Simultaneous Determination of Signal and Background Asymmetries

Jörg Pretz *

Physikalisches Institut, Universität Bonn, 53115 Bonn, Germany

Jean-Marc Le Goff

CEA Saclay IRFU/SPhN, 91191 Gif-sur-Yvette, France

Abstract

This article discusses the determination of asymmetries. We consider a sample of events consisting of a peak of signal events on top of some background events. Both signal and background have an unknown asymmetry, e.g. a spin or forward-backward asymmetry. A method is proposed which determines signal and background asymmetries simultaneously using event weighting. For vanishing asymmetries the statistical error of the asymmetries reaches the minimal variance bound (MVB) given by the Cramér-Rao inequality and it is very close to it for large asymmetries. The method thus provides a significant gain in statistics compared to the classical method of side band subtraction of background asymmetries. It has the advantage with respect to the unbinned maximum likelihood approach, reaching the MVB as well, that it does not require loops over the event sample in the minimization procedure.

Key words: event weighting, minimal variance bound, Cramér-Rao inequality, asymmetry extraction, optimal observables, side band subtraction
PACS: 02.70.Rr, 13.88.+e

1 Introduction

Asymmetries of cross sections, e.g. spin-asymmetries and forward-backward asymmetries, are often interesting physics quantities. For concreteness let us

* corresponding author
Email address: jorg.pretz@cern.ch (Jörg Pretz).

consider a situation as shown in Fig. 1, where the asymmetry of the signal events, in the central Gaussian peak with a width of σ, should be determined from data taken in two different spin configurations. The number density of events as a function of some kinematic variable, x, (typically a reconstructed mass) is given by

$$n^{\pm}(x) = a(x)(\sigma_S(x) + \sigma_B(x)) \left(1 \pm A_S \frac{\sigma_S(x)}{\sigma_S(x) + \sigma_B(x)} \pm A_B \frac{\sigma_B(x)}{\sigma_S(x) + \sigma_B(x)}\right)$$

with $\sigma_{S,B} = \frac{1}{2}(\sigma^+_{S,B} + \sigma^-_{S,B})$. Here $\sigma^{\pm}_S$ ($\sigma^{\pm}_B$) denotes the cross section of the signal (background) events in the two different spin configurations $+$ and $-$. The factor a is a luminosity and acceptance factor, assumed to be the same for the two spin configurations. The goal is to determine from spectra as shown in Fig. 1, and taken in two spin configurations, the two unknown asymmetries $A_S = (\sigma^+_S - \sigma^-_S)/(\sigma^+_S + \sigma^-_S)$ and $A_B = (\sigma^+_B - \sigma^-_B)/(\sigma^+_B + \sigma^-_B)$, assumed to be independent of x. It is of course not known event-by-event whether a particular event is signal or background; one only knows the fraction of signal events as a function of x, from a fit to the event spectrum as in Fig. 1.

Section 2 presents the simplest method, based on counting rate asymmetries. Section 3 describes the unbinned likelihood method which is known to yield the smallest possible variance of all unbiased estimators in the limit of an infinite number of events, thus reaching the minimal variance bound (MVB) given by the Cramér-Rao inequality. Section 4 presents a new asymmetry estimator, based on weighted events. This estimator is also unbiased in the large N limit, i.e. it is consistent, and it is very close to reach the minimal variance bound. The advantage is that it can also be used in cases where the unbinned likelihood method is cumbersome because of large number of events. Event weighting to extract the number of signal and background events was discussed in Ref. [1] but extraction of asymmetries is not discussed in this reference. The different methods are compared in section 5.

2 Estimator based on counting rate asymmetries

A method often found in the literature [2,3] is to determine the asymmetry in a k-standard-deviation region around the peak, a region which includes both signal and background; then to measure the background asymmetry in some side bands around the signal peak ($-k_{max}\sigma < x < -k_{min}\sigma$ and $k_{min}\sigma < x < k_{max}\sigma$) and to use the result to correct the asymmetry measured in the peak region. For sake of simplicity we will set $\sigma = 1$, so that everywhere below we can write k instead of $k\sigma$.

