${}_s\mathcal{P}lot$: a statistical tool to unfold data distributions

M. Pivk[a] and F.R. Le Diberder[b]

[a] *CERN, CH-1211 Geneva 23, Switzerland*

[b] *Laboratoire de l'Accélérateur Linéaire, IN2P3-CNRS et Université de Paris-Sud, F-91898 Orsay, France*

Abstract

The paper advocates the use of a statistical tool dedicated to the exploration of data samples populated by several sources of events. This new technique, called ${}_s\mathcal{P}lot$, is able to unfold the contributions of the different sources to the distribution of a data sample in a given variable. The ${}_s\mathcal{P}lot$ tool applies in the context of a Likelihood fit which is performed on the data sample to determine the yields of the various sources.

Contents

1 Introduction

This paper describes a new technique to explore a data sample when the later consists of several sources of events merged into a single sample of events. The events are assumed to be characterized by a set of variables which can be split into two components. The first component is a set of variables for which the distributions of all the sources of events are known: below, these variables are collectively referred to as a (unique) *discriminating* variable. The second component is a set of variables for which the distributions of some sources of events are either truly unknown or considered as such: below, these variables are collectively referred to as a (unique) *control* variable.

The new technique, termed ${}_s\mathcal{P}lot$, allows to reconstruct the distributions for the control variable, independently for each of the various sources of events, without making use of any *a priori* knowledge on this variable. The aim is thus to use the knowledge available for the discriminating variable to be able to infer the behavior of the individual sources of events with respect to the control variable. An essential assumption for the ${}_s\mathcal{P}lot$ technique to apply is that the control variable is uncorrelated with the discriminating variable.

The ${}_s\mathcal{P}lot$ technique is developed in the context of a data sample analyzed using a maximum Likelihood method making use of the discriminating variable. Section 2 is dedicated to the definition of fundamental objects necessary for the following. Section 3 presents an intermediate technique, simpler but inadequate, which is a first step towards the ${}_s\mathcal{P}lot$ technique. Section 4 is the core of the document where the ${}_s\mathcal{P}lot$ formalism is developed (Section 4.1) and its properties explained in detail (Section 4.2). Section 4.3 then gives instructions about how to implement and use ${}_s\mathcal{P}lot$. Finally, illustrations of ${}_s\mathcal{P}lots$ are provided with simulated events (Section 4.4) and an application for branching ratios measurements (Section 4.5) is briefly described.

To provide some intuitive understanding of how and why the ${}_s\mathcal{P}lot$ formalism works, the problem of reconstructing the true distributions is reconsidered in Appendix A, in a simpler analysis framework. An extension of the ${}_s\mathcal{P}lot$ technique is presented in Appendix B.

2 Basics and definitions

A common method used to extract parameters from a data sample is the maximum Likelihood method which is briefly reviewed in Section 2.1 since it constitutes the foundation of the ${}_s\mathcal{P}lot$ technique. Section 2.2 discusses the need for checks of an analysis based on the Likelihood method and introduces more precisely the goal of the ${}_s\mathcal{P}lot$ technique.

2.1 Likelihood method

One considers an extended Likelihood analysis of a data sample in which are merged several species of events. These species represent various signal components (ie. sources of events in which one is interested) and background components (ie. irrelevant sources of events accompagnying the signal components) which all together account for the data

sample. The log-Likelihood is expressed as:

$$\mathcal{L} = \sum_{e=1}^{N} \ln \Big\{ \sum_{i=1}^{\mathrm{N_s}} N_i \mathrm{f}_i(y_e) \Big\} - \sum_{i=1}^{\mathrm{N_s}} N_i \ , \tag{1}$$

where

- N is the total number of events in the data sample,
- $\mathrm{N_s}$ is the number of species of events populating the data sample,
- N_i is the number of events expected on the average for the i^{th} species,
- y is the set of discriminating variables,
- f_i is the Probability Density Function (PDF) of the discriminating variables for the i^{th} species,
- $\mathrm{f}_i(y_e)$ denotes the value taken by the PDFs f_i for event e, the later being associated with a set of values y_e for the set of discriminating variables,
- x is the set of control variables which, by definition, do not appear in the above expression of $\mathcal{L}$.

The log-Likelihood $\mathcal{L}$ is a function of the $\mathrm{N_s}$ yields N_i and, possibly, of implicit free parameters designed to tune the PDFs on the data sample. These parameters as well as the yields N_i are determined by maximizing the above log-Likelihood.

