Subtracting and Fitting Histograms using Profile Likelihood

F.M.L. de Almeida Jr. and A.A. Nepomuceno
Instituto de Física, Universidade Federal do Rio de Janeiro,RJ,Brazil

Abstract

It is known that many interesting signals expected at LHC are of unknown shape and strongly contaminated by background events. These signals will be difficult to detect during the first years of LHC operation due to the initial low luminosity. In this work, one presents a method of subtracting histograms based on the profile likelihood function when the background is previously estimated by Monte Carlo events and one has low statistics. Estimators for the signal in each bin of the histogram difference are calculated so as limits for the signals with 68.3% of Confidence Level in a low statistics case when one has a exponential background and a Gaussian signal. The method can also be used to fit histograms when the signal shape is known. Our results show a good performance and avoid the problem of negative values when subtracting histograms.

1 Introduction

The search for signals of low statistics has led to a strong development on statistics methods for high energy physics. Recently, methods based on profile likelihood has been widely used in problems related to setting limits to a signal and to test hypotheses. This approach shows very good performance in extracting signal information in the presence of nuisance parameters [1].

In this work one considers a χ^2-function obtained from the profile likelihood for subtracting histograms where the signal to backgrounds ration is small, and both distribution have unknown shape. One also shows that, when the signal distribution is known, one can use this χ^2-function to fit the signal without fitting the background. It is presented in the next Section the road map to this new χ^2-function. Section 3 presents the results for extracting signal information by subtracting histograms, and limits to signal are computed using the proposed χ^2-function. Section 4 shows an example on the fit method.

2 Likelihood and Profile Likelihood

Let us assume a counting experiment such that the signal and background events are completely independent and both obey to Poisson distributions. The background events are first estimated using the Monte Carlo method, running the experiment in "idle" mode or by any other technique. Suppose that during the experiment k data events are obtained and that m background events were previously estimated using Monte Carlo(MC) techniques. Since the number of previously estimated MC events depends on computational resources, it is possible to generate τ samples, such that

$$\tau = \mathcal{L}_{\mathcal{MC}}/\mathcal{L}_{\mathcal{EXP}}, \tag{1}$$

where $\mathcal{L}_{\mathcal{EXP}}$ and $\mathcal{L}_{\mathcal{MC}}$ are the experimental and MC luminosities, respectively, and $\tau > 0$. When one has limited computer resources, τ may be restricted τ to the range $0 < \tau < 1$. Any information about the background is helpful in order to extract as clean a signal as possible. The likelihood corresponding to the above discussion is

$$L(s, b; k, m, \tau) \propto (s+b)^k e^{-(s+b)} (\tau b)^m e^{-\tau b}, \tag{2}$$

where s and b are related to the signal and background distributions, respectively.

To obtain a b independent likelihood, one can find the maximum likelihood estimator of the background $\widehat{b}$ as a function of s and replace the true value b by $\widehat{b}$ in Eq. (2). Taking the derivative of that equation, and solving it for $b \geq 0$, one gets

$$\widehat{b}(s) = max\left(0, \frac{k+m-(1+\tau)s+\Delta(s)}{2(1+\tau)}\right), \tag{3}$$

where

$$\Delta(s) = \sqrt{[k+m-(1+\tau)s]^2 + 4m(1+\tau)s} \geq 0. \tag{4}$$

Replacing b by $\widehat{b}(s)$ in Eq. (2), one obtains the profile likelihood $L_P(s;k,m,\tau)$, which does not dependent on b [2].

$$L_P(s;k,m,\tau) \propto (s+\widehat{b}(s))^k e^{-(s+\widehat{b}(s))}(\tau\widehat{b}(s))^m e^{-\tau\widehat{b}(s)}. \tag{5}$$

The maximum value of L_P and the most probable value of s, $\widehat{s}$, are obtained by solving Eq (5). The simple analytical solution for $\widehat{s}$ is an unbiased value

$$\widehat{s} = max\left(0, k - \frac{m}{\tau}\right), \tag{6}$$

since $s \geq 0$ due to physical constrains. The parameter $\widehat{s}$ is just the maximum profile likelihood estimator of s.

