sFit Improved: Maximum Likelihood Fitting of Unbinned Data with Weights

Richard T. Jones[1]

[1] *University of Connecticut, Storrs, CT 06269*[*]

(Dated: November 22, 2016)

Abstract

The unbinned maximum-likelihood method provides a very powerful tool for estimating the values of model parameters in fits to experimental data. The method defines a log likelihood function on the data as a sum over all events in the sample, whose maximum in terms of the model parameters provides a statistical estimator for the parameters and their errors. The LHCb collaboration has published results based in part on an extension to the standard unbinned ML method called *sFit*, which allows events in the sample to be assigned weights in order to cancel out background contributions. This paper examines the statistical basis for the expressions employed in *sFit*, and introduces modifications that lead to improved statistical properties and error estimation.

I. OBJECTIVES

The decays of unstable resonances under the strong interaction into n-particle final states with $n \geq 2$ are generally described by multi-parameter probability density functions defined on a space of $3n - 4$ dimensions. The correlations between the degrees of freedom in this space are critical to distinguishing the contributions of competing processes in these models. Ideally one would like to histogram these data in $3n - 4$ dimensions and employ a chi-squared analysis to test the goodness of fit and extract values for the model parameters. However, this proves impractical most of the time because making the bins small enough for the PDFs to be approximately constant over a bin leads to far more bins than there are events in the sample. This is exacerbated by the fact that generally the experimental data have already been binned in two or more dimensions (s, t, mass of parent resonance) prior to the fitting step, which renders prohibitive their further partitioning in multiple dimensions with a resolution high enough to fully describe the details of the model PDFs. The unbinned maximum likelihood method [1] is a standard alternative to chi-squared analysis that leads to statistical estimators for model parameters with many desirable properties (consistency, minimum-variance, Gaussian errors). Although rigorous proof of these properties is only available in the large-N limit, experience shows that they extend to a very good approximation down to sample sizes of practical interest to particle experiments for a broad class of models.

Most of the time, the events that the model is supposed to describe (signal) are mixed together in the experimental sample with events of another origin (background). There are three general means to deal with background in an experiment: (a) eliminate it with cuts, (b) subtract it using weights, and (c) model it together with the signal and fit their sum. In the past, many particle physics experiments have sought to get away with using (a) and (c) alone, and while this may be possible in some special cases, in general the capability to employ all three should be included in the analysis framework for an experiment. Ref. [2] provides an example of the use of both approaches by the LHCb collaboration. The former approach they call *cFit* and the latter *sFit*. The *sFit* method [3] takes the standard unbinned maximum likelihood tool *sPlot* [4] and extends it to incorporate weighted event sampling. The objective of this paper is to critically examine the statistical properties of the *sFit* scheme and compare them with the standard ML approach. A number of improvements to *sFit* are proposed which resolve some problems with its error estimation and improve its statistical errors.

II. EVENT WEIGHTING

Event weights are a standard method for estimating the amount and properties of a signal in the presence of background. For example, suppose one had a sample of K^+K^- events around the mass of the $\phi(1020)$ and one wanted to determine their decay angular distribution in ϕ decays. Suppose the sample contains a sizable admixture of misidentified $\pi^+\pi^-$ events that appear in the invariant mass plot as a background under the ϕ peak. In such a case, one might take a window of width Δm_p that contains the peak, and around that on either side define 'sideband' regions of width $\Delta m_s/2$. To events inside the peak region, one assigns the an event weight $w = 1$, and to regions in either sideband one assigns the weight $w = -\Delta m_p/\Delta m_s$ [5]. Obviously the sum of the weights thus defined gives an estimate of the total number of counts in the peak minus the background extrapolated under the peak, but this is not the real point of the procedure. If a histogram is formed of the decay angles weighted with w_i for each event then one obtains an estimate for the decay distribution of the ϕ alone; the weights effectively cancel out the background contribution.

This familiar sideband subtraction procedure described above is a special case of a more general method that employs event weights to cancel the effects of background present in a sample. This description adopts the notation used in Ref. [3] but seeks to remain as open and general as possible when that treatment is more prescriptive than the method requires. Let the relevant measured variables that define an event be divided into two multidimensional quantities x and y. Like the decay angles in the above example, x and its internal correlations is what the parameterized model seeks to describe, whereas y is like the invariant mass of the K^+K^- pair for each event whose distribution allows the signal and background components to be distinguished.

For sideband subtraction to work, it is essential that the x and y random variables be statistically independent, governed by a product of separate probability density functions (PDFs) $f(x)\,F(y)$ for the signal component and $h(x)\,H(y)$ for the background. However this condition is difficult to ensure under realistic conditions, and it is *not required* for the more general method of event weighting presented in this paper to be valid. Instead of describing the background by a single product $h(x)\,H(y)$, a general form $\sum_j h_j(x)\,H_j(y)$ is adopted for the joint PDF $h(x, y)$. Such a form is sufficient to describe any normalizable joint PDF.

One may also wish to adopt a similarly general expression for the signal distribution $f(x, y)$, and certainly this is allowed. Such a treatment could be called a "multi-signal analysis" which seeks to simultaneously fit several signal components that are deemed to be present in the same sample. This situation is common in particle physics experiments, where multiple resonances show up in the final state as overlapping peaks or structures in a mass plot. However, the joint PDF in such situations cannot be decomposed into a simple sum of products $f_j(x)\,F_j(y)$ because the different channels add coherently and the interfere with one another. Therefore, when multiple interfering resonances are present in the same region of a mass variable, that variable must be included in x, and the joint dependence of the signal on all of the variables in x be considered together in writing down the theoretical model for $f(x)$. If multiple signals are indeed present in a sample, the simplest way to proceed is to conduct separated single-channel analyses for each one, and let the signals that are not the focus of each analysis branch be lumped into the background sum. Hence without any loss of generality, this paper adopts the convention of a single separable signal PDF $f(x)\,F(y)$ and a general non-separable background PDF represented by $\sum_j h_j(x)\,H_j(y)$, where the sum is over $j = 1 \ldots J$ and J represents the number of product terms needed to describe the background.