The expectation value of the counting rate asymmetry, A^{cnt}, in the range

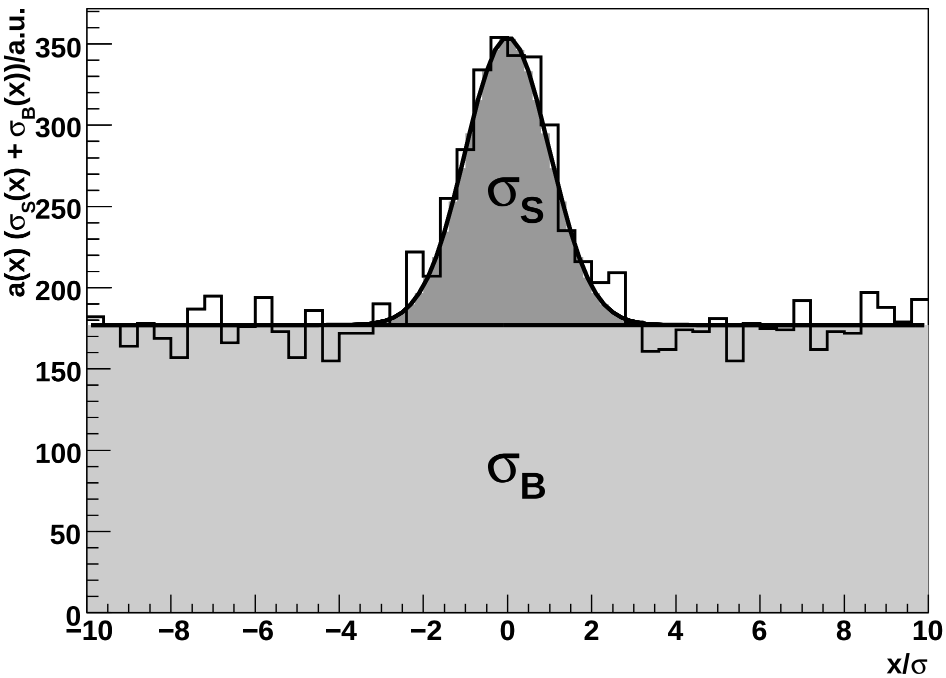

Fig. 1. Example of signal events originating from a Gaussian distribution centered at $x = 0$ and width $\sigma = 1$ sitting on a constant background.

$-k < x < k$ is related to A_S and A_B in the following way:

$$\langle A^{cnt} \rangle = \frac{\langle N^+ \rangle - \langle N^- \rangle}{\langle N^+ \rangle + \langle N^- \rangle} = A_S \frac{\int_{-k}^{k} a\sigma_S \mathrm{d}x}{\int_{-k}^{k} a(\sigma_S + \sigma_B)\mathrm{d}x} + A_B \frac{\int_{-k}^{k} a\sigma_B \mathrm{d}x}{\int_{-k}^{k} a(\sigma_S + \sigma_B)\mathrm{d}x} \,, \quad (1)$$

where we used $\langle N^+ \rangle = \int n^+(x)\mathrm{d}x$ and $\langle N^- \rangle = \int n^-(x)\mathrm{d}x$. An estimator for A_S is given by:

$$\tilde{A}_S = \frac{\int_{-k}^{k} a(\sigma_S + \sigma_B)\mathrm{d}x}{\int_{-k}^{k} a\sigma_S \,\mathrm{d}x} \left(A^{cnt} - \frac{\int_{-k}^{k} a\sigma_B \mathrm{d}x}{\int_{-k}^{k} a(\sigma_S + \sigma_B)\mathrm{d}x} A_B \right) . \quad (2)$$

Note that, strictly speaking, the first equality in Eq. (1) is valid only in the large N limit. In this limit Eqs. (1) and (2) indicate that $\langle \tilde{A}_S \rangle = A_S$, i.e. $\tilde{A}_S$ is a consistent estimator.

The corresponding figure of merit, FOM=$1/\sigma^2_{\tilde{A}_S}$, reads

$$\mathrm{FOM} = \left(\frac{\int_{-k}^{k} a\sigma_S \mathrm{d}x}{\int_{-k}^{k} a(\sigma_S + \sigma_B)\mathrm{d}x} \right)^2 \left(\sigma^2_{A^{cnt}} + \left(\frac{\left(\int_{-k}^{k} a\sigma_B \mathrm{d}x\right)}{\int_{-k}^{k} a(\sigma_S + \sigma_B)\mathrm{d}x} \right)^2 \sigma^2_{A_B} \right)^{-1} . \quad (3)$$

Here and in the following we assume small asymmetries, such that for the error calculation the approximation $\langle N^+ \rangle \approx \langle N^- \rangle$ is valid. In this case one finds

