E1-2003-103

S. Stoynev

FITTING EXPERIMENTAL DATA BY USING WEIGHTED MONTE CARLO EVENTS

1. Introduction

The purpose of this paper is to present a method for fitting experimental data using MC sample. It concerns the case when we have a single MC source and a number of unknown parameters to modify the primary MC distribution. As the distribution itself can be of an arbitrary (in principle) dimension, and as it is usually binned, some bins might contain very few events. Since both, experimental and MC data are finite, a careful treatment when evaluating the quantity to be minimized in search for "best fit" parameters and parameters errors should be applied.

A general approach for analyzing data involves fitting experimental data distribution (N-dimensional) with one or a composition of more MC based distributions. Simulated and the experimental data are both subject to statistical fluctuations. Usually the number of events in some bins is small, and χ^2 minimization is inappropriate. Then accounting for Poisson statistics a Maximum Likelihood Technique is preferable.

Let us denote the number of events in some N-dimensional experimental data distribution falling into bin i by d_i and the number of expected events from MC source j in the same bin as A_{ji} (which has unknown value). For each A_{ji} the corresponding simulated a_{ji} is generated by Poisson distribution (if $A_{ji} << N_j$ which is the case). Then the predicted number of events f_i in a given bin i is just the sum of the expected events from all MC sources taken in some unknown proportions P_j:

$$f_i = N_D \sum P_j A_{ji}/N_j \equiv \sum p_j A_{ji} \tag{1}$$

where N_D is the total number of events in the experimental data, and N_j is the the total number of events in the j-th MC source. For convenience we have introduced notation $p_j = N_D P_j / N_j$.

The logarithm of the Likelihood function to be maximized is the combined probability of the observed $\{d_i\}$ and the observed $\{a_{ji}\}$:

$$\ln L = \sum (d_i \ln f_i - f_i) + \sum \sum (a_{ji} \ln A_{ji} - A_{ji}) . \tag{2}$$

In this way we account correctly for small numbers of events in the bins for both data and MC distributions. The technique is called "binned maximum likelihood" fit [1]. The values of unknown parameters p_j and A_{ji} are found by maximizing this likelihood function.

In many cases it is necessary to apply weights to the MC distribution(s) before comparing them with the real data. In this case the predicted number of events is modified and eq. (1) becomes

$$f_i = \sum p_j w_{ji} A_{ji} . \tag{3}$$

The quantity w_{ji} does not need to be the same for all events from a given source in a given bin and is usually not. If one assumes it is the same (that is the ideal average weight for source j in bin i) then one should be worried about the discrepancy between the average actual weight and the true average weight. Fluctuations in the bins with small number of events with large weights will overwhelm the information obtained from low weight events. For details see [1], [2], [3].

In the case when we have to obtain the parameters of the matrix element for some physical process usually we have only one MC source. Then modifying the MC distribution according to some theory and fitting the experimental data we are extracting the values to these parameters. However, in this case the schema described above, as well as the standard HBOOK [3] routines are not applicable directly and require some modification.

In section 2 a way out to solve this problem is proposed. In section 3 an example of the application of the method developed is presented.

2. Fitting experimental data with modified MC sample

In the standard HBOOK realization P_j are considered as fractions of one primary source, i.e. $\sum P_j = 1$. The number of independent parameters P_j in the data/MC fit is of course exactly equal to the number of MC sources and the latter must be greater than 1.

Sometimes a little bit different case arise. Let us suppose we have only one MC source which we have to modify (following some theory) by the factor of $1 + \sum w_\alpha \lambda_\alpha$ or even

$$1 + \sum w_\alpha \lambda_\alpha + \sum v_\alpha \lambda_\alpha^2 + \sum \sum v_{\alpha\beta} \lambda_\alpha \lambda_\beta \tag{4}$$

where λ_α are parameters (for example some form factors), α and β are parameter indexes, w and v are some weights (for example kinematic variables). What we are looking for are the values of the parameters best fitting the experimental data. In this case, applying the underlying theory, eq. (3) obtains the following form:

$$f_i = A_{0i} \sum w_{ji} P'_j \,. \tag{5}$$

However the meaning of the quantities taking part in it is different. There is no need to use the total numbers of events N_D and N_j here, w_{ij} are just redefined ws and vs from eq. (4) ($w_{0i} \equiv 1$), P'_j are functions of λ and $\sum P'_j$ is not normalized to 1, j stands for a term number in eq. (4) rather than for a MC source. Eq. (2) then is modified to:

$$lnL = \sum (d_i ln f_i - f_i) + \sum (a_{0i} ln A_{0i} - A_{0i}) \,. \tag{6}$$

To find the maximum of lnL we differentiate eq. (6) and set the derivatives to zero. Differentials with respect to P'_j give (for each j):

$$\sum_i \left(\frac{d_i A_{0i} w_{ji}}{f_i} - A_{0i} w_{ji}\right) = 0 \,. \tag{7}$$