The functions $f(x)$ and $h_j(x)$ are unknown. The chief advantage of event weighting over other schemes for dealing with background is that it is completely agnostic as to the form of the $h_j(x)$; in fact the functions $h_j(x)$ never appear explicitly anywhere in the formalism. This maintains complete freedom in the analysis results from any assumptions about the $h_j(x)$ apart from the fact that they exist and are normalizable. They are normalized such that the integral of the joint distributions $\int f(x)F(y)\,dx\,dy$ and $\sum_j \int h_j(x)H_j(y)\,dx\,dy$ represent the total yields of signal and background events in the sample, respectively. By contrast, in order to assign weights to the events, something more must be known about the PDFs $F(y)$ and $H_j(y)$. Three distinct cases are considered.

I. A pure signal region A exists in y where $\sum_j \int_A H_j(y)\,dy = 0$ and $\int_A F(y)\,dy$ is large.

II. A pure background region B exists in y where $\int_B F(y)\,dy = 0$ and $\sum_j \int_B H_j(y)\,dy$ is large.

III. Both signal and background distributions overlap over the full y domain of the signal, but they have different distributions in y.

In case I, the way to proceed is to set $w = 1/|A|$, $|A| = \int_A F(y)\,dy$ for all events $y \in A$ and $w = 0$ elsewhere, and proceed with an unweighted ML analysis. This is nothing more than a cut. If the price paid in statistics for this cut is too high, one can also try enlarging the allowed region in y and weighting all events as prescribed for case III below. However, the results should always be compared with those from the analysis with uniform weighting, which is free of any systematics that arise from the presence of the background in the sample.

In case II the simplest way to proceed is to define a signal region A where both signal and background coexist, where A and B are disjoint. One then assigns a uniform weight $w = 1/|A|$ to all events with $y \in A$ and $w = -\sum_j |A'_j|/|B_j|$ for all events with $y \in B$, where $|A| = \int_A F(y)\,dy$, $|B_j| = \int_B H_j(y)\,dy$, and $|A'_j| = \int_A H_j(y)\,dy$. In the case of $J = 1$, this is nothing more than sideband subtraction. This scheme relies only on knowing how to scale the background from region B to estimate its contribution in A, and not on the detailed shapes of the H_j functions. Sometimes, as in the case of experiments with tagged photon beams, time translational invariance allows the ratio $|A'_j|/|B_j|$ to be determined simply from the ratio of widths for intervals A and B. In such cases, one should take B to be as large as possible for the sake of statistical precision, and restrict A to as small a region as possible without giving up too much in signal statistics. If any of the values of $|A'_j|/|B_j|$ are of order unity or larger, this approach will introduce significant statistical error and the procedure for case III below should be attempted. However, the case II procedure is less sensitive to any assumptions about the exact shape of the $H_j(y)$ and so it should always be used as a consistency check on the results from an analysis using a generalized weighting scheme.

Case III is the most general of the three, but it requires the most knowledge of the F and H PDFs. If the assumed forms for $F(y)$ and and $H_j(y)$ are correct, the following weighting procedure provides the optimum statistical precision in the fit results for a given sample size. Normally parameterized forms for F and the H_j can be extracted from empirical fits to histograms of y formed on the unweighted event sample, eg. a double-Gaussian signal peak on top of a polynomial background. Once this is done, each event in the sample should be assigned the following weight based only on its value of y,

$$w_e = \frac{1}{D(y_e)} \left(V_{ss}\, F(y_e) + \sum_j V_{js}\, H_j(y_e) \right) \tag{1}$$

where e labels an individual event in the sample. The constants V_{ss} and V_{js} are the elements of the first column of a square matrix whose inverse is given by

$$V_{ij}^{-1} = \sum_e \frac{G_i(y_e)\, G_j(y_e)}{D(y_e)} \tag{2}$$

where G_i refers either to F when $i = 0$ or H_i when $i = 1 \ldots J$. The sum is over all events in the sample [6]. This is the general form of the weights that lead to unbiased estimators for the values of the function $f(x)$, independent of the amount of background present in the sample. A derivation of this result is given in Appendix A.

The one thing that remains to be specified is the form of the denominators $D(y)$ in Eqs. 1,2. These one is freely able to choose, subject to the requirement that V remains non-singular. It is shown in Appendix A that the form of $D(y)$ that leads to a minimum-variance estimator for the histogram of the signal distribution $f(x)$ is

$$D(y) = N_s\, F(y) + \sum_j N_{bj}\, H_j(y) \tag{3}$$

where N_s and the N_{bj} are the total number of signal and j-type background events in the sample, respectively. Note that the normalization of $D(y)$ cancels out in computing the weights w_e using Eq. 1 so only the relative size of N_s and the N_{bj} matters. Of course, in general the signal/background ratio will vary across the full domain in x, so it will not be possible to be statistically optimal everywhere. However it turns out that the weight function does not depend very strongly on the exact value of this ratio, and near-optimal performance can be often obtained using weights with the values of N_s and N_{jb} set to the approximate total number of events of each type in the sample. In any case, whatever choice is made for $D(y)$, all histograms of x variables weighted according to Eq. 1 are unbiased estimators for $f(x)$, even if their variance is not the absolute minimum possible everywhere in x.

A couple of limiting cases are interesting in this regard. If the value N_s in Eq. 3 is chosen such that $N_s \gg \sum_j N_{bj}$ then the weights in Eq. 1 tend toward a two-value limit similar to case II above: a positive constant close to unity inside the signal region and a negative constant outside. This corresponds for $J = 1$ to the standard sideband subtraction scheme. It agrees with the intuition that sideband subtraction makes sense when the signal/background ratio is significantly larger than one. The other limit $N_s \ll \sum_j N_{bj}$ leads to a weight function which tends to follow the shape of the signal $F(y)$ inside the signal region, rising to greater than one at the maximum, but which then drops off to a negative constant value outside. This amounts effectively to smoothing out the sharp edges of the sideband subtraction window and narrowing the signal region. In this scheme, the signal events near the peak in $F(y)$ get over-counted, while those that fall in the tails of the peak count increasingly toward the background compensation. Even though this may sound counter-intuitive, it actually leads to lower statistical errors in the background-subtracted spectra than the standard sideband subtraction scheme produces.