2.2 Analysis Validation

The crucial point for such an analysis of the data sample to be reliable is to use an exhaustive list of sources of events combined with an accurate description of all the PDFs f_i.

To assess the quality of the fit, one may rely on an evaluation of the goodness of fit based on the actual value obtained for the maximum of $\mathcal{L}$, but this is rarely convincing enough. A complementary quality check is to explore further the data sample by examining the distributions of control variables. If the distributions of these control variables are known for at least one of the sources of events, one can compare the expected distribution for this source to the one extracted from the data sample. In order to do so, one must be able to unfold from the distribution of the whole data sample, the contribution arising from the source under scrutiny.

In some instances of control variables, the PDF might even be known for all the sources of events. Such a control variable can be obtained for instance by removing one of the discriminating variables from the set y before performing again the maximum Likelihood fit, and considering the removed variable as a control variable x. Another example is provided by a discriminating variable for which the distributions are known for all sources of events, but which does not improve significantly the accuracy fo the fit, and is not incorporated in the set y, for the sake of simplicity.

In an attempt to have access to the distributions of control variables, a common method consists in applying cuts which are designed to enhance the contributions to the data sample of particular sources of events (typically of signal species). Having enforced this enhancement, the distribution of x for the reduced data sample can be used to probe

the quality of the fit through a comparison with a Monte Carlo simulated distribution. However, the result is frequently unsatisfactory: firstly because it can be used only if the signal has prominent features to be distinguished from the background, and secondly because of the cuts applied, a sizeable fraction of signal events can be lost, while a large fraction of background events may remain. Therefore, the resulting data distribution concerns a reduced subsample for which statistical fluctuations, or true anomalies, cannot be attributed unambiguously, neither to the signal, nor to the background. For example, one can be tempted to misinterpret an anomaly in the distribution of x coming from the signal as a harmless background fluctuation.

The aim of the ${}_s\mathcal{P}lot$ formalism developed in this paper is to provide a convenient method to unfold the overall distribution of a mixed sample of events in a control variable x into the sub-distributions of the various species which compose the sample. It is a statistical technique which allows to keep all signal events while getting rid of all background events, and keeping track of the statistical uncertainties per bin.

More formally, one is interested in the true distribution (denoted in boldface $\mathbf{M}_{\rm n}(x)$) of a control variable x for events of the $\rm n^{th}$ species, the later being any one of the $\rm N_s$ signal and background species. The purpose of this paper is to demonstrate that one can reconstruct $\mathbf{M}_{\rm n}(x)$ from the sole knowledge of the PDFs of the discriminating variables $\rm f_i$, the first step being to proceed to the maximum Likelihood fit to extract the yields N_i.

As an introduction, in Section 3, the case is considered where the variable x actually belongs to the set of y discriminating variables. That is to say that one makes the assumption opposite to the interesting one: x is assumed to be totally correlated with y. Because of this total correlation, there exists a function of the y parameters which fully determines the 'control' variable, $x = x(y)$. In that case, while performing the fit, an *a priori* knowledge of the x-distributions is implicitly used, thus x cannot play the role of a control variable. Although the technique presented in the following Section is inadequate, it provides a natural first step towards ${}_s\mathcal{P}lot$.

Section 4, dedicated to the ${}_s\mathcal{P}lot$ formalism, treats the interesting case, where x is truly a control variable uncorrelated with y. In that case, while performing the fit, no *a priori* knowledge of the x-distributions is used.

3 First step towards ${}_s\mathcal{P}lot$: ${}_{\rm in}\mathcal{P}lot$

In this Section, one is considering a variable x which can be expressed as a function of the discriminating variables y used in the fit. A fit having been performed to determine the yields N_i for all species, from the knowledge of the PDFs $\rm f_i$ and of the values of the N_i, one can define naively, for all events, the weight [1]

$$\mathcal{P}_{\rm n}(y_e) = \frac{N_{\rm n}{\rm f}_{\rm n}(y_e)}{\sum_{k=1}^{\rm N_s} N_k {\rm f}_k(y_e)} \ , \tag{2}$$

[1] It was pointed out to the authors that a weight similar to the naive one of Eq. (2) was introduced long ago in [1].

which can be used to build the x-distribution $\tilde{\mathrm{M}}_{\mathrm{n}}$ defined by:

$$N_{\mathrm{n}}\tilde{\mathrm{M}}_{\mathrm{n}}(\bar{x})\delta x \;\equiv\; \sum_{e\subset\delta x}\mathcal{P}_{\mathrm{n}}(y_e)\;, \tag{3}$$

where the sum $\sum_{e\subset\delta x}$ runs over the $N_{\delta x}$ events for which x_e (i.e. the value taken by the variable x for event e) lies in the x-bin centered on $\bar{x}$ and of total width δx.