The ones with respect to A_{0i} are:

$$\frac{d_i}{f_i} \sum_j (w_{ji} P'_j) - \sum_j (w_{ji} P'_j) + \frac{a_{0i}}{A_{0i}} - 1 = 0 \,. \tag{8}$$

Multiplying eq. (8) by A_{0i} and taking into account (5) we obtain

$$d_i - A_{0i}\sum_j w_{ji}P'_j + a_{0i} - A_{0i} = 0 \tag{9}$$

or

$$A_{0i} = \frac{d_i + a_{0i}}{1 + \sum_j w_{ji}P'_j} \ . \tag{10}$$

If we define (taking into account that $w_{0i} \equiv 1$)

$$h_i \equiv \sum_j w_{ji}P'_j = C(1 + \sum_{j>0} w_{ji}P_j) \tag{11}$$

we arrive to:

$$f_i = A_{0i}h_i \qquad \text{and} \qquad \mathrm{A_{0i}} = \frac{\mathrm{d_i + a_{0i}}}{\mathrm{1 + h_i}} \tag{12}$$

where C is a normalization factor for MC-data and P_j are known functions of the parameters (form factors) to be found. The maximization of the likelihood function is equivalent to solving eqs. (7). Having P_j it is straightforward to obtain the parameters λ_k.

It can be shown [5] that the maximization of the log-likelihood functions $lnL_1 = \sum(d_i ln f_i - f_i)$ and $lnL_2 = \sum(a_{0i} ln A_{0i} - A_{0i})$ is equivalent to the minimization of the quantities $\chi_1^2 = 2[\sum(f_i - d_i + d_i ln\frac{d_i}{f_i})]$ and $\chi_2^2 = 2[\sum(A_{0i} - a_{0i} + a_{0i} ln\frac{a_{0i}}{A_{0i}})]$ correspondingly. Both functions asymptotically obey χ^2-distribution. The sum of this two χ^2-distributions

$$L_{gof} = 2[\sum(f_i - d_i + d_i ln\frac{d_i}{f_i}) + \sum(A_{0i} - a_{0i} + a_{0i} ln\frac{a_{0i}}{A_{0i}})] \tag{13}$$

is a χ^2-distribution itself and can be used as an estimator of the fit quality [5], [6].

In general, one needs not to use one and the same quantity for determination of the parameters, their errors and the goodness-of-fit.

3. A practical example

Let us suppose we want to determine the values of the possible form factors in the decay $K \to \pi e \nu$ (it does not matter here if the kaon is charged or neutral). This can be done by exploring the experimental distribution of the Dalitz plot density:

$$\rho(\chi_1, \ldots, \chi_n) \equiv \frac{\mathrm{d}^n N}{\mathrm{d}\chi_1 \ldots \mathrm{d}\chi_n} \tag{14}$$

where χ_i are corresponding independent kinematic variables describing the decay.

In our particular case there are two independent variables. The most general form of the Dalitz plot density in the kaon rest frame can be written as:

$$\rho(E_\pi, E_e) \sim B[1 + W_1\lambda_+ + W_2\lambda_+^2 + W_3(\frac{f_S}{f_+(0)})^2 + W_4(\frac{f_T}{f_+(0)})^2 + W_5\frac{f_S}{f_+(0)}\frac{f_T}{f_+(0)}] \tag{15}$$

where E_π is the energy of the pion, E_e is the energy of the charged lepton, B and W_j ($j = 1, \ldots, 5$) are some known functions of E_π and E_e. $f_+(0) \approx 1$ is a constant in which we are not really interested now. The parameters λ_+, $F_S \equiv \frac{f_S}{f_+(0)}$ and $F_T \equiv \frac{f_T}{f_+(0)}$ are the form factors we are searching for. According to the V-A theory of the weak interactions, F_S and F_T are equal to zero and our investigation can be considered as a test for this theory. Fig. 1 represents the shape of the Dalitz plot.