Whatever choice is made for the shape of $D(y)$, any weighted histogram on the x degrees of freedom is an unbiased estimator for the true parent PDF $f(x)$. The fact that valid results can be obtained using different choices for $D(y)$ is useful for exploring systematic errors associated with the parametrizations taken for $F(y)$ and the $H_j(y)$.

III. STEP-BY-STEP METHOD FOR DEFINING WEIGHTS

In summary, the procedure for defining a proper weighting scheme is as follows.

1. Identify one or more event variables x on which the physics model PDF $f(x)$ depend, and exclude these from the list of potential components to be included in y.

2. Identify one or more event variables y not in the above list which have discriminatory power between signal and background in the event sample. Good candidates for these might be invariant masses of parent or daughter particles, opening angles and energy differences that should be zero for signal but not for background, decay times for detached vertices, and tagging coincidence time differences. These should be statistically uncorrelated with x in that they are independent physical quantities.

3. Use unweighted histograms of the event sample in y to find parameterized forms for the functions $F(y)$ and $H_j(y)$. If y contains more than one variable, consider whether the PDFs factorize into independent 1D functions or whether a joint PDF is needed. Most of the time a product PDF of independent 1D functions should be adequate.

4. Check whether there is any evidence that the x and y variables are correlated. This correlation may come from the experimental acceptance or from the underlying physical processes; it does not matter. If your y variables have been chosen properly, any correlation which exists must be presumed to come from background sources in the sample. Use the correlation to break up the background into several factorizable components, so that the $H(y)$ is split into a set of functions $H_j(y)$ that describe the joint background distribution as $h(x, y) = \sum_j H_j(y)\, h_j(x)$ [7].

5. Select a restricted window in y around the peak in F that contains as much of the total signal strength as possible without making it too wide. If the signal/background ratio in the sample is high, this range can be made somewhat wider than the peak, otherwise making it wider will entail a cost in terms of statistical errors. Summing over sample events within this restricted window, use Eq. 2 to obtain values for V_{ss} and the V_{js} in Eq. 1.

6. Choose values for the constants N_s and N_{bj} in Eq. 1. The method does not specify values for these constants, but minimal statistical errors on the fit results are obtained when they approximately match the total numbers of signal and background events in the sample. They

appear in Eq. 1 as scale factors next to $F(y)$ and $H_j(y)$ respectively, so it is only the relative scales of $N_s F(y)$ and $\sum_j N_{bj} H_j(y)$ within the restricted window chosen in the previous step that matters. This is the only way that the normalization chosen for $F(y)$ and $H_j(y)$ has any consequences for the method.

7. Perform the maximum likelihood fit to the theoretical model using these weights, as described in the next section. Only events within the restricted region of y chosen above in step (5) are used in the fit; all others have zero weight and so will not be counted. The results of the fit include a parameterized form for $f(x)$ with the model parameters adjusted to best describe the signal component within the event sample, ignoring the background.

8. Plot various projections of the x variables as weighted histograms of the sample events. Overlay on these histograms the corresponding projection of the fit function $f(x)$ obtained in the last step. If it is a good fit, they should agree within statistical errors as indicated by a reasonable chi-squared value.

IV. METHOD FOR FITTING WEIGHTED EVENTS

Weighted event analysis is much more general than fitting and subtracting background in measured spectra. Sometimes there are experimental variables that are measured for the specific purpose of being able to use them in y as a basis for eliminating background from a mixed sample of events. A classic example of this is a tagged beam experiments, where a sample of accidental coincidences between the tagger and a detected events is deliberately injected with negative weights into the event sample, and analyzed analyzed together with the in-time events whose weight is normally set to 1. In such cases, the function $w(y)$ is determined by experimental considerations alone, and not by the presumed shape of any background distribution. This is an example of case II in Sec. II above. There is no general fixed rule regarding how events should be weighted. The general guidance given in Sec. III should only be applied when the observed form in the distribution of variables y is *the only clear way* to distinguish signal from background in the sample, the reason being that it introduces systematic errors due to the limited accuracy with which the empirical function $h(x, y)$ can be extracted from the data.

This section provides a step-by-step method for carrying out step (c) using an unbinned maximum likelihood technique in the case where step (b) has already been used to apply non-uniform weights to the sample. A general expression for the negative log likelihood function to be used when fitting a parameterized model PDF $Q(x)$ to a weighted event sample is derived in Appendix B, with the result given by Eq. B.9. Once a weight w_e has been assigned to each event e in the sample as described in the previous section, the following steps may be used to implement the fitting procedure.

1. Form histograms of x weighted by w_e and w_e^2. Form a third histogram to represent $R(x)$ which is computed as the bin-by-bin ratio of the first two. Study the R histogram and decided on an appropriate bin size which captures the variation of the weights over the sample without introducing excessive statistical noise due to small bin size, then go back and make all three again using the revised bin size. If significant statistical fluctuations remain in either of the first two histograms, a smoothing procedure may be applied before taking the ratio.

2. Carry out a minimization search on the function $\mathcal{X}$ defined in Eq. B.9, using the R histogram created in the previous step as a lookup table for evaluating $R(x_e)$ on each event. The normalization integrals, which also involve $R(x)$, may be computed as a sum over reconstructed Monte Carlo events, weighted using the $R(x)$ factor obtained by lookup up its value in the same lookup table.

3. As usual, the values of the model parameters at the minimum of $\mathcal{X}$ are the best-fit values and the inverse of the Hessian matrix of $\mathcal{X}$ at the same point gives their covariances. The 1σ error ellipse in the parameters is given by the contour $\mathcal{X} = \mathcal{X}_{\mathrm{min}} + 0.5$.