In other words, $N_{\mathrm{n}}\tilde{\mathrm{M}}_{\mathrm{n}}(\bar{x})\delta x$ is the x-distribution obtained by histogramming events, using the weight of Eq. (2).

This procedure reproduces, on average, the true distribution $\mathbf{M}_{\mathrm{n}}(x)$. In effect, on average, one can replace the sum in Eq. (3) by the integral

$$\left\langle \sum_{e\subset\delta x} \right\rangle \longrightarrow \int \mathrm{d}y \sum_{j=1}^{\mathrm{N_s}} N_j \mathrm{f}_j(y)\delta(x(y)-\bar{x})\delta x\;. \tag{4}$$

Similarly, identifying the number of events N_i as determined by the fit to be the expected number of events, one obtains:

$$\begin{aligned}
\left\langle N_{\mathrm{n}}\tilde{\mathrm{M}}_{\mathrm{n}}(\bar{x})\right\rangle &= \int \mathrm{d}y \sum_{j=1}^{\mathrm{N_s}} N_j \mathrm{f}_j(y)\delta(x(y)-\bar{x})\mathcal{P}_{\mathrm{n}}(y) \\
&= \int \mathrm{d}y \sum_{j=1}^{\mathrm{N_s}} N_j \mathrm{f}_j(y)\delta(x(y)-\bar{x})\frac{N_{\mathrm{n}}\mathrm{f}_{\mathrm{n}}(y)}{\sum_{k=1}^{\mathrm{N_s}} N_k \mathrm{f}_k(y)} \\
&= N_{\mathrm{n}}\int \mathrm{d}y\delta(x(y)-\bar{x})\mathrm{f}_{\mathrm{n}}(y) \\
&\equiv N_{\mathrm{n}}\mathbf{M}_{\mathrm{n}}(\bar{x})\;.
\end{aligned} \tag{5}$$

Therefore, the sum over events of the naive weight $\mathcal{P}_{\mathrm{n}}$ provides a direct estimate of the x-distribution of events of the n^{th} species. Plots obtained that way are referred to as ${}_{\mathrm{in}}\mathcal{P}lots$: they provide a correct means to reconstruct $\mathbf{M}_{\mathrm{n}}(x)$ only insofar as the variable considered is **in** the set of discriminating variables y. These ${}_{\mathrm{in}}\mathcal{P}lots$ suffer from a major drawback: x being correlated to y, the PDFs of x enter implicitly in the definition of the naive weight, and as a result, the $\tilde{\mathrm{M}}_{\mathrm{n}}$ distributions cannot be used easily to assess the quality of the fit, because these distributions are biased in a way difficult to grasp, when the PDFs $\mathrm{f}_i(y)$ are not accurate. For example, let us consider a situation where, in the data sample, some events from the n^{th} species show up far in the tail of the $\mathrm{M}_{\mathrm{n}}(x)$ distribution which is implicitly used in the fit. The presence of such events implies that the true distribution $\mathbf{M}_{\mathrm{n}}(x)$ must exhibit a tail which is not accounted for by $\mathrm{M}_{\mathrm{n}}(x)$. These events would enter in the reconstructed ${}_{\mathrm{in}}\mathcal{P}lot$ $\tilde{\mathrm{M}}_{\mathrm{n}}$ with a very small weight, and they would thus escape detection by the above procedure: $\tilde{\mathrm{M}}_{\mathrm{n}}$ would be close to M_{n}, the distribution assumed for x. Only a mismatch in the core of the x-distribution can be revealed with ${}_{\mathrm{in}}\mathcal{P}lots$. Stated differently, the error bars which can be attached to each individual bin of $\tilde{\mathrm{M}}_{\mathrm{n}}$ cannot account for the systematical bias inherent to the ${}_{\mathrm{in}}\mathcal{P}lots$.

4 The ${}_{s}\mathcal{P}lot$ technique

It was shown in the previous Section that if the 'control' variable x belongs to the set y of discriminating variables, one can reconstruct the expected distribution of x with ${}_{\mathrm{in}}\mathcal{P}lots$.

However, the ${}_{\rm in}\mathcal{P}lots$ are not easy to decipher because knowledge of the x distribution enters in their construction.