$1/\sigma^2_{A^{cnt}} = \int_{-k}^{k} [n^+(x) + n^-(x)]\,\mathrm{d}x$ and $1/\sigma^2_{A_B} = \int_{-k_{max}}^{-k_{min}} [n^+(x) + n^-(x)]\mathrm{d}x + \int_{k_{min}}^{k_{max}} [n^+(x) + n^-(x)]\mathrm{d}x$. Introducing these values of $\sigma^2_{A^{cnt}}$ and $\sigma^2_{A_B}$ in Eqs. 3 shows that the FOM depends on the choice of both the signal region (k) and the background region (k_{min} and k_{max}). The solid line in Fig. 2 shows the FOM as a function of k_{max}, for $k_{min} = 3$ which is a reasonable value to make sure that the side bands include a negligible amount of signal. The signal region, i.e. the value for k, is chosen in order to maximize the FOM for the given k_{max}. The FOM depends also on the signal to background ratio, here chosen to be 1:1 at $x = 0$, as in Fig. 1.

3 Maximum Likelihood asymmetry estimators

In the large N limit, the unbinned maximum likelihood method is known to provide an unbiased estimator for the parameters A_S and A_B, which reaches the minimal variance bound. Since the numbers of events N^+ and N^- are not fixed, an extended maximum likelihood method has to be used [4]. With the definitions $S_i = \sigma_S(x_i)/(\sigma_S(x_i) + \sigma_B(x_i))$, $B_i = \sigma_B(x_i)/(\sigma_S(x_i) + \sigma_B(x_i))$ and $\alpha_i = a(x_i)\,(\sigma_S(x_i) + \sigma_B(x_i))$ the log likelihood function reads:

$$
\begin{aligned}
l = \ln \mathcal{L} = &\sum_1 \ln\left(\alpha_i(1 + S_i A_S + B_i A_B)\right) - \langle N^+\rangle(A_S, A_B) \\
&+ \sum_2 \ln\left(\alpha_i(1 - S_i A_S - B_i A_B)\right) - \langle N^-\rangle(A_S, A_B)\,,
\end{aligned}
$$

where Σ_1 (Σ_2) runs over all events in the $+$ ($-$) configuration and in the range $-k_{max} < x < k_{max}$, while $\langle N^\pm\rangle(A_S, A_B) = \int n^\pm(x)\mathrm{d}x$. The first derivative is

$$
\frac{\partial l}{\partial A_S} = \sum_1 \frac{S_i}{1 + S_i A_S + B_i A_B} - \sum_2 \frac{S_i}{1 - S_i A_S - B_i A_B}\,, \qquad (4)
$$

with a similar expression for A_B. Note that the terms with $\langle N^+\rangle$ and $\langle N^-\rangle$ cancel each other because the same a is assumed for the two configurations. The set of equations $\partial l/\partial A_{S,B} = 0$ can be solved for A_S and A_B.

For small asymmetries a first order expansion in A_S and A_B gives the set of equations

$$
\begin{aligned}
\left(\sum_1 S_i^2 + \sum_2 S_i^2\right) A_S + \left(\sum_1 S_i B_i + \sum_2 S_i B_i\right) A_B &= \sum_1 S_i - \sum_2 S_i\,, \\
\left(\sum_1 S_i B_i + \sum_2 S_i B_i\right) A_S + \left(\sum_1 B_i^2 + \sum_2 B_i^2\right) A_B &= \sum_1 B_i - \sum_2 B_i
\end{aligned} \qquad (5)
$$

and the covariance matrix of the two parameters A_S and A_B reads :

$$\mathrm{cov}^{-1}(A_S, A_B) = -\begin{pmatrix} \frac{\partial^2 l}{\partial A_S^2} & \frac{\partial^2 l}{\partial A_S \partial A_B} \\ \frac{\partial^2 l}{\partial A_S \partial A_B} & \frac{\partial^2 l}{\partial A_B^2} \end{pmatrix} = \begin{pmatrix} \sum S_i^2 & \sum S_i B_i \\ \sum S_i B_i & \sum B_i^2 \end{pmatrix}, \qquad (6)$$

For the FOM of A_S one finds

$$\mathrm{FOM} = (1-\rho^2) \sum S_i^2 \quad \text{with} \quad \rho = \frac{\sum S_i B_i}{\sqrt{\sum S_i^2 \sum B_i^2}} . \qquad (7)$$

Note that, if not otherwise stated, all sums run over both event samples, 1 and 2.

The dotted line in Fig. 2 shows this FOM as a function of k_{max}, i.e. for events in the region $-k_{max} < x < k_{max}$. For a given range of data available, defined by k_{max}, it is always larger than the FOM obtained with the side band subtraction method shown by the solid line. The latter method does not reach the minimal variance bound.