V. A CONCRETE EXAMPLE

To illustrate the use of unbinned fitting of weighted events, a simple experimental sample is proposed. In this toy model there is just one x variable denoted θ which lies within the interval $[-\pi, \pi]$, and one y variable denoted E which lies within the interval $[0, 1]$. The PDFs governing the signal in these two random variables are

$$f(\theta) = a_0 + a_1 \sin^2\theta \tag{4}$$

$$F(E) = \exp\left(-\frac{[E-\mu_E]^2}{2\sigma_E^2}\right)(2\pi\sigma_E^2)^{-\frac{1}{2}} \tag{5}$$

The PDFs governing the background in these same two random variables are

$$h(\theta) = b_0 + b_1 \cos^3\theta \tag{6}$$

$$H(E) = \alpha \exp(-\alpha E)\left(1 - e^{-\alpha}\right)^{-1} \tag{7}$$

TABLE I: Values of the constants used to generate Monte Carol model samples for these tests.

constant	value
a_0	0.2
a_1	1.0
μ_E	0.75
σ_E	0.05
b_0	3.0
b_1	3.0
E_0	1.5

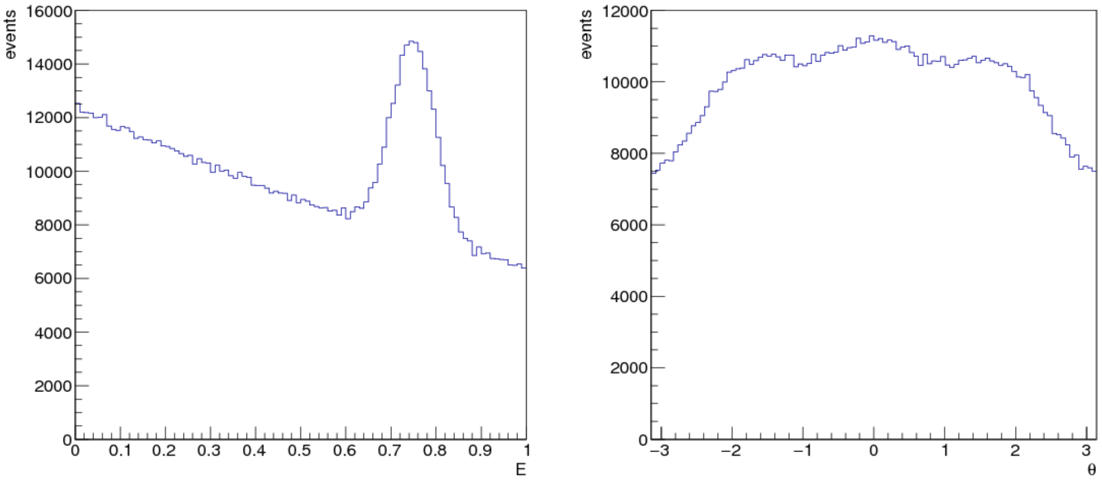

FIG. 1: Raw histograms of the event variables E (left) and θ (right) for a Monte Carlo sample of the toy model consisting of one million events. Note that the minimum in the signal θ distribution at $\theta = 0$ is completely obscured by the presence of the background.

This form has been chosen to depict a generic decay of a resonance with energy E and observed width σ_E sitting on top of a background that is flat in E and distorted in θ. The theoretical model will be taken to be of the form Eq. 4 with parameters a_0 and a_1 to be determined by analysis of the experimental data. The values of the constants that were used to generate Monte Carlo model samples are given in Table. I. Unweighted histograms of a sample of 10^6 events are shown in Fig. 1. The weights for this sample computed using Eq. 1 are shown in the left panel of Fig. 2. Even though the raw θ distribution in Fig. 1 has effectively obscured the angular distribution under the dominant background modulation, the weighted distribution shown in the right panel of Fig. 2 strongly resembles Eq. 4.

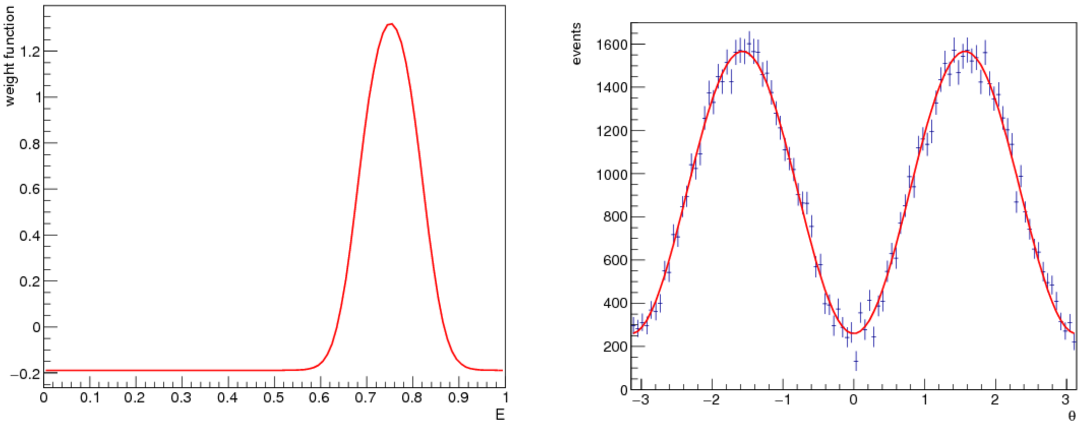

FIG. 2: Event weight function computed on the Monte Carlo sample using Eq. 1 with $N_s/N_b = 0.1$ and $J = 1$. Replotting the right histogram in Fig. 1 using these weights gives the plot on the right in this figure. The red curve in the right-hand plot is a chi-squared fit of the weighted histogram to the generic form of $f(\theta)$ in Eq. 4. The values for the parameters returned from the fit, rescaled to the number of events in the sample, are $a_0 = 0.198 \pm 0.006$ and $a_1 = 1.003 \pm 0.011$, in excellent agreement with the input values in Table I. The chi-squared of the fit is 103 with 98 degrees of freedom.

The results shown in Fig. 2 show that the weighted spectrum of x comes out with the background removed, but the weights were explicitly constructed to do this so the only conclusion one can draw so far is that the Monte Carlo is functioning correctly. Next, the same weighted events are fed into an unbinned maximum likelihood fit using the procedure outlined in the previous section. A sample of one million events was generated using the same Monte Carlo generator, the events were weighted using the same scheme as was used to make Fig. 2, and the $R(x)$ function constructed based on the distribution of these weights. The TMinuit optimization tool within the ROOT framework was used to find the maximum likelihood estimators for parameters a_0 and a_1, together with their fit errors. This was then repeated 10,000 times. The results for the best-fit parameters are shown in Fig. 3. Notice that not only the mean values but also their spread are in agreement with the results of the binned fit shown in Fig. 2. It is a well-known feature of maximum likelihood estimators that they are not zero-bias [1], but the bias revealed in the offset between the means in Fig. 3 and the true Monte Carlo parameter values in Table I is a small fraction of their statistical error reflected in the width of the distributions.