In this Section is considered the more interesting case where the variable x is truly a control variable, i.e. where x does not belong to y. More precisely, the two sets of variables x and y are assumed to be uncorrelated: hence, the total PDFs $\mathrm{f}_i(x,y)$ all factorize into products $\mathbf{M}_i(x)\mathrm{f}_i(y)$.

4.1 The ${}_s\mathcal{P}lot$ formalism

One may still consider the above distribution $\tilde{\mathrm{M}}_{\rm n}$, but this time the naive weight is no longer satisfactory: as shown below, Eq. (5) does not hold. This is because, when summing over the events, the x-PDFs $\mathbf{M}_j(x)$ appear now on the right hand side of Eq. (4), while they are absent in the Likelihood function. However, a simple redefinition of the weights allows to overcome this difficulty.

Considering the naive weight of Eq. (2):

$$\begin{aligned}
\left\langle N_{\rm n}\tilde{\mathrm{M}}_{\rm n}(\bar{x})\right\rangle &= \int\int \mathrm{d}y\mathrm{d}x \sum_{j=1}^{\mathrm{N_s}} N_j\mathbf{M}_j(x)\mathrm{f}_j(y)\delta(x-\bar{x})\mathcal{P}_{\rm n} \\
&= \int \mathrm{d}y \sum_{j=1}^{\mathrm{N_s}} N_j\mathbf{M}_j(\bar{x})\mathrm{f}_j(y)\frac{N_{\rm n}\mathrm{f}_{\rm n}(y)}{\sum_{k=1}^{\mathrm{N_s}} N_k\mathrm{f}_k(y)} \\
&= N_{\rm n}\sum_{j=1}^{\mathrm{N_s}} \mathbf{M}_j(\bar{x})\left(N_j\int \mathrm{d}y\frac{\mathrm{f}_{\rm n}(y)\mathrm{f}_j(y)}{\sum_{k=1}^{\mathrm{N_s}} N_k\mathrm{f}_k(y)}\right) \qquad (6)\\
&\neq N_{\rm n}\mathbf{M}_{\rm n}(\bar{x})\ . \qquad (7)
\end{aligned}$$

Indeed, as announced, the previous procedure does not apply. In effect, the correction term appearing in Eq. (6)

$$N_j\int \mathrm{d}y\frac{\mathrm{f}_{\rm n}(y)\mathrm{f}_j(y)}{\sum_{k=1}^{\mathrm{N_s}} N_k\mathrm{f}_k(y)} \qquad (8)$$

is not identical to the kroenecker symbol $\delta_{j\rm n}$. The ${}_{\rm in}\mathcal{P}lot$ distribution $N_{\rm n}\tilde{\mathrm{M}}_{\rm n}$ obtained using the naive weight is a linear combination of the true distributions $\mathbf{M}_j$. Only if the y variable was totally discriminating would one recover the correct answer. In effect, for a total discrimination, $\mathrm{f}_{j\neq\rm n}(y)$ vanishes if $\mathrm{f}_{\rm n}(y)$ is non zero. Thus, the product $\mathrm{f}_{\rm n}(y)\mathrm{f}_j(y)$ is equal to $\mathrm{f}^2_{\rm n}(y)\delta_{j\rm n}$, and one gets:

$$N_j\delta_{j\rm n}\int \mathrm{d}y\frac{\mathrm{f}^2_{\rm n}(y)}{N_{\rm n}\mathrm{f}_{\rm n}(y)} = \delta_{j\rm n}\ . \qquad (9)$$

But this is purely academic, because, if y was totally discriminating, the obtention of $\mathbf{M}_{\rm n}(x)$ would be straightforward: one would just apply cuts on y to obtain a pure sample of events of the $\mathrm{n}^{\rm th}$ species and plot them to get $\mathbf{M}_{\rm n}(x)$.

However, in the case of interest where y is not totally discriminating, one observes that the correction term is related to the inverse of the covariance matrix, given by the second derivatives of $-\mathcal{L}$, which the analysis minimizes:

$$\mathbf{V}^{-1}_{{\rm n}j} = \frac{\partial^2(-\mathcal{L})}{\partial N_{\rm n}\partial N_j} = \sum_{e=1}^{N}\frac{\mathrm{f}_{\rm n}(y_e)\mathrm{f}_j(y_e)}{(\sum_{k=1}^{\mathrm{N_s}} N_k\mathrm{f}_k(y_e))^2}\ . \qquad (10)$$

