure 11: (a) show the $\cos\theta$ distribution for the separated data set. (b) shows $\cos\theta$ weighted by the Q-factor and (c) shows the $\cos\theta$ weighted by $1 - Q$.

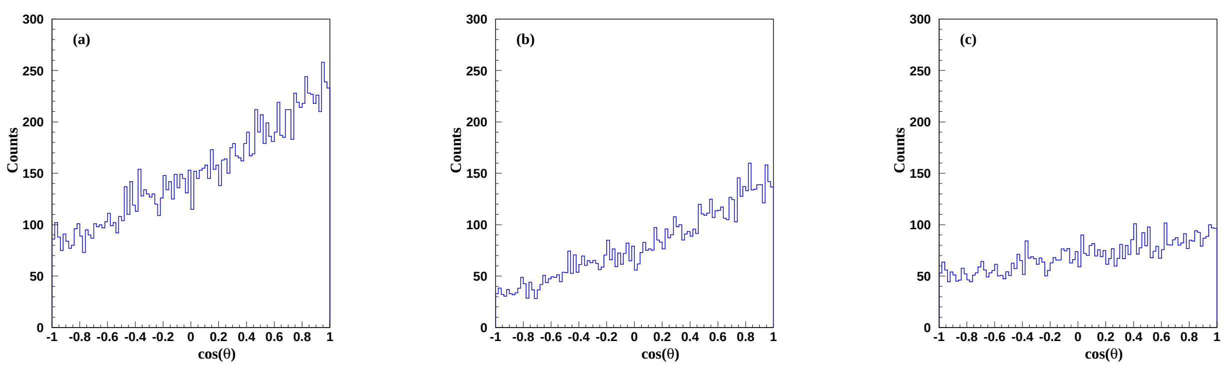

Figure 12: (a) show the $\cos\theta$ distribution for the non separated data set. (b) shows the $\cos\theta$ weighted by the Q-factor and (c) shows the $\cos\theta$ weighted by $1 - Q$.

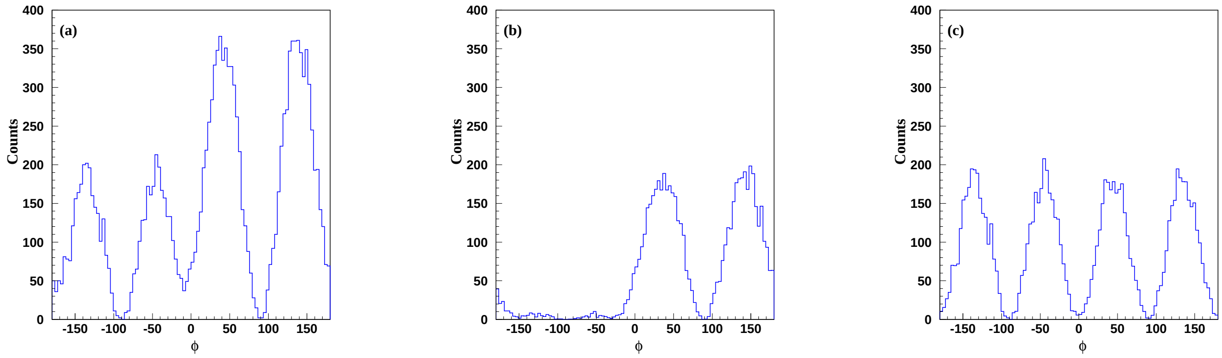

Figure 13: (a) show the ϕ distribution for the separated data set. (b) shows the ϕ weighted by the Q-factor and (c) shows the ϕ weighted by $1 - Q$.

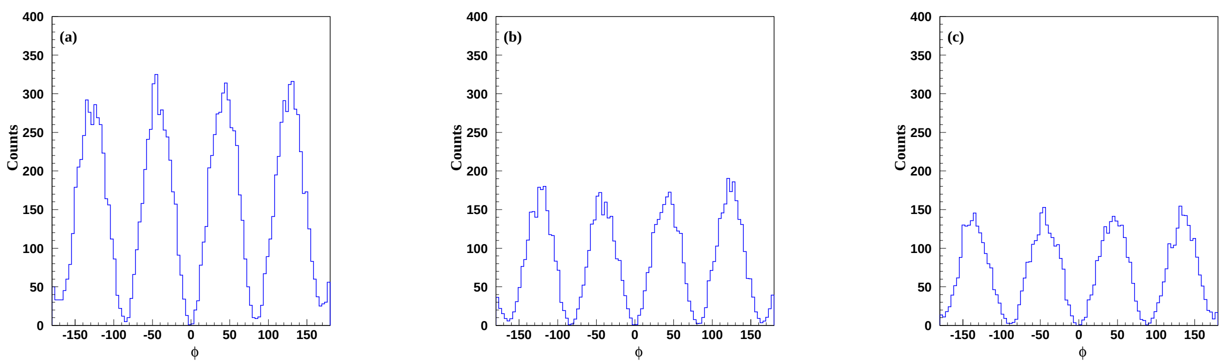

Figure 14: (a) show the ϕ distribution for the non separated data set. (b) shows the ϕ weighted by the Q-factor and (c) shows the ϕ weighted by $1 - Q$.