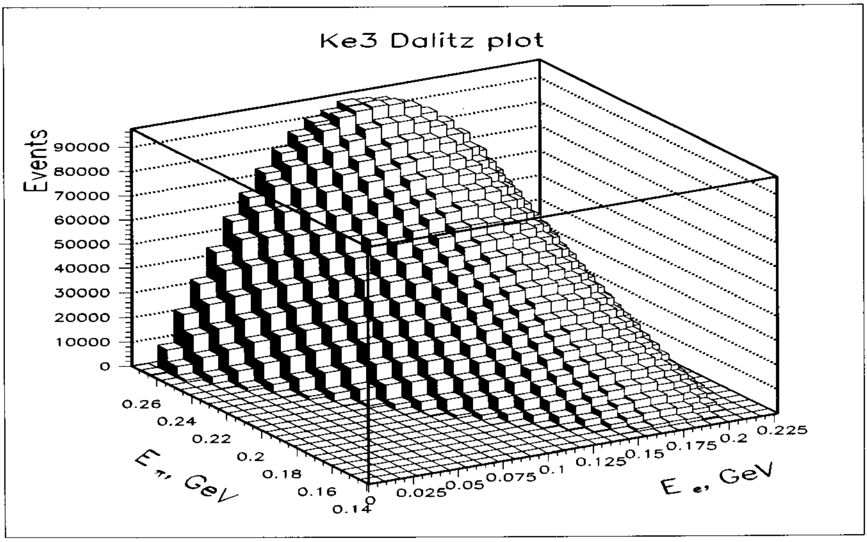

Figure 1. Representation of the Ke3 Dalitz plot

We generate a MC sample in which all the three parameters are set to zero. Then we evaluate the Ws (for each event) and weighte the binned MC Dalitz plot a_{0i} according to eq. (15) with three unknown parameters, i.e. eq. (11) in our case become:

$$h_i \equiv C(1 + W_1\lambda_+ + W_2\lambda_+^2 + W_3F_S^2 + W_4F_T^2 + W_5F_SF_T) \tag{16}$$

a_{0i} is obtained from the reconstructed E'_π and E'_e. They are not equal to the E_π and E_e used to calculate Ws due to the resolution of the experimental setup. The bins for which $a_{0i} = 0$ or $d_i = 0$ are rejected from the Log-likelihood sum. We account for this approximation by varying the width of the bins over the Dalitz plot.

Fitting the modified Dalitz plot to the experimental one using eqs. (16) and (6) (or (13)), we obtain results for the three form factors λ_+, F_S and F_T [7]. The number of degrees of freedom (DOF) in the fit is equal to the number of bins used minus the number of the free parameters (which is four). χ^2 from eq. (13) and DOF are a quantitative estimation of the goodness-of-fit.

The Log-likelihood function (with "minus" sign) in the F_S-F_T space is plotted on fig. 2. It is normalized so as the minimum of the function is at zero (there are 2 minima there which are equivalent to each other as can be seen from eq. (16)). Fig. 3 shows the same plot as confidence level contours. For goodness-of-fit estimation we used eq. (13) and the fit gave $\chi^2/DOF = 3010/2915$.

To check the method we use MC source with embedded (as proper weights) form factors instead of the experimental source. The values of the form factors are set to, as follows: $\lambda_+ = 0.032$, $|f_S/f_+(0)| = 0.02$ and $|f_T/f_+(0)| = 0.01$. Values for the scalar and tensor form factors are chosen to be roughly equal to the experimentally obtained ones. Results from the Log-likelihood fit are $\lambda_+ = 0.0320 \pm 0.0009$, $|f_S/f_+(0)| = 0.022 \pm 0.012$, $|f_T/f_+(0)| = 0.01 \pm 0.05$. When we are fitting without scalar and tensor parameters the result is $\lambda_+ = 0.0321 \pm 0.0005$. The form factor values are statistically coincident with the embedded ones.

