Crystal Barrel Note 192

Fits to Unbinned Data

Kay Königsmann

Introduction

To determine the intermediate resonance structure in final states produced for example in $\bar{p}p$ reactions we have employed so far the standard χ^2 fits to binned Dalitz plots. This method is not very practical for decays to more than three particles. For example, a fit to the five-dimensional phase space for 4 particles would require some 10^5 bins, with most of the bins empty. Also many of the bins would be at the border of the phase space and would thus need to be excluded from the fit.

The solution to this dilemma is to fit to individual events instead of binned histograms. This is done with the standard likelihood technique, which is, for example, described in the book by W.T. Eadie et al., North Holland 1971, or by A.G. Frodesen et al., Columbia Univ. 1979. For a novice in this field I recommend the book by L. Lyons, Cambridge Univ. 1986. In the following I will list the relevant formulae and discuss three methods, of which one is not easily found in the literature. For a discussion on how to include background in likelihood fits, see the recent CBar note by C. Amsler.

The Standard Likelihood Method

The standard definition of a likelihood function is

$$\mathcal{L} = \prod_{i=1}^{\mathrm{N}} \mu_i \tag{1}$$

where N is the total number of observed events and μ is the probability density function (PDF) we want to fit to. In a fit to a Dalitz plot the PDF contains complex amplitudes with interfere, $\mu = |\sum_k \alpha_k A_k|^2$. The problem arises from the fact, that such a PDF is *not* properly normalized, i.e. the integral over all phase space $\int \mu \, \mathrm{d}\Omega$ is not 1. But the PDF can easily be normalized and we define

$$\mathcal{L} = \prod_{i=1}^{\mathrm{N}} \frac{\mu_i}{\int \mu \, \mathrm{d}\Omega} \tag{2}$$

as our standard likelihood function, which has to be maximized. Since the product consists of many small numbers, it is common to take the negative of the logarithm,

$$-\ln \mathcal{L} = \mathrm{N} \ln \left(\int \mu \, \mathrm{d}\Omega \right) - \sum_{i=1}^{\mathrm{N}} \ln \mu_i \ , \tag{3}$$

which is to be minimized (typically with minimization packages like MINUIT). I will show some easy examples below.

The Extended Likelihood Method

The standard likelihood method is typically used in fits to *shapes* of distributions, where the absolute normalization is fixed to the number of observed events (see also the discussion at the end of the next section). It is, however, possible to also let the fit provide an estimate on the absolute normalization. This can be achieved with the extended likelihood function, which is just the product of the standard likelihood function with an extra factor to account for the probability to obtain an event sample size N. This factor could be

a Gaussian or Poisson distribution. For example, the extended likelihood function would be written as

$$\mathcal{L} = \frac{\mathrm{e}^{-\phi}\phi^{\mathrm{N}}}{\mathrm{N}!} \prod_{i=1}^{\mathrm{N}} \frac{\mu_i}{\int \mu \, \dot{\Omega}} \ , \tag{4}$$

where $\phi = \int \mu \, \dot{\Omega}$ is the normalization which is allowed to vary according to Poisson statistics around the measured number of events N. For Gaussian statistics this factor would be $\exp(-(\phi - \mathrm{N})^2/2\mathrm{N}$.

The Generalized Likelihood Method

Another way to define a likelihood function is e.g.

$$\mathcal{L} = \exp(-\int \mu \, \dot{\Omega}) \ \prod_{i=1}^{\mathrm{N}} \mu_i \ , \tag{5}$$

which yields the following function to be minimized

$$-\ln \mathcal{L} = \int \mu \, \dot{\Omega} - \sum_{i=1}^{\mathrm{N}} \ln \mu_i \ . \tag{6}$$

The difference with the 'standard' method is obvious. A justification for this ansatz can be derived from Poisson statistics. Take the likelihood function used for a fit to a binned histogram which has small statistics, $\mathcal{L} = \prod_{k=1}^{\mathrm{K}} \mathrm{e}^{-\mu_k} \mu_k^n / n!$ and decrease the bin size to zero and the number K of bins to infinity. Then n (the number of events in bin k) will be either zero or one. Therefore the likelihood can be written as $\mathcal{L} = \exp(-\sum_k \mu_k) \prod_k \mu_k$, which is the desired form. In this formulation the PDF μ does *not* need to be normalized; it is the integral, which takes care of the normalization. A half page discussion of the generalized likelihood function can be found in the book by A.G. Frodesen et al.

What is the difference between the standard and the generalized likelihood functions? Using the standard likelihood function will force the normalization to be strictly correct, i.e. $\int \mu \, \dot{\Omega} = \mathrm{N}$, where N is the total number of

events. On the other hand, in the generalized maximum likelihood method this is not necessarily true. Thus the generalized likelihood method could give different estimates of parameters which, in principle, should be better. This is similar to the use of multinomial or Poisson statistics in fits to binned histograms. In the first case the total number of fitted events is rigorously fixed to those measured, whereas in the latter case the number of fitted events will fluctuate around the number of measured events according to Poisson statistics. Note that the generalized likelihood method was derived from Poisson statistics and thus will show the same behavior regarding the number of fitted events.