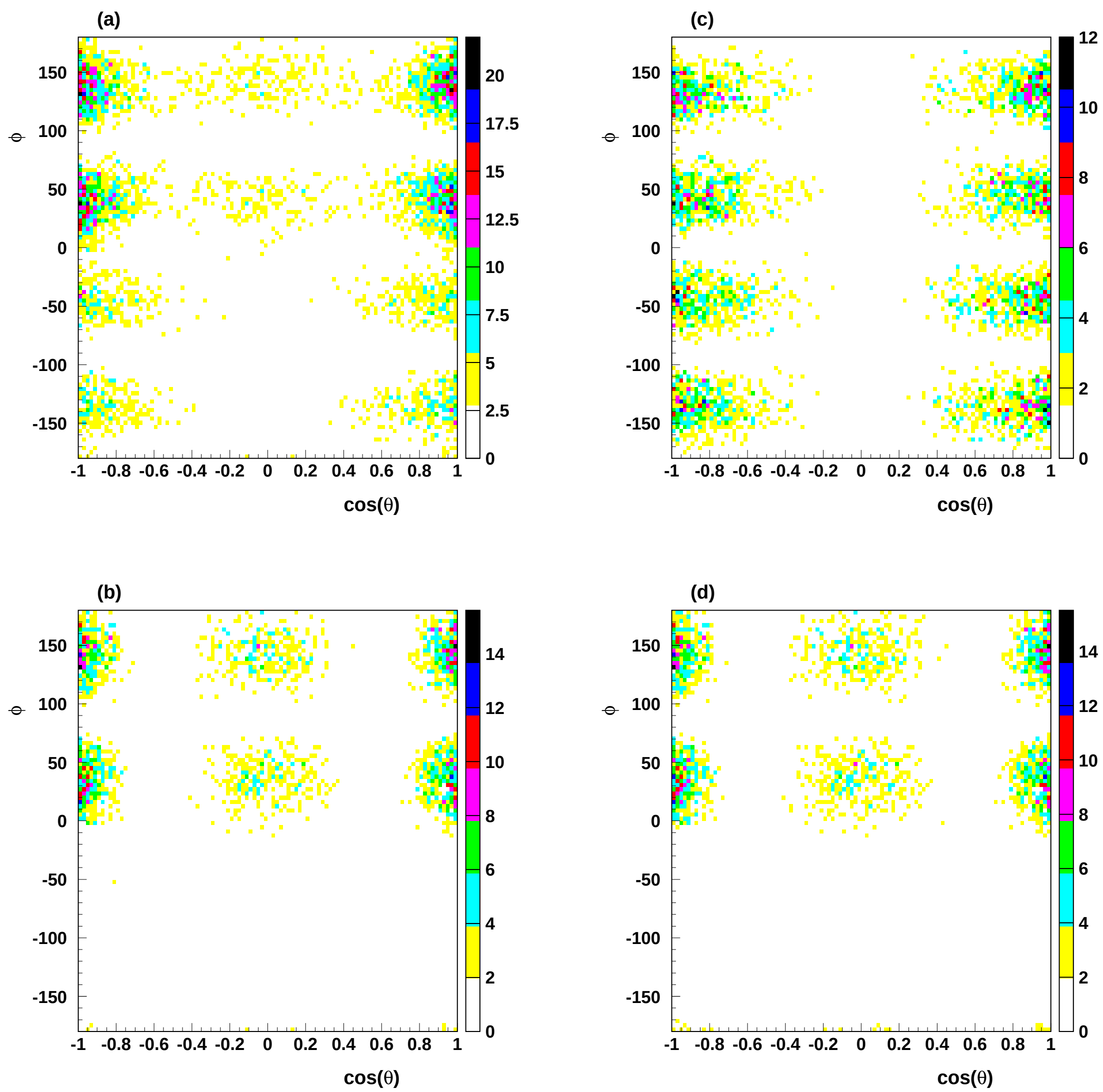

Figure 15: Plots of ϕ versus $\cos\theta$ for the separated data set. (a) shows all the data, (b) shows the data weighted by $1-Q$ (the background). (c) Shows the data weighted by Q (the signal) and (d) shows the true signal distribution.