4 Extracting the asymmetries using event weighting

In this section a method to extract A_S (and simultaneously A_B) using event weighting is developed. It is clear that the estimator based on the counting rate asymmetry is not statistically optimal since it gives the same weight to all events. Better estimators can be obtained by weighting each event by the signal strength, S_i, and by the background strength, B_i. These weight factors coincide with the optimal weights found in [1] to extract the number of signal events. They are used to build the following asymmetries:

$$a_S = \frac{\sum_1 S_i - \sum_2 S_i}{\sum_1 S_i + \sum_2 S_i}, \qquad a_B = \frac{\sum_1 B_i - \sum_2 B_i}{\sum_1 B_i + \sum_2 B_i}. \qquad (8)$$

In the large N limit, the expectation values of a_S and a_B are

$$\langle a_S \rangle = A_S \frac{\int \alpha S^2 \mathrm{d}x}{\int \alpha S \mathrm{d}x} + A_B \frac{\int \alpha BS \mathrm{d}x}{\int \alpha S \mathrm{d}x}, \qquad (9)$$

$$\langle a_B \rangle = A_S \frac{\int \alpha BS \mathrm{d}x}{\int \alpha B \mathrm{d}x} + A_B \frac{\int \alpha B^2 \mathrm{d}x}{\int \alpha B \mathrm{d}x}, \qquad (10)$$

where $\alpha = a(x)\,(\sigma_S(x) + \sigma_B(x))$, as in section 3. The ratios of integrals can easily be obtained from the event sample, e.g. $\int \alpha S^2 \mathrm{d}x \,/\, \int \alpha S \mathrm{d}x \approx \sum_{1,2} S_i^2 \,/\, \sum_{1,2} S_i$, which results exactly in the set of equations (5) found for the likelihood method

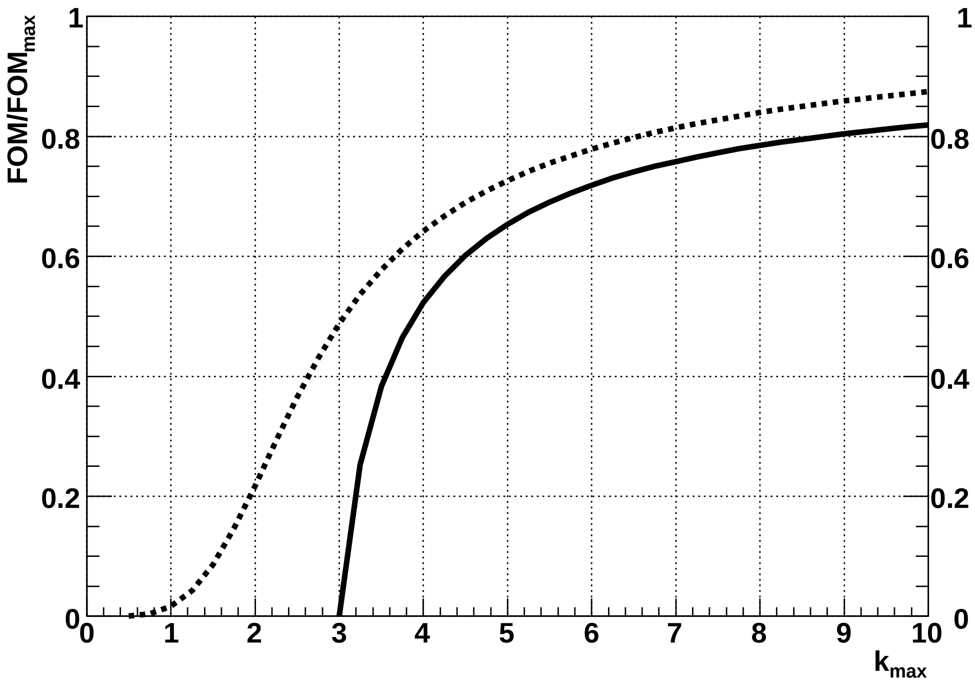

Fig. 2. FOM of A_S as a function of the maximum range of data available defined by k_{max}, for the classical method of side band subtraction (solid line) and for the likelihood or weighting method (dotted line). In the side band subtraction method $k_{min} = 3$ and for each value of k_{max} the value of k, defining the signal region, is chosen in order to maximize the FOM. The figures of merit are normalized to the maximum FOM reachable in the likelihood or weighting method in the limit $k_{max} \to \infty$. In this case FOM $= \sum_{1,2} S_i^2$.

in the small asymmetry limit. So the FOM is still $1/\sigma_{A_S}^2 = (1-\rho^2)\sum S_i^2$. This result can of course also be obtained directly, by simple error propagation using the expressions found for A_S and A_B from Eqs. (9) and (10). Appendix A shows that the factor ρ is actually the correlation coefficient between $\sum S_i$ and $\sum B_i$.