Let us construct now an approximate χ^2-function using Eq (5). The maximum profile likelihood ratio is given by

$$\lambda_P = \frac{L_P(s,k,m,\tau)}{L_P(\widehat{s},k,m,\tau)}, \tag{7}$$

where the denominator is the maximum profile likelihood, which occurs when $s = \widehat{s}$. According to the maximum likelihood ratio theorem $\chi^2_P \approx -2\log\lambda_P$ and hence, the profile χ^2_P-function is written as

$$\chi^2_P = 2\left\{(s-\widehat{s}) + (\tau+1)\left(\widehat{b}(s) - \widehat{b}(\widehat{s})\right) + k\ln\left(\frac{\widehat{s}+\widehat{b}(\widehat{s})}{s+\widehat{b}(s)}\right) + m\ln\left(\frac{\widehat{b}(\widehat{s})}{\widehat{b}(s)}\right)\right\}, \tag{8}$$

where $\widehat{b}(s)$ and $\widehat{s}$ are given by Eqs (3,6), respectively, so as $\widehat{b}(\widehat{s})$.

3 Subtracting Histograms and Setting Limits

In order to show the applicability of the χ^2_P-function obtained we generated 500 Toy Monte Carlo events, such that 50 were signal and 450 were background, distributed in a histogram of 50 bins. The signal and background were generated according to Gaussian and Exponential functions, respectively,

$$S \sim Gauss(1.2, 0.2),\ B \sim Exp(-x). \tag{9}$$

The number of background events in each bin was previously estimated by generating 2250 background events, corresponding to $\tau = 5$. It is useful to mention at this point that there is no advantage in taking $\tau > 5$ when one estimates the background from MC, since there is no relevant change in the χ^2_P-function for $\tau > 5$. Figure 1 shows the 'data', the background previously estimated and the signal. To

extract the signal histogram from the 'data', one can use Eq (6), which give us the signal estimated for each bin. Its limits (s_{min}, s_{max}) are obtained by solving the system

$$\begin{cases} \int_{s_{min}}^{s_{max}} f_P(s;k,m,\tau)ds = 1-\alpha \\ \\ \chi^2_P(s_{min}) = \chi^2_P(s_{max}) \\ \\ 0 \leq s_{min} < s_{max} \end{cases} \tag{10}$$

where $f_P(s;k,m,\tau)$ is the normalized probability distribution of s given k, m and τ obtained normalizing $L_P(s;k.m,\tau)$ with respect to s, and α depends on the chosen confidence level.

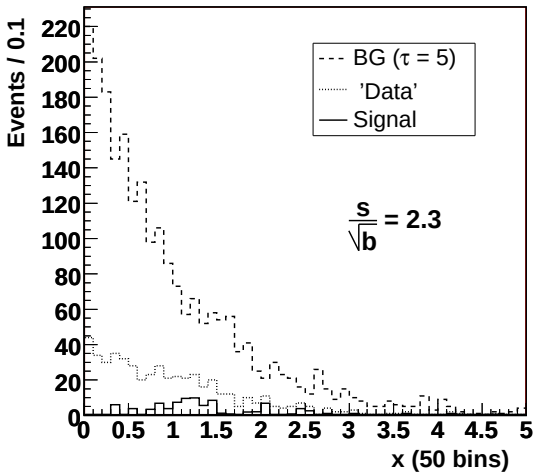

Fig. 1: Toy Monte Carlo Example. The full line represents the signal contained in the 'data'. The background was previously estimated with $\tau = 5$.

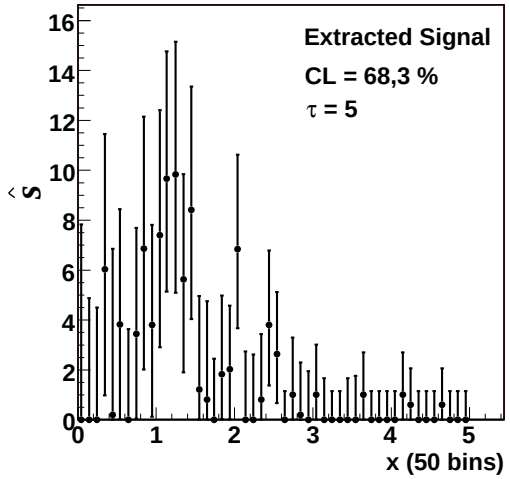

Fig. 2: Extracted signal. The signal limits were calculated for a confidence level of 68.3%. The constraining $\widehat{s} > 0$ avoid bins with negative values.