3.4 Error Estimation

We have applied the error estimation procedure as described in section 2.2 to the two distributions examined here. For each event, we have determined four fit parameters, $\eta_1, \cdots, \eta_4$ and the covariance matrix, C_η for a fit to the signal and background functions,

$$S(m) = \eta_1 \cdot V(m) \tag{22}$$

$$B(m) = \eta_2 + \eta_3(m - m_0) + \eta_3(m - m_0)^2 \tag{23}$$

where $V(m)$ is our voigtian function with centroid m_0. At a particular value of m, we have the Q-factor as

$$Q(m) = \frac{S(m)}{S(m) + B(m)} .$$

In order to get the errors on Q, we need to invert the covariance matrix, C_η, to obtain the error matrix, C_η^{-1}. The error on Q is then computed as:

$$\sigma_Q^2 = \sum_{i,j} \frac{\partial f}{\partial \eta_i} (C_\eta^{-1})_{ij} \frac{\partial f}{\partial \eta_i}$$

For a fixed Voigtian and a quadratic background, the derivatives are straight forward to compute. As an example, we have

$$\begin{aligned}
\frac{\partial Q}{\partial \eta_1} &= \frac{\partial}{\partial \eta_1}\left(\frac{S}{S+B}\right) \\
\frac{\partial Q}{\partial \eta_1} &= \frac{\partial S}{\partial \eta_1}\frac{1}{S+B} - s\frac{\partial S}{\partial \eta_1}\frac{1}{(S+B)^2} - s\frac{\partial B}{\partial \eta_1}\frac{1}{(S+B)^2} \\
\frac{\partial Q}{\partial \eta_1} &= \frac{B}{(S+B)^2}\frac{\partial S}{\partial \eta_1} \\
\frac{\partial Q}{\partial \eta_1} &= \frac{B}{(S+B)^2} V(\xi_r)
\end{aligned}$$

The other derivatives can be computed in a similar fashion to yield that

$$\begin{aligned}
\frac{\partial Q}{\partial \eta_2} &= \frac{-S}{(S+B)^2} \\
\frac{\partial f}{\partial \eta_3} &= \frac{-S}{(S+B)^2}\xi_r \\
\frac{\partial f}{\partial \eta_4} &= \frac{-S}{(S+B)^2}\xi_r^2
\end{aligned}$$

In order to get the total error on a measured quantity, we need to add the statistical error on the signal in quadrature to the to fit error. In Figure 16 are shown the errors on the mass distribution for the separated and non separated samples. Similarly, Figure 17 and 18 shows the errors for the $\cos\theta$ and ϕ distributions of the signals. We believe that the fit errors are an over estimate of the true error, so combining these would also yield an over estimate of the total error.

We note here that if one has a distribution where R becomes close to zero or one, then one needs to be careful in computing the errors. In particular, we have found it better to fit the data as all background or all signal and use the resulting errors from the simpler fits. However, even with this, one can see in Figures 16, 17 and 18 that the fit errors tend to come out larger in the cases where we have separated distributions.

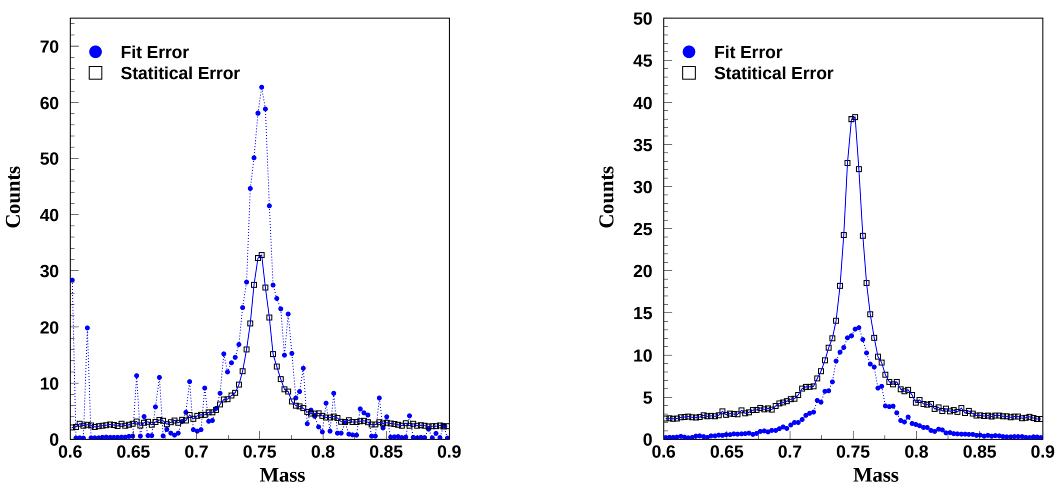

Figure 16: Plots of the computed error (sold blue) and actual error (dashed magenta) for fits to the mass distribution. The left-hand plot shows the separated distribution while the right-hand plot is for the non separated distributions.