The top row of plots in Fig. 4 shows the same results as residuals, where the true value of the parameter has been subtracted off and the difference normalized to the error returned by the fit. This shows that not only does the weighted likelihood method produce consistent parameter

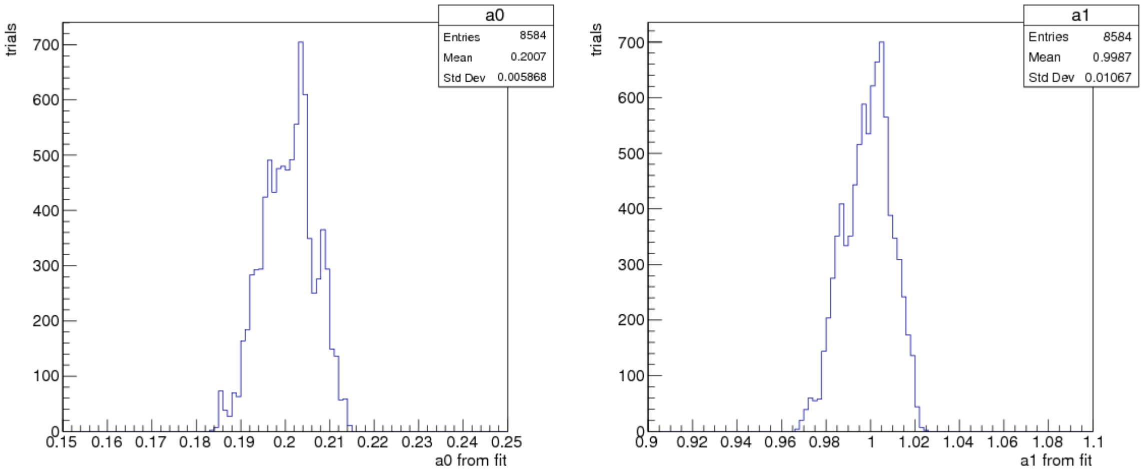

FIG. 3: Best-fit values for signal model parameters a_0 (left) and a_1 (right) if Eq. 4. A total of 10,000 samples of one million events each were generated and fitted with the weighted maximum likelihood method. About 15% of the fits failed to converge within the number of calls to the log likelihood function, so they were omitted from the plots.

values, but also that the error values returned by the fit are consistent. This is in sharp contrast to what is seen using the weighted maximum likelihood method presented in Ref. [3]. The bottom row of plots in Fig. 4 shows what is obtained using that procedure, where the factor $R(x)$ is replaced with 1. Comparison of the top and bottom plots in Fig. 4 shows the impact of using the correct prescription when including weight factors in a maximum likelihood fit. Ref. [2] points out that the errors returned by sFit are too small, and adopts a global rescaling factor on the errors to account for this. However, Fig. 4 shows that an overall scale factor is not an adequate way to deal with this, as the factors needed to rescale the residuals in the bottom row of plots to unit standard deviation are different for a_0 and a_1.

VI. CONCLUSIONS

Background subtraction using an event weighting scheme is an essential tool for particle physics data analysis, in particular for experiments that rely on a tagged beam. However, event weighting is normally considered to be applicable only to a binned data analysis, whereas the analysis of final states with many particles demands the use of an unbinned maximum likelihood approach, due to the sheer number of correlated variables involved. A prescription has been published called sFit [3] for how the same weights that are used in a binned analysis can be applied in a weighted maximum