The subtracted histogram result is shown in Fig. 2. The points are the signal estimated for each bin and the error bars were calculated using Eq (10) for a confidence level of 68.3%. Notice that we have no bin with negative values due the constraint $\widehat{s} > 0$. It is important to mention also that one did not need to know the true background rate b in order to get signal limits, since the χ^2_P-function, given by Eq. (8), does not depend on that parameter since it has been replaced by an estimate.

The signal significance can be obtained by looking at the *P*-value under the hypothesis that one has no signal. Taking into account just the bins between $x = 0.85$ and $x = 2.5$, one gets a *P*-value of 0.022.

4 Fitting Histograms

When the signal shape is known, one can use Eq. (8) to fit histograms. In such case, the χ^2_P-function that will be minimized is given by the sum of all $\chi^2_{P_i}(s_i, k_i, m_i, \tau)$ which correspond to N bin contributions, where s_i must be substituted by the function $f(x_i, \boldsymbol{\theta})$ to be fit, x_i being the corresponding ordinate in the i^{th} bin and $\boldsymbol{\theta}$ the parameter vector to be fitted.

One can apply this approach to fit the signal in the Monte Carlo sample shown before, but now the events are distributed in a histogram of 100 bins, since one knows now the signal distribution shape, as shown in Fig. 3. The number of previously estimated background events in each bin m_i is given by the

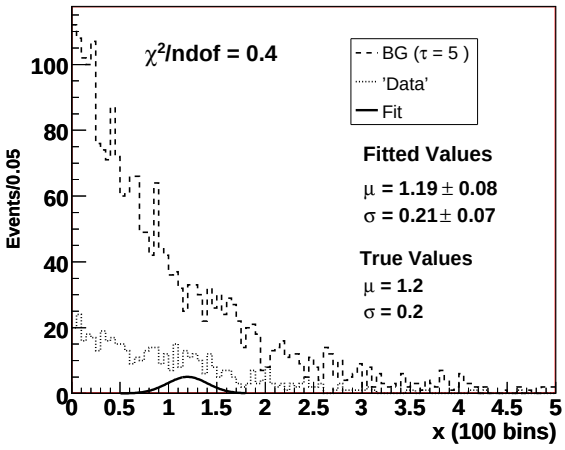

Fig. 3: Previously estimated background, 'data' and fitted curve.

histogram labeled BG in Fig. 3, and k_i is the number of 'data' events in each bin. The signal distribution s_i is substituted by a Gaussian function, and $\tau = 5$. By the minimization of the χ^2_P-function, one gets the fitted parameters $\mu = 1.19 \pm 0.08$ and $\sigma = 0.21 \pm 0.07$, which are in very good agreement with the "true" values 1.2 and 0.2, respectively. The full line in Fig. 3 shows the fitted curve.

Notice that as Eq. (8) depends just on $f(x_i, \boldsymbol{\theta})$, k_i, m_i and τ, one did not need to fit the background distribution, and the only necessary information from background was its number of events m_i estimated by MC. This is the great advantage of this method. The χ^2_P-function already incorporates the background statistical fluctuations. Besides reducing the numbers of fitted parameters, this method presents no problems when one has few or no events in one or more bins as can occur in data with long tails. Even the bins with $k_i = 0$ and/or $m_i = 0$ contributes to the χ^2_P-function. It is is only necessary to fit the signal function parameters which will allow us to obtain a much cleaner and less noisy analysis. This will affect in a positive way the parameter covariance matrix. A systematic study of this method was done for different τ values and different signal and background distributions, and in all cases the method showed very good performance.

5 Summary

The proposed χ^2_P-function can be used to extract signal information without need to know the background distribution shape. The fact that one just needs to fit the signal reduce the number of parameters to be fitted and avoid the uncertainties carrying by the lack of knowledge of the exact background parameters. The method works well even in situation where there is very low statistics.

Acknowledgments

This work was partially supported by the Brazilian Agencies CNPq and CAPES and the HELEN project.

References

[1] W. A. Rolke, A.M. Lopez, Nuclear Instruments and Methods **A551** (2005) 493.

[2] W. A. Rolke, A.M. Lopez, Nuclear Instruments and Methods **A458** (2001) 745.