4 Summary

In this document we have defined two quantites, the Q-factor and the R-factor which can be used on an event-by-event basis to separate signal from background in an event sample. The Q-factor behaves like the probability of a given event being signal, while the R-factor measures the probability that events near a given event come from the signal distribution.

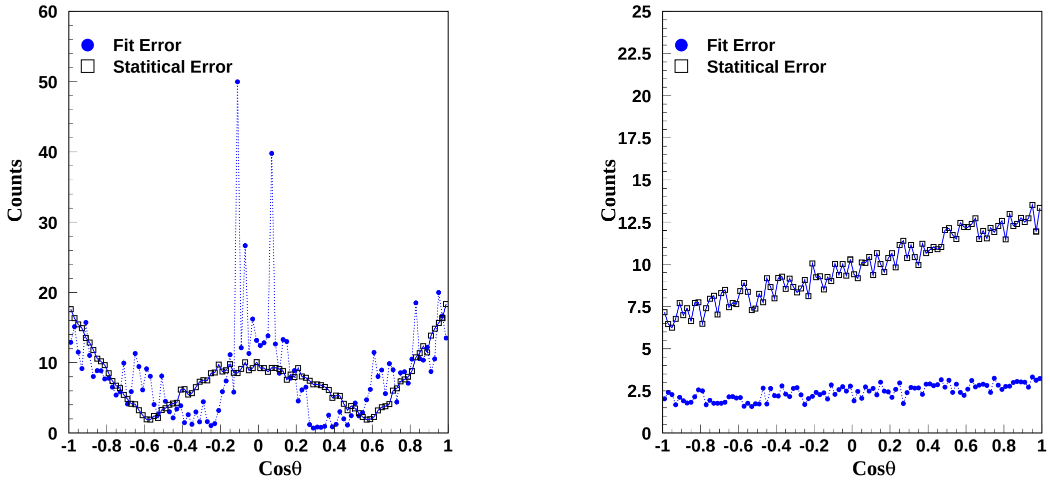

Figure 17: Plots of the computed error (sold blue) and actual error (dashed magenta) for fits to the $\cos\theta$ distribution. The left-hand plot shows the separated distribution while the right-hand plot is for the overlapping distributions.

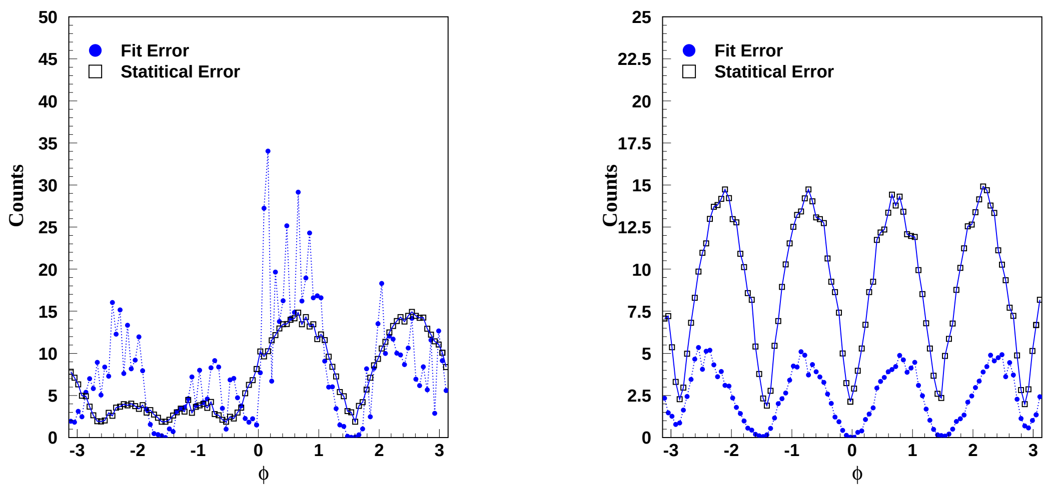

Figure 18: Plots of the computed error (sold blue) and actual error (dashed magenta) for fits to the ϕ distribution. The left-hand plot shows the separated distribution while the right-hand plot is for the overlapping distributions.