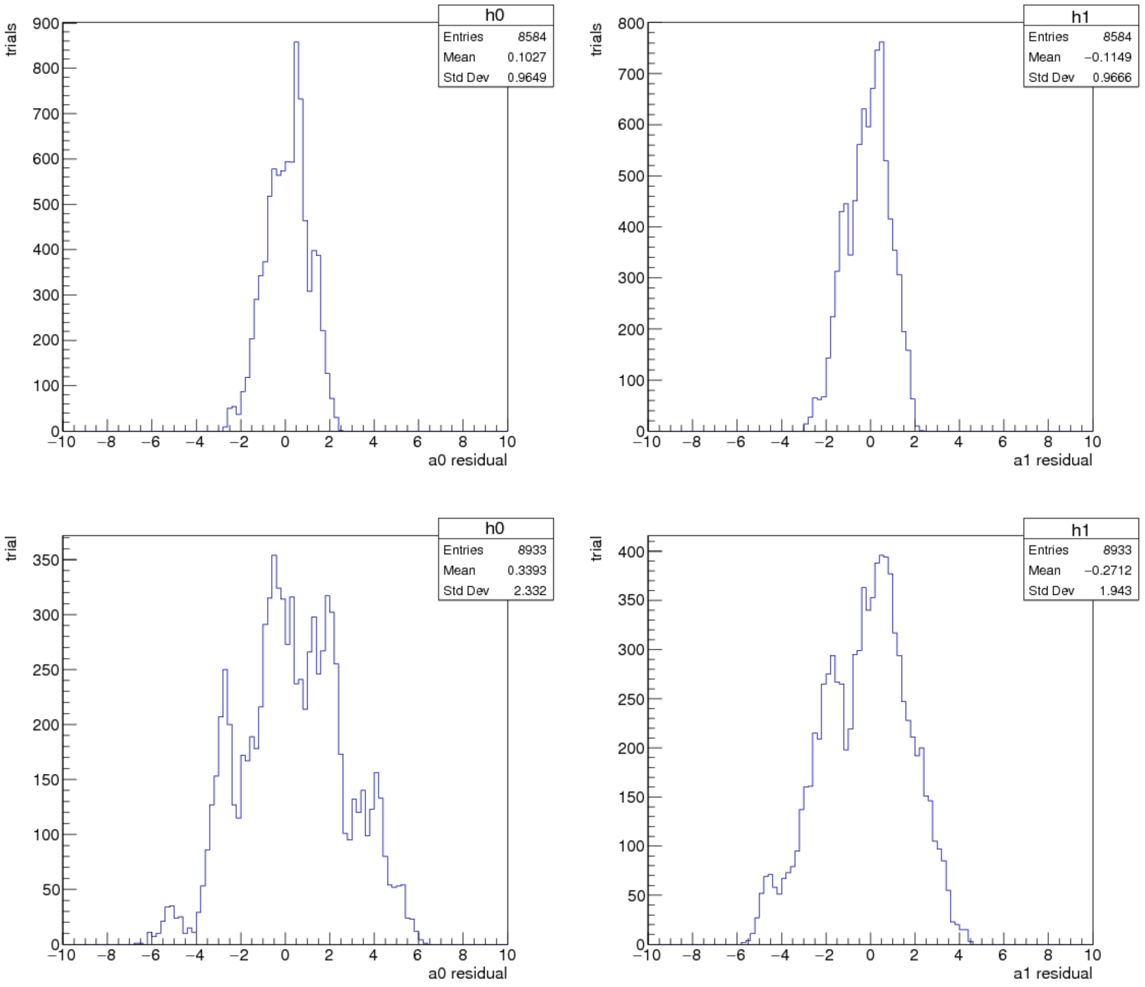

FIG. 4: Residuals on the fitted parameters, defined as the fit value minus the true value divided by the error returned by the fit. These distributions should converge to Normal(0,1) in the limit of large sample size. The top row shows the same results as Fig. 3, but now plotted as a residual to show that not only the fit values but also the fit errors are consistent. The second row shows the results when the same Monte Carlo exercise is repeated using the different maximum likelihood prescription presented in Ref. [3].

likelihood analysis. This prescription has been examined and found not to be valid as implemented in sFit. A modified scheme is proposed which achieves the same goal with limited additional computational effort. A general method for removing background from a sample using weights is presented, together with a procedure for incorporating these weights into a maximum likelihood fitting algorithm. Proofs are given that the proposed weights have the usual properties (generate histograms that are minimum-variance, unbiased estimators of the true parent distributions) in a binned analysis, and that the derived likelihood function of the unbinned event variables with weights converges to the binned likelihood with the same weights in the large-N limit.

* Electronic address: richard.t.jones@uconn.edu

[1] W.T. Eadie, D. Drijard, F.E. James, M. Roos, B. Sadoulet, *Statistical Methods in Experimental Physics*, North-Holland, Amsterdam (1971).

[2] R. Aaij et al. (LHCb Collaboration), "Observation of J/ψ resonances consistent with pentaquark states in $\Lambda_b^0 \to J/\psi K p$ decays", Phys. Rev. Lett. 115, 072001 (2015).

[3] Yuehong Xie, "sFit: a method for background subtraction in maximum likelihood fit", arXiv:0905.0724v1 (2009).

[4] M. Pivk and F. R. Le Diberder, "sPlot: A statistical tool to unfold data distributions", Nucl. Instrum. Meth. A555 (2005) 356.

[5] This assumes that the background distribution under the peak is flat in the mass plot, otherwise an adjustment to the sideband weight would be necessary to make the subtraction complete.

[6] Note that Eq. 2 differs from Eq. 5 in Ref. [3] in that the latter has the factor $D(y_e)$ squared in the denominator of the summand. This seems to be a typographical error in that paper.

[7] If correlations persist between x and y that cannot be accounted for in this way, one or more of the y component variables must be moved over to the x side and incorporated into $f(x)$. Either the correlation comes from the experimental acceptance (acceptance is a part of $f(x)$) or it must be incorporated into the physics model; either way rules out the use of this variable as a component of y in a weighted sample analysis.

Appendix A: Minimum-variance weighting

The goal of event weighting is to define a signal filter through which, when the events of a mixed sample are passed, the background is blocked and the signal passes with 100% strength. The idea is that in addition to a primary set of event variables denoted by compound quantity x whose probability distribution the experiment is designed to determine, there is an auxiliary set of variables denoted y which contain information that can distinguish signal events from background on a statistical basis. The distributions of signal and background in y are overlapping, otherwise one could simply use a cut to eject the background from the sample. Ordinarily when making a histogram in x, one would project out y, but with weighted sampling one constructs a factor $w(y)$ for each event and increments the given x bin with the value $w(y)$, instead of 1, each time an event falls in its corresponding x interval. This operation defines a linear filter that maps a y distribution onto a number. For finite samples, there are statistical fluctuations about this mean value, but the mean value is the same for samples of all sizes. Just three criteria are sufficient to uniquely define the coefficients to be used in this filter.

(i) For every signal event entering the filter at a given x, on average one count emerges from the filter at this same x.

(ii) For event background event entering the filter at a given x, on average zero counts emerge from the filter.

(iii) The statistical fluctuations in the output be the smallest possible for any choice of linear filter.

To solve for the filter coefficients $w(y)$, the expression given by Eq. A.1 is taken as a starting point for the global sample PDF,

$$g(x,y) = f(x)F(y) + \sum_j h_j(x)H_j(y) \tag{A.1}$$

The normalization of $g(x,y)$ is taken to equal the total number of events in the sample, both signal and background. In what follows, $F(y)$ will be assumed to be normalized to unity, but the normalization of $H_j(y)$ will depend on how the domain in y is defined, and so will be left unspecified. Conditions (i) and (ii) above reduce to the following relations,

$$\begin{aligned} \int w(y)\, F(y)\, dy &= 1 \\ \int w(y)\, H_j(y)\, dy &= 0\ \forall j = 1 \ldots J \end{aligned} \tag{A.2}$$

Subject to these conditions, in the limit of large sample size, the sum of weights w for all events (x_e, y_e) with x_e within the interval $[x, x + dx]$ converges to $f(x)\,dx$ independent of the presence or amount of background. At finite sample size, this remains the mean value for the weight sum in each bin, but now the question arises of what the size of the statistical fluctuations might be. Here simple Poisson statistics provides the answer. Defining s_x as the sum of sample weights in the interval $[x, x + dx]$, the variance on this quantity is given by Eq. A.3.