This shows that the weighting method and the unbinned likelihood method are identical for small asymmetries. The advantage of the weighting method is that the estimators derived from Eqs. (9) and (10) can also be used for arbitrary asymmetries, whereas the likelihood method requires in this case a numerical maximization of $\ln \mathcal{L}$ with loops over all events. For sake of simplicity, the error calculation was only presented for small asymmetries. Extending it to arbitrary asymmetries is straightforward but lengthy; it shows that the FOM of the weighting method is only slightly smaller than the FOM of the unbinned LH method. For example for a signal to background ratio as given in Fig. 1 and asymmetries smaller than 50% the decrease in the FOM is less than 1%.

The weighting method can also be extended to more complicated cases where for example the acceptance factors a are not the same in the two spin configurations or even when the asymmetries have to be determined from four counting rates in order to cancel differences of acceptances and flux factors for the two spin configurations, as in Ref. [5].

5 Discussion of the results & summary

A comparison of the two curves in Fig. 2 shows that the FOM of the likelihood or event weighting method is always larger than the corresponding FOM for the classical method. For a signal-to-background ratio of 1:1 at $x = 0$, as in Fig. 1, the gain is 23% for $k_{max} = 4$ and 7% for $k_{max} = 10$. For $k_{max} = 10$ the gain is 2% and 10% for a signal-to-background ratio of 10:1 and 1:10, respectively. Apart from the gain in statistics it should also be noted that the weighting method avoids the arbitrary choice of the background region which starts here at 3σ. For Breit-Wigner distributions for example this choice is less obvious.

In summary, a new set of two estimators was presented to determine simultaneously signal and background asymmetries. These estimators are unbiased in the large N limit, i.e. they are consistent. For small asymmetries they are also efficient, i.e. they reach the minimal variance bound, like the statistically optimal unbinned likelihood method. This is in contrast to the classical method of side band subtraction. These estimators can actually be derived from the likelihood method in the case of vanishing asymmetries. For large asymmetries their variances are still very close to the minimal variance bound. The advantage of the method is its applicability in cases where the likelihood method is cumbersome.

A Derivation of the covariance matrix cov(a_S, a_B) and correlation coefficient ρ

Consider two weight factors S and B. The covariance between $\sum_i S_i$ and $\sum_j B_j$ is given by:

$$
\begin{aligned}
&\mathrm{cov}(\sum_i S_i, \sum_j B_j) \\
&= \langle \sum_i S_i \sum_j B_j \rangle - \langle \sum_i S_i \rangle \langle \sum_j B_j \rangle \\
&= \langle \sum_{i=j} S_i B_i + \sum_{i \neq j} S_i B_j \rangle - \langle \sum_i S_i \rangle \langle \sum_j B_j \rangle \\
&= \langle N \rangle \langle SB \rangle + \langle N(N-1) \rangle \langle S \rangle \langle B \rangle - \langle N \rangle^2 \langle S \rangle \langle B \rangle \\
&= \langle N \rangle \langle SB \rangle + (\langle N^2 \rangle - \langle N \rangle - \langle N \rangle^2) \langle S \rangle \langle B \rangle \,.
\end{aligned}
$$

If the number of events N is Poisson distributed, i.e. $\langle N^2 \rangle - \langle N \rangle - \langle N \rangle^2 = 0$, one finds $\mathrm{cov}(\sum_i S_i, \sum_j B_j) = \langle N \rangle \langle SB \rangle \approx \sum_i S_i B_i$. The error on the sums of weights is given by $\sigma_S^2 = \mathrm{cov}(\sum_i S_i, \sum_j S_j) = \sum_i S_i^2$. Thus the correlation coefficient is

$$
\rho = \frac{\mathrm{cov}(\sum_i S_i, \sum_j B_j)}{\sigma_S \sigma_B} = \frac{\sum_i S_i B_i}{\sqrt{\sum_i S_i^2 \sum_i B_i^2}} \,. \tag{A.1}
$$

References

[1] R. J. Barlow, J. Comput. Phys. **72** (1987) 202.

[2] R. Barate et al. [ALEPH Collaboration], Phys. Lett. B **434** (1998) 415

[3] A. Adare et al. [PHENIX Collaboration], Phys. Rev. D **76** (2007) 051106

[4] R. J. Barlow, Statistics, Wiley 1989

[5] COMPASS, M. Alekseev et al., CERN-PH-EP/2008-003 (arXiv:0802.3023)