Tests of the two Likelihood Functions

I have generated 200 events randomly distributed between 0 and 10. Added to this distribution were two Gaussians centered at 2.5 and 7.5 with widths of 0.2 and 150 events each, see Fig.1a. This distribution was then fitted to the function

$$\mu(x) = \frac{A}{\int \dot{x}} + \frac{Bx}{\int x\dot{x}} + \frac{C\,\mathrm{e}^{-(x-\nu_1)^2/2\sigma_1^2}}{\sqrt{2\pi}\sigma_1} + \frac{D\,\mathrm{e}^{-(x-\nu_2)^2/2\sigma_2^2}}{\sqrt{2\pi}\sigma_2} \tag{7}$$

using the three likelihood methods discussed above.

Note that all terms are individually normalized, as needs to be done for the amplitudes used in fits to Dalitz plots. The overall normalization $\int \mu(x)\,\dot{x}$ is in this case simply $A+B+C+D$. However, such a normalization cannot be evaluated analytically in the case of fits to Dalitz plots. Therefore I have also included in the program the possibility to evaluate the normalization integral with Monte Carlo methods. Note that in this case the sum over weights needs to be multiplied with the total phase space volume Ω and divided by the number of dialed Monte Carlo points M. This multiplicative factor Ω/M is not crucial when fitting with the standard likelihood technique,

Figure 1: The left histogram shows the data used in log-likelihood fits to the unbinned data. The right histogram shows the result of a fit using the generalized likelihood method.

as it contributes only an additive constant to the log-likelihood function:

$$-\ln\mathcal{L} = \mathrm{N}\ln\left(\int \mu\,\dot{\Omega}\right) - \sum_{i=1}^{\mathrm{N}} \ln\mu_i \tag{8}$$

$$= \mathrm{N}\ln\left(\sum_{j}^{\mathrm{M}} \mu_j\right) + \mathrm{N}\ln(\Omega/\mathrm{M}) - \sum_{i=1}^{\mathrm{N}} \ln\mu_i \ . \tag{9}$$

However, it is crucial when using the generalized or extended likelihood methods. In the first case the factor is multiplicative in the log-likelihood and thus cannot be omitted. In the latter case the normalization is needed in the Gaussian factor $\exp(-(\int \mu\,\dot{\Omega} - \mathrm{N})^2/2\mathrm{N})$ in the likelihood.

Results of Fits

The results from the fits to the unbinned data as shown in Fig. 1a are summarized in Table 1. First of all it is evident, that the standard log-likelihood fit yields by design exactly the total number of events fitted. Note that in this case D was calculated in the program as $D = \mathrm{N} - A - B - C$ and thus has no error. Both, the extended and the generalized log-likelihood fits yield a normalization of 491 events, short of the 500 events to be fitted. Using in the latter fit the analytical form of the normalization yields 499 events. This shows the dependence of the fit on the Monte Carlo sampling for the total phase space. Note that I used only 1000 Monte Carlo events, just twice as many as data events. In fact, using 10 times as many Monte Carlo events yield fit results very similar to the ones obtained with the analytical normalization. As an example I show in Fig. 1b the result of the fit with the generalized likelihood method.

For a second example I generated 200 events distributed flatly between 0 and 10. On top were generated 300 events following an exponential decay with decay time $t = 1$. Data and fit result (using the generalized likelihood method) are shown in Fig. 2. The fit result was: $y = (205 \pm 20) + (300 \pm 28) \times \exp(-(0.95 \pm 0.09)t)$.

Figure 2: The left histogram shows the data used in log-likelihood fits to the unbinned data. The right histogram shows the result of a fit using the generalized likelihood method.

Table 1: Results from fits to the unbinned data shown in Fig. 1a. The different fits are: SLL = standard log-likelihood fit (normalization calculated with Monte Carlo); ELL = extended log-likelihood fit (normalization calculated with Monte Carlo); GLL1 = generalized log-likelihood fit (normalization calculated with Monte Carlo); GLL2 = generalized log-likelihood fit (normalization calculated analytically).

Parameter/Fit	SLL	ELL	GLL1	GLL2
A	173.8 ± 31.3	170.8 ± 31.6	170.9 ± 31.4	152.7 ± 30.2
B	27.1 ± 28.3	26.6 ± 27.5	26.6 ± 27.4	44.4 ± 27.3
C	149.2 ± 11.7	146.6 ± 13.1	146.6 ± 13.1	156.0 ± 13.7
σ_1	0.22 ± 0.02	0.22 ± 0.02	0.22 ± 0.02	0.23 ± 0.02
ν_1	2.47 ± 0.02	2.47 ± 0.02	2.47 ± 0.02	2.50 ± 0.02
D	$149.9 \pm 0.$	147.3 ± 13.5	147.3 ± 13.4	146.8 ± 13.4
σ_2	0.22 ± 0.02	0.22 ± 0.02	0.22 ± 0.02	0.19 ± 0.02
ν_2	7.46 ± 0.02	7.46 ± 0.02	7.46 ± 0.02	7.47 ± 0.02
$A + B + C + D$	500.0	491.3	491.4	499.1