$$V(s_x) = \int w^2(y)\, g(x, y)\, dy \tag{A.3}$$

The filter affects not just this bin at x but all bin across the complete x space, so it is not just this $V(s_x)$ that should be minimized, but its sum over all bins. This leads directly to a global optimization functional $\mathcal{F}$ of the weight coefficients $w(y)$,

$$\begin{aligned}\mathcal{F} \;=\; & \int w^2(y)\, g(x, y)\, dx\, dy - \\ & 2\mu_0 \left(\int w(y)\, F(y)\, dy - 1 \right) - \\ & 2 \sum_{j=1}^{J} \mu_j \int w(y)\, H_j(y)\, dy \end{aligned} \tag{A.4}$$

where Lagrange multipliers μ_j have been introduced to enforce the constraints Eq. A.2. The standard Euler-Lagrange equations yield a unique extremum for $\mathcal{F}$ when the function $w(y)$ satisfies the relation

$$w(y)D(y) - \mu_0 F(y) - \sum_{j=1}^{J} \mu_j H_j(y) = 0 \tag{A.5}$$

where $D(y) = \int g(x, y)\, dx = N_s F(y) + \sum_j N_{bj}\, H_j(y)$. Substitution of the solution to $w(y)$ given by Eq. A.5 into Eq. A.2 gives coupled linear equations for μ and λ.

$$\begin{bmatrix} \int \frac{F^2(y)}{D(y)}\, dy & \int \frac{F(y)\, H_1(y)}{D(y)}\, dy & \cdots \\ \int \frac{F(y)\, H_1(y)}{D(y)}\, dy & \int \frac{H_1^2(y)}{D(y)}\, dy & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix} \begin{pmatrix} \mu_0 \\ \mu_1 \\ \vdots \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ \vdots \end{pmatrix} \tag{A.6}$$

Inverting Eq. A.6 leads to unique values for the μ_i, which completes the solution for the minimum-variance weights.

$$w(y) = \frac{\mu_0\, F(y) + \sum_{j=1} \mu_j\, H_j(y)}{N_s\, F(y) + \sum_{j=1} N_{bj}\, H_j(y)} \tag{A.7}$$

Appendix B: Maximum likelihood with weights

Standard method with uniform weighting

The most straight-forward way to derive the correct expression to use in the extraction of Bayesian maximum-likelihood estimators for model parameters from an unbinned sample of events indexed by $e = 1 \ldots N$ is to start with the classical expression for the probability that the N independent events that were measured in the experiment would have emerged in the order they were seen from a parent distribution governed by an (a priori unknown) set of model parameters denoted collectively by the variable a.

$$\mathcal{P}(\{x_e\}|N, a) = \prod_{e=1}^{N} P_1(x_e|a) \tag{B.1}$$

assuming that the experiment ran until N events were collected and then stopped. It is customary to suppose that instead the experiment ran for a fixed time period instead of a fixed number of events, in which case the expression the net probability of the experimental observations becomes

$$\mathcal{P}(\{x_e\}, N|a) = e^{-\Lambda} \frac{\Lambda^N}{N!} \prod_{e=1}^{N} P_1(x_e|a) \tag{B.2}$$

where Λ is the expected total number of expected events in the experiment based on the model parameters a. In the spirit of Bayesian parameter estimation, one switches from viewing the $\mathcal{P}$ in Eq. B.2 from a probability of the observations given fixed values for the model parameters to viewing it as a likelihood function of the model parameters for fixed values of the experimental data the $\{x_e\}$ and N. Finding values of the model parameters at which the likelihood achieves its maximum value is the same as seeking the minimum of the function $\mathcal{X}$ defined as

$$\mathcal{X} = \int Q(x)\, A(x)\, dx - \sum_{e=1}^{N} \log(Q(x_e)) + \mathcal{X}_0 \tag{B.3}$$

where $P_1(x)$ has been refactored into the product of the model function $Q(x)$, which contains all of the dependence on the a parameters, and the experimental acceptance $A(x)$. Several additive terms that do not depend on a have been lumped into the irrelevant constant offset $\mathcal{X}_0$, which gets quietly ignored. The first term on the right-hand side in Eq. B.3 is called the "normalization integral" for $Q(x)$. The dependence of this term on the model parameters a is readily estimated using Monte Carlo simulation. It is the second term that contains all of the information from the experimental data regarding the shape and magnitude of Q.

Generalization to non-uniform weights

The previous section ignored any mention of event weights, implicitly assuming that all of the weights were 1. Following the same line of reasoning as Sec. B-1, it is tempting to suppose that Eq. B.1 can be adapted to the non-uniform weighting case by simply raising the $P_1(x_e|a)$ to the power of the event weight w_e. In fact, this is exactly what is assumed in Ref. [3] (see Eq. 5 in that article). This prescription has the intuitively attractive feature that background events in the sample enter the product as the reciprocals of the corresponding signal events at the same x, effectively canceling them out as they should. However, this does not prove to be the statistically correct expression, and it leads to faulty results when it is translated into a maximum-likelihood prescription for model parameter estimation, as demonstrated in Ref. [3].

The correct expression (derived below) for the likelihood on a weighted sample is very similar in form to Eq. 5 of Ref. [3], only the exponents to be used on the individual P_1 factors in the likelihood product are not simply the weight factors $w(y)$ derived in Appendix A, but a closely related quantity denoted $u(x, y)$. Although $u(x, y)$ which is not the same as $w(y)$, it can be computed based on $w(y)$ and how it varies across the sample.

The derivation begins by noting how weights are first introduced in Appendix A, as a means for generating unbiased estimates for the pure signal PDF $f(x)$ out of histograms of mixed-sample data. A bridge must be constructed between a chi-squared function of binned data in the large-N limit and the negative log likelihood function that appears in the equivalent unbinned ML analysis. If this correspondence can be carried out with weights then the correct expression for the likelihood on a weighted sample will be immediately apparent.