On average, replacing the sum over events by an integral (Eq. (4)) the variance matrix reads:

$$\begin{aligned}\left\langle \mathbf{V}_{\mathrm{n}j}^{-1} \right\rangle &= \int\int \mathrm{d}y\mathrm{d}x \sum_{l=1}^{\mathrm{N_s}} N_l \mathbf{M}_l(x) \mathrm{f}_l(y) \frac{\mathrm{f_n}(y)\mathrm{f}_j(y)}{(\sum_{k=1}^{\mathrm{N_s}} N_k \mathrm{f}_k(y))^2} \\ &= \int \mathrm{d}y \sum_{l=1}^{\mathrm{N_s}} N_l \mathrm{f}_l(y) \frac{\mathrm{f_n}(y)\mathrm{f}_j(y)}{(\sum_{k=1}^{\mathrm{N_s}} N_k \mathrm{f}_k(y))^2} \int \mathrm{d}x \mathbf{M}_l(x) \\ &= \int \mathrm{d}y \frac{\mathrm{f_n}(y)\mathrm{f}_j(y)}{\sum_{k=1}^{\mathrm{N_s}} N_k \mathrm{f}_k(y)} \ . \end{aligned} \tag{11}$$

Therefore, Eq. (6) can be rewritten:

$$\left\langle \tilde{\mathrm{M}}_{\mathrm{n}}(\bar{x}) \right\rangle = \sum_{j=1}^{\mathrm{N_s}} \mathbf{M}_j(\bar{x}) N_j \left\langle \mathbf{V}_{\mathrm{n}j}^{-1} \right\rangle \ . \tag{12}$$

Inverting this matrix equation, one recovers the distribution of interest:

$$N_{\mathrm{n}} \mathbf{M}_{\mathrm{n}}(\bar{x}) = \sum_{j=1}^{\mathrm{N_s}} \left\langle \mathbf{V}_{\mathrm{n}j} \right\rangle \left\langle \tilde{\mathrm{M}}_j(\bar{x}) \right\rangle \ . \tag{13}$$

Hence, if the control variable x is uncorrelated with the discriminating variable, the true distribution of x can still be reconstructed using the naive weight of Eq. (2), through a linear combination of the ${}_{\mathrm{in}}\mathcal{P}lots$. This result is better restated as follows. When x does not belong to the set y, the appropriate weight is not given by Eq. (2), but is the covariance-weighted quantity (thereafter called sWeight) defined by:

$$\boxed{{}_s\mathcal{P}_{\mathrm{n}}(y_e) = \frac{\sum_{j=1}^{\mathrm{N_s}} \mathbf{V}_{\mathrm{n}j} \mathrm{f}_j(y_e)}{\sum_{k=1}^{\mathrm{N_s}} N_k \mathrm{f}_k(y_e)}} \ . \tag{14}$$

With this sWeight, the distribution of the control variable x can be obtained from the ${}_s\mathcal{P}lot$ histogram:

$$N_{\mathrm{n}}\ {}_s\tilde{\mathrm{M}}_{\mathrm{n}}(\bar{x})\delta x \ \equiv \ \sum_{e \subset \delta x} {}_s\mathcal{P}_{\mathrm{n}}(y_e) \ , \tag{15}$$

which reproduces, on average, the true distribution:

$$\left\langle N_{\mathrm{n}}\ {}_s\tilde{\mathrm{M}}_{\mathrm{n}}(x) \right\rangle \ = \ N_{\mathrm{n}} \mathbf{M}_{\mathrm{n}}(x) \ . \tag{16}$$

If the control variable x exhibits significant correlation with the discriminating variable y, the ${}_s\mathcal{P}lots$ obtained with Eq. (15) cannot be compared directly with the pure distributions of the various species. In that case, one must proceed to a Monte-Carlo simulation of the procedure to obtain the expected distributions to which the ${}_s\mathcal{P}lots$ should be compared with.

The fact that the matrix $\mathbf{V}_{ij}$ enters in the definition of the sWeights is enlightening, and, as discussed in the next Section, this confers nice properties to the ${}_s\mathcal{P}lots$. But this is not the key point. The key point is that Eq. (6) is a matrix equation which can be inverted

using a numerical evaluation of the matrix based only on data, thanks to Eq. (10). Rather than computing the matrix by this direct sum over the events, on can use the covariance matrix resulting from the fit, but this option is numerically less accurate than the direct computation[2].

4.2 ${}_s\mathcal{P}lot$ Properties

Beside satisfying, on the average, the essential asymptotic property Eq. (16), ${}_s\mathcal{P}lots$ bear properties which hold even under non-asymptotic conditions.