In summary, all three likelihood methods work fine for unbinned data. The question arises which one to choose. If the calculation of the total phase space volume is no problem, I would suggest the generalized likelihood method, which is the easiest and fastest method (fewer calculations of logarithms). This method yields the same fit results as the extended likelihood method. In case the total phase space volume is difficult to calculate, then the standard likelihood method is appropriate. However, this method requires to constrain all amplitudes to add to the total number of events. I did this by calculating $D = \mathrm{N} - A - B - C$ and therefore get no error estimate on D. Note that without this constraint the fit does not converge at all! I also did some timing tests for all fits and found that the standard likelihood method is the fastest: 13.6 (CERN accounting) seconds. This is due to the

constraint N = $A + B + C + D$, which requires one less parameter to be minimized. All other fits took about 16.2 sec. However, if I introduce in the generalized likelihood fit this constraint as well, the time reduces to 13.4 sec. Summarizing, there is no big difference in time consumption for 500 events.

The test program UNBINFIT FORTRAN and the input data UNBINFIT INPUT can be found on my VM-disk.

Dalitz Plot Fits

We now turn to fits of Dalitz plots using the unbinned data. The probability density function μ consists of a coherent sum of n complex amplitudes, weighted by the density of states w in phase space and the detector efficiency ϵ :

$$\mu = \epsilon w \left| \sqrt{f_1} \mathrm{e}^{i\phi_1} \frac{A_1}{\sqrt{N_1}} + \cdots + \sqrt{f_n} \mathrm{e}^{i\phi_n} \frac{A_n}{\sqrt{N_n}} \right|^2 \tag{10}$$

$$= \epsilon w \sum_{k=1}^{n} \sum_{l=1}^{n} \sqrt{f_k f_l} \, \mathrm{Re} \left\{ \mathrm{e}^{i(\phi_k - \phi_l)} \frac{A_k A_l^*}{\sqrt{N_k N_l}} \right\} , \tag{11}$$

where the relative fractions f_n and the phases ϕ_n are varied in the fit. One phase can of course arbitrarily be set to zero. The complex amplitudes A_n are evaluated as usual with the helicity formalism or the Lorentz-invariant (Rarita-Schwinger) formalism or the Zemach formalism. N_n are the normalizations of the squared amplitudes over phase space

$$N_n = \int A_n A_n^* \, \Omega \, , \tag{12}$$

which can be obtained with standard Monte Carlo summation once before the fit, $N_n = \sum_{j=1}^{\mathrm{M}} A_n A_n^* \, \Omega / \mathrm{M}$. This requires the knowledge of the total phase space volume.

Since the efficiency ϵ and the weight w are multiplicative factors, they can be factored out in the log-likelihood. Defining $\mu = \epsilon w \hat{\mu}$ we obtain for

the standard likelihood method and the generalized likelihood method the following functions to be minimized:

$$-\ln \mathcal{L}_{\rm S} \;=\; {\rm N} \ln \left(\int \mu \, \dot{\Omega} \right) - \sum_{i=1}^{\rm N} \ln \hat{\mu}_i \;\; \text{and} \tag{13}$$

$$-\ln \mathcal{L}_{\rm G} \;=\; \int \mu \, \dot{\Omega} - \sum_{i=1}^{\rm N} \ln \hat{\mu}_i \; . \tag{14}$$

There is no need to know the efficiencies and weights of the real events!

The normalization $\int \mu \, \dot{\Omega}$ needs to be calculated in each iteration of MINUIT. It is best done with Monte Carlo summation: $\int \mu \, \dot{\Omega} = \sum_{j=1}^{\rm M} \mu_j \, \Omega / {\rm M}$. Since μ_j contains the efficiency one either has to know the efficiency function over the phase space or the events have to be passed through a complete detector simulation. In the latter case the efficiency is either one or zero, depending on whether the event pass or does not pass the cuts. In this case the integral reduces to

$$\int \mu \, \dot{\Omega} = \sum_{j=1}^{\rm M} w \, \sum_{k=1}^{n} \sum_{l=1}^{n} \sqrt{f_k f_l} \, {\rm Re} \left\{ {\rm e}^{i(\phi_k - \phi_l)} \frac{A_k A_l^*}{\sqrt{N_k N_l}} \right\} \, \frac{\Omega}{\rm M} \, . \tag{15}$$

However, a complete detector simulation for tens of thousands of Monte Carlo events may be too time consuming, and one may have to assume the efficiency to be constant in phase space. Another possibility is to simulate the detector efficiency by effective cuts, which can be applied at the 4-vector level. Note that such cuts or a complete detector simulation needs to be done only once before the fit; all what is needed is to store for events which pass the cuts those variables which describe the position of the event in phase space.