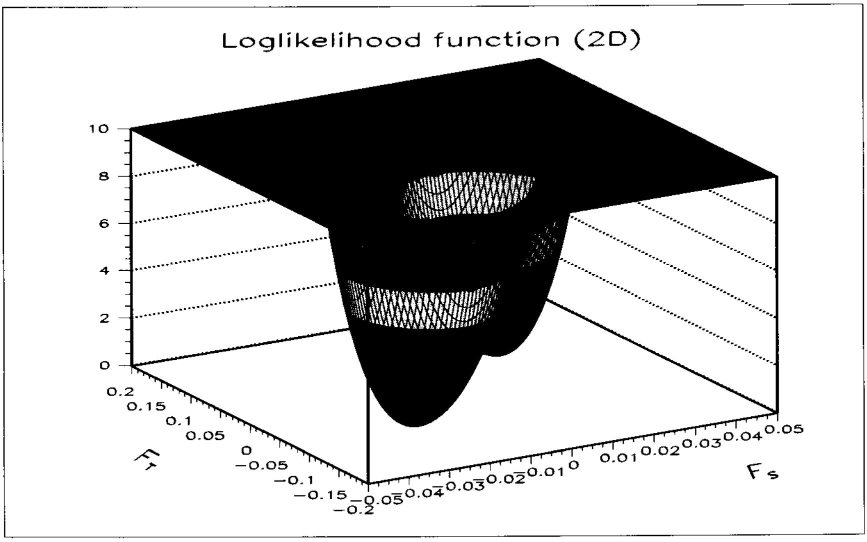

Figure 2. A view of the Log-likelihood function in two dimensions

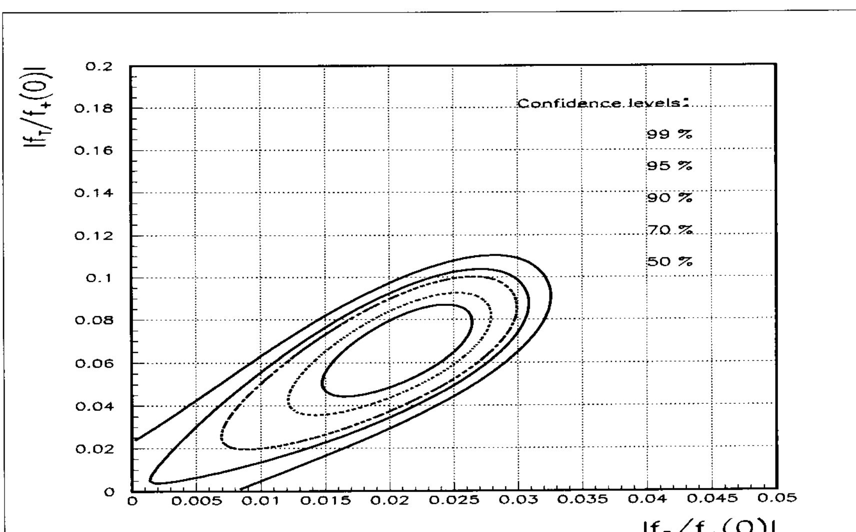

Figure 3. Confidence level contours in the $|\frac{f_S}{f_+(0)}|$-$|\frac{f_T}{f_+(0)}|$ plane at $\lambda_+ = 0.0288$ (value at the minimum)

4. Conclusions

A method for fitting experimental data with single finite MC sample has been proposed. Based on some underlying theory a Log-likelihood function has been deduced to help in fitting MC to experimental data. It accounts correctly for small numbers of events in the bins for both data and MC distributions. The method using this function has been checked by introducing some values of the fitted parameters in a MC sample and treating this sample like an experimental one. It has been shown that the method can be successfully applied in concerned regions of data analysis. A full practical example has been presented.

5. Acknowledgments

The author would like to thank Cvetan Cheshkov, Leandar Litov and Peter Hristov for the careful reading of the manuscript and the valuable and helpful remarks. Special thanks to Cvetan Cheshkov for the very close co-operation.

Reference

[1] Roger Barlow and Christine Beeston, Comp. Phys. Comm., 77:219, 1993.

[2] Roger Barlow, J. Comp. Phys., 72:202, 1987.

[3] CN/ASD Group. HBOOK users guide (Version 4.24), nProgram Library Y250. CERN, 1995.

[4] CN/ASD Group. MINUIT - Users guide (Version 94.1), nProgram Library D506. CERN, 1994.

[5] S. Baker and R. Cousins, NIM 221, 437 (1984).

[6] K. Hagiwara et al., Phys. Rev. D 66 (2002) 010001

[7] The NA48 Collaboration, "Experimental measurement of $Ke3$ form factors", in preparation.

Received on June 2, 2003.

Стойнев С. E1-2003-103
Фитирование экспериментальных данных
с использованием взвешенных событий Монте-Карло

Разработан метод для фитирования экспериментальных данных с помощью событий, полученных методом Монте-Карло. Метод позволяет извлечь величины параметров из теории, описывающей данный физический процесс, используя ограниченное число моделированных событий. На основе метода максимального правдоподобия проводится измерение величин искомых параметров, их ошибок, а также величины, которая характеризует качество фита.

Работа выполнена в Лаборатории физики частиц ОИЯИ.

Сообщение Объединенного института ядерных исследований. Дубна, 2003

Stoynev S. E1-2003-103
Fitting Experimental Data
by Using Weighted Monte Carlo Events

A method for fitting experimental data using modified Monte Carlo (MC) sample is developed. It is intended to help when a single finite MC source has to fit experimental data looking for parameters in a certain underlying theory. The extraction of the searched parameters, the errors estimation and the goodness-of-fit testing is based on the binned maximum likelihood method.

The investigation has been performed at the Laboratory of Particle Physics, JINR.

Communication of the Joint Institute for Nuclear Research. Dubna, 2003

Макет *Т. Е. Попеко*

Подписано в печать 16.06.2003.
Формат 60 × 90/16. Бумага офсетная. Печать офсетная.
Усл. печ. л. 0,68. Уч.-изд. л. 1,34. Тираж 365 экз. Заказ № 53955.

Издательский отдел Объединенного института ядерных исследований
141980, г. Дубна, Московская обл., ул. Жолио-Кюри, 6.
E-mail: publish@pds.jinr.ru
www.jinr.ru/publish/