4.2.1 Normalization

The distribution ${}_s\tilde{\mathrm{M}}_\mathrm{n}$ defined by Eq. (15) is guaranteed to be normalized to unity and the sum over the species of the ${}_s\mathcal{P}lots$ reproduces the data sample distribution of the control variable. These two properties are not obvious because, from expression Eq. (14), neither is it obvious that the sum over the x-bins of $N_\mathrm{n}\ {}_s\tilde{\mathrm{M}}_\mathrm{n}\delta x$ is equal to N_n, nor is it obvious that, in each bin, the sum over all species of the expected numbers of events equates to the number of events actually observed. The demonstration uses the three sum rules below.

1. Maximum Likelihood Sum Rule
   The Likelihood Eq. (1) being extremal for N_j, one gets the first sum rule:

$$\sum_{e=1}^{N} \frac{\mathrm{f}_j(y_e)}{\sum_{k=1}^{\mathrm{N_s}} N_k \mathrm{f}_k(y_e)} = 1 \ , \ \forall j \ . \tag{17}$$

2. Variance Matrix Sum Rule
   From Eq. (10) and Eq. (17) one derives:

$$\sum_{i=1}^{\mathrm{N_s}} N_i \mathbf{V}^{-1}_{ij} \ = \ \sum_{i=1}^{\mathrm{N_s}} N_i \sum_{e=1}^{N} \frac{\mathrm{f}_i(y_e)\mathrm{f}_j(y_e)}{(\sum_{k=1}^{\mathrm{N_s}} N_k \mathrm{f}_k(y_e))^2} \ = \ \sum_{e=1}^{N} \frac{\mathrm{f}_j(y_e)}{\sum_{k=1}^{\mathrm{N_s}} N_k \mathrm{f}_k(y_e)} = 1 \ . \tag{18}$$

3. Covariance Matrix Sum Rule
   Multiplying both sides of Eq. (18) by $\mathbf{V}_{jl}$ and summing over j one gets the sum rule:

$$\sum_{j=1}^{\mathrm{N_s}} \mathbf{V}_{jl} \ = \ \sum_{j=1}^{\mathrm{N_s}} \mathbf{V}_{jl} \sum_{i=1}^{\mathrm{N_s}} N_i \mathbf{V}^{-1}_{ij} \ = \ \sum_{i=1}^{\mathrm{N_s}} \left(\sum_{j=1}^{\mathrm{N_s}} \mathbf{V}^{-1}_{ij} \mathbf{V}_{jl} \right) N_i \ = \ \sum_{i=1}^{\mathrm{N_s}} \delta_{il} N_i \ = \ N_l \ . \tag{19}$$

It follows that:

- Each x-distribution is properly normalized (cf. Eq. (17) and Eq. (19)):

$$\sum_{[\delta x]} N_\mathrm{n}\ {}_s\tilde{\mathrm{M}}_\mathrm{n}(x)\delta x \ = \ \sum_{e=1}^{N} {}_s\mathcal{P}_\mathrm{n}(y_e) \ = \ \sum_{e=1}^{N} \frac{\sum_{j=1}^{\mathrm{N_s}} \mathbf{V}_{\mathrm{n}j} \mathrm{f}_j(y_e)}{\sum_{k=1}^{\mathrm{N_s}} N_k \mathrm{f}_k(y_e)} \ = \ \sum_{j=1}^{\mathrm{N_s}} \mathbf{V}_{\mathrm{n}j} \ = \ N_\mathrm{n} \ . \tag{20}$$

[2]Furthermore, when parameters are fitted together with the yields N_j, in order to get the correct matrix, one should take care to perform a second fit, where these parameters are frozen.

- The contributions ${}_s\mathcal{P}_j(y_e)$ add up to the number of events actually observed in each x-bin. In effect, for any event (cf. Eq. (19)) :

$$\sum_{l=1}^{\mathrm{N_s}} {}_s\mathcal{P}_l(y_e) \;=\; \sum_{l=1}^{\mathrm{N_s}} \frac{\sum_{j=1}^{\mathrm{N_s}} \mathbf{V}_{lj}\mathrm{f}_j(y_e)}{\sum_{k=1}^{\mathrm{N_s}} N_k \mathrm{f}_k(y_e)} \;=\; \frac{\sum_{j=1}^{\mathrm{N_s}} N_j \mathrm{f}_j(y_e)}{\sum_{k=1}^{\mathrm{N_s}} N_k \mathrm{f}_k(y_e)} \;=\; 1\ . \tag{21}$$

Therefore, an ${}_s\mathcal{P}lot$ provides a consistent representation of how all events from the various species are distributed in the control variable x. The contributions to the data sample distribution in x from the various species are disentangled according to a fit based on the discriminating variable y, provided x and y are uncorrelated. Summing up the $\mathrm{N_s}$ ${}_s\mathcal{P}lots$, one recovers the data sample distribution in x, and summing up the number of events entering in a ${}_s\mathcal{P}lot$ for a given species, one recovers the yield of the species, as it is provided by the fit.