Constructing this bridge is mathematically subtle, as can be seen by comparing the equivalent but surprisingly different-looking expressions for binned and unbinned likelihoods in the case of uniform weighting. On the unbinned side, one has Eq. B.2. The equivalent expression in terms of a binned analysis is given by

$$\mathcal{P}(\{n_b\}|a) = \prod_{b=1}^{B} \exp\left(-\frac{[n_b - \lambda_b(a)]^2}{2\sigma_b^2}\right) \tag{B.4}$$

where the sum over bin index b extends from 1 to the total number B of bins. The theoretical model factors $\lambda_b(a)$ are given in terms of the PDFs introduced in Appendix A as $\lambda_b(a) = f(x_b)\,dx$ with the $f(x)$ PDF implicitly depending on the model parameters a. To build a bridge to Eq. B.2, one takes the negative log of Eq. B.4 and lumps additive terms that do not depend on a into an irrelevant offset $\mathcal{X}_0$.

$$\mathcal{X} = \frac{1}{2}\sum_{b=1}^{B} \frac{\lambda_b^2}{\sigma_b^2} \; - \; \sum_{b=1}^{B} \frac{n_b\lambda_b}{\sigma_b^2} \; + \; \mathcal{X}_0 \tag{B.5}$$

The next step is subtle, as there does not seem to be any obvious way to get a sum of logs from this expression. The key to the next step is to see that the two expressions for $\mathcal{X}$ do not need to agree where they are far from the minimum because there the likelihoods themselves can be very close together near zero even while their logs remain widely separated in the region of large negative log values. The only region where the likelihood functions need to be compared is when $\mathcal{X}$ is in the vicinity of its minimum where the likelihood is significantly different from zero, which corresponds to the values of n_b in each bin being a limited number of σ_b from the mean value λ_b. The exact value of the likelihood cutoff does not matter, whether one places it at 10^{-10} or 10^{-50}, the fact remains that the two expressions need only agree over a limited range in the difference $1 - n_b/\lambda_b$. In this spirit, one defines an expansion parameter $z_b = 1 - n_b/\lambda_b$ and considers that in the large-N limit only values of $|z_b| \ll 1$ will matter when computing the sum in the first term of Eq. B.5.

$$\begin{aligned} \mathcal{X} \; &= \; \frac{1}{2}\sum_{b=1}^{B} \frac{n_b\lambda_b}{\sigma_b^2(1-z_b)} \; - \; \sum_{b=1}^{B} \frac{n_b\lambda_b}{\sigma_b^2} \; + \; \mathcal{X}_0 \\ &= \; \frac{1}{2}\sum_{b=1}^{B} \frac{n_b\lambda_b}{\sigma_b^2}[1 + z_b + z_b^2 + \ldots] \; - \; \sum_{b=1}^{B} \frac{n_b\lambda_b}{\sigma_b^2} \; + \; \mathcal{X}_0 \end{aligned} \tag{B.6}$$

For unweighted events $\sigma_b^2 = n_b$. The expression in square brackets in Eq. B.6 is closely related to the Taylor expansion of the log function about 1.

$$\log(1-z) = -z - \frac{1}{2}z^2 - \frac{1}{3}z^3 - \ldots$$

Still more strongly convergent for small z is the expression,

$$z + (1-z)\log(1-z) = \frac{1}{2}z^2 + \frac{1}{6}z^3 - \ldots$$

Discarding all terms cubic and higher in z in $\mathcal{X}$ leads to

$$\begin{aligned} \mathcal{X} \; &\simeq \; \frac{1}{2}\sum_{b=1}^{B} \lambda_b[1 + 3z_b + 2(1-z_b)\log(1-z)] \; - \; \sum_{b=1}^{B} \lambda_b \; + \; \mathcal{X}_0 \\ &= \; \sum_{b=1}^{B} \lambda_b \; + \; \sum_{b=1}^{B} n_b \log(1/\lambda_b) \; + \; \mathcal{X}_1 \\ &= \; \int f(x)\, dx \; - \; \sum_{b=1}^{B} n_b \log(\lambda_b) \; + \; \mathcal{X}_1 \\ &= \; \int Q(x)\, A(x)\, dx \; - \; \sum_{e=1}^{N} \log(Q(x_e) \; + \; \mathcal{X}_2 \end{aligned} \tag{B.7}$$

In the last step of Eq. B.7, the replacement $f(x) = Q(x)A(x)$ was made, followed by discarding an extra additive constant proportional to $\log(A(x_b))$. A sum over bins with a factor n_b is also replaced with the equivalent sum over events in the sample, thus completing the bridge via Eq. B.3 from Eq. B.4 to Eq. B.2.

The remaining task is to retrace these same steps for the case of non-uniform weights. In this case, the argument proceeds in a very similar fashion, except that the expression for the chi-squared likelihood is modified to included weighted values for the mean bin content ν_b and its variance τ_b^2.

$$\begin{aligned} \mathcal{P}(\{n_b\}|a) &= \prod_{b=1}^{B} \exp\left(-\frac{[\nu_b - \lambda_b(a)]^2}{2\tau_b^2}\right) \quad \text{where} \\ \nu_b &= \sum_{e=1}^{n_b} w_e = n_b \langle w \rangle_b \\ \tau_b &= \sum_{e=1}^{n_b} w_e^2 = n_b \langle w^2 \rangle_b \end{aligned} \tag{B.8}$$

From here, the argument follows the same track as for the uniform weighting case, and all expressions remain nearly the same except that the ratio n_b/σ_b^2 that canceled inside the sums on the right-hand side of Eq. B.6 becomes instead $\nu_b/\tau_b^2 = \langle w \rangle_b \,/\, \langle w^2 \rangle_b$, leading to a modified form of Eq. B.7,

$$\mathcal{X} = \int R(x)\, Q(x)\, A(x)\, dx \quad - \quad \sum_{e=1}^{N} w_e\, R(x_e) \;\log(Q(x_e) \quad + \quad \mathcal{X}_1 \tag{B.9}$$

where $R(x) = \langle w \rangle \,/\, \langle w^2 \rangle$ for events with x_e in the immediate neighborhood of point x. From this negative log likelihood, one can read off the master likelihood function that gives rise to it,

$$\begin{aligned} \mathcal{P}(\{x_e\}, N|a) &= e^{-\Lambda} \frac{\Lambda^M}{M!} \prod_{e=1}^{N} P_1(x_e|a)^{u_e} \quad \text{where} \\ \Lambda &= \int R(x)\, Q(x)\, A(x)\, dx \\ u_e &= w(y_e)\, R(x_e) \\ M &= \sum_{e=1}^{N} u_e \quad . \end{aligned} \tag{B.10}$$

This is the correct statistical expression for the likelihood function, rather than Eq. 5 of [3], in the case of a non-uniform weighted sample. The function $R(x)$ can be constructed directly from the event sample by forming histograms of w_e and w_e^2 and then taking their ratio.