For instance, if one observes an excess of events for a particular $\mathrm{n^{th}}$ species, in a given x-bin, this excess is effectively accounted for in the number of event N_n resulting from the fit. To remove these events (for whatever reason and by whatever means) implies a corresponding decrease in N_n. It remains to gauge how significant is an anomaly in the x-distribution of the $\mathrm{n^{th}}$ species. This is the subject of the next Section.

4.2.2 Statistical uncertainties

The statistical uncertainty on $N_\mathrm{n}\ {}_s\tilde{\mathrm{M}}_\mathrm{n}(x)\delta x$ can be defined in each bin by

$$\sigma[N_\mathrm{n}\ {}_s\tilde{\mathrm{M}}_\mathrm{n}(x)\delta x] \;=\; \sqrt{\sum_{e\subset\delta x} ({}_s\mathcal{P}_\mathrm{n})^2}\ . \tag{22}$$

The proof that Eq. (22) holds asymptotically goes as follows:

$$\begin{aligned}
\left\langle \left(\sum_{e\subset\delta x} {}_s\mathcal{P}_\mathrm{n} \right)^2 \right\rangle - \left\langle \sum_{e\subset\delta x} {}_s\mathcal{P}_\mathrm{n} \right\rangle^2 &= \langle N_{\delta x}\rangle \left\langle {}_s\mathcal{P}_\mathrm{n}^2 \right\rangle + \langle N_{\delta x}\left(N_{\delta x}-1\right)\rangle \langle {}_s\mathcal{P}_\mathrm{n}\rangle^2 - \langle N_{\delta x}\rangle^2 \langle {}_s\mathcal{P}_\mathrm{n}\rangle^2 \\
&= \langle N_{\delta x}\rangle \left\langle {}_s\mathcal{P}_\mathrm{n}^2 \right\rangle + \left(\left\langle N_{\delta x}^2 \right\rangle - \langle N_{\delta x}\rangle \right) \langle {}_s\mathcal{P}_\mathrm{n}\rangle^2 - \langle N_{\delta x}\rangle^2 \langle {}_s\mathcal{P}_\mathrm{n}\rangle^2 \\
&= \langle N_{\delta x}\rangle \left\langle {}_s\mathcal{P}_\mathrm{n}^2 \right\rangle \\
&\quad + \left(\langle N_{\delta x}\rangle + \langle N_{\delta x}\rangle^2 - \langle N_{\delta x}\rangle \right) \langle {}_s\mathcal{P}_\mathrm{n}\rangle^2 - \langle N_{\delta x}\rangle^2 \langle {}_s\mathcal{P}_\mathrm{n}\rangle^2 \\
&= \langle N_{\delta x}\rangle \left\langle {}_s\mathcal{P}_\mathrm{n}^2 \right\rangle = \left\langle \sum_{e\subset\delta x} ({}_s\mathcal{P}_\mathrm{n})^2 \right\rangle = \left\langle \sigma^2[N_\mathrm{n}\ {}_s\tilde{\mathrm{M}}_\mathrm{n}\delta x] \right\rangle\ .
\end{aligned} \tag{23}$$

The above asymptotic property is completed by the fact that the sum in quadrature of the uncertainties Eq. (22) reproduces the statistical uncertainty on the yield N_n, as it is provided by the fit: $\sigma[N_\mathrm{n}] \equiv \sqrt{\mathbf{V}_{\mathrm{nn}}}$. The sum over the x-bins reads:

$$\begin{aligned}
\sum_{[\delta x]} \sigma^2[N_\mathrm{n}\ {}_s\tilde{\mathrm{M}}_\mathrm{n}\delta x] &= \sum_{[\delta x]} \sum_{e\subset\delta x} ({}_s\mathcal{P}_\mathrm{n})^2 = \sum_{e=1}^{N} \left(\frac{\sum_{j=1}^{\mathrm{N_s}} \mathbf{V}_{\mathrm{n}j}\mathrm{f}_j(y_e)}{\sum_{k=1}^{\mathrm{N_s}} N_k \mathrm{f}_k(y_e)} \right)^2 \\
&= \sum_{j=1}^{\mathrm{N_s}} \sum_{l=1}^{\mathrm{N_s}} \mathbf{V}_{\mathrm{n}l}\mathbf{V}_{\mathrm{n}j} \sum_{e=1}^{N} \frac{\mathrm{f}_l(y_e)\mathrm{f}_j(y_e)}{(\sum_{k=1}^{\mathrm{N_s}} N_k \mathrm{f}_k(y_e))^2}
\end{aligned}$$

$$= \sum_{j=1}^{\mathrm{N_s}}\sum_{l=1}^{\mathrm{N_s}} \mathbf{V}_{\mathrm{n}l}\mathbf{V}_{\mathrm{n}j}\mathbf{V}^{-1}_{lj} = \sum_{l=1}^{\mathrm{N_s}} \mathbf{V}_{\mathrm{n}l}\delta_{\mathrm{n}l}$$
$$= \mathbf{V}_{\mathrm{nn}} \ , \tag{24}$$

and more generally, the whole covariance matrix is reproduced:

$$\sum_{e=1}^{N}({}_s\mathcal{P}_i)({}_s\mathcal{P}_j) = \mathbf{V}_{ij} \ . \tag{25}$$

Therefore, for the expected number of events per x-bin indicated by the ${}_s\mathcal{P}lots$, the statistical uncertainties are straightforward to compute using Eq. (22). The later expression is asymptotically correct, and it provides a consistent representation of how the overall uncertainty on N_{n} is distributed in x among the events of the n^{th} species. Because of Eq. (25), and since the determination of the yields is optimal when obtained using a Likelihood fit, one can conclude that the ${}_s\mathcal{P}lot$ technique is itself an optimal method to reconstruct distributions of control variables. [3]

4.2.3 Merging ${}_s\mathcal{P}lots$

As a result of the above, two species i and j can be merged into a single species $(i+j)$ without having to repeat the fit and recompute the sWeights. The ${}_s\mathcal{P}lot$ of the merged species is just the sum of the two ${}_s\mathcal{P}lots$ obtained by adding the sWeights on an event-by-event basis:

$$N_{(i+j)}\tilde{\mathrm{M}}_{(i+j)}\delta x = \sum_{e\subset\delta x}({}_s\mathcal{P}_i + {}_s\mathcal{P}_j) \ . \tag{28}$$

The resulting ${}_s\mathcal{P}lot$ has the proper normalization and the proper error bars (Eqs. (20) and (25)):

$$N_{(i+j)} = \sum_{e=1}^{N}({}_s\mathcal{P}_i + {}_s\mathcal{P}_j) = N_i + N_j \tag{29}$$

$$\sigma^2[N_{(i+j)}] = \sum_{e=1}^{N}({}_s\mathcal{P}_i + {}_s\mathcal{P}_j)^2$$
$$= \mathbf{V}_{ii} + \mathbf{V}_{jj} + 2\mathbf{V}_{ij} = \mathbf{V}_{(i+j)(i+j)} \ . \tag{30}$$

[3] This is not the case for ${}_{\mathrm{in}}\mathcal{P}lots$ for which one gets:

$$\sum_{e=1}^{N}(\mathcal{P}_i)(\mathcal{P}_j) = N_iN_j\mathbf{V}^{-1}_{ij} \ . \tag{26}$$

Hence, using the fact, that contrary to sWeights, the ${}_{\mathrm{in}}\mathcal{P}lot$ weights of Eq. (2) are positive definite, one gets:

$$\sum_{e=1}^{N}(\mathcal{P}_i)^2 \le \sum_{e=1}^{N}(\mathcal{P}_i)(\sum_{j=1}^{\mathrm{N_s}}\mathcal{P}_j) = N_i\sum_{j=1}^{\mathrm{N_s}}N_j\mathbf{V}^{-1}_{ij} = N_i \le \mathbf{V}_{ii} \ . \tag{27}$$

That is to say that the statistical uncertainties attached to the ${}_{\mathrm{in}}\mathcal{P}lots$ are always not only smaller than the ones resulting from the fit, but even smaller than the statistical uncertainties obtained in a backgound free situation.

4.3 ${}_s\mathcal{P}lot$ implementation

This Section is meant to show that using ${}_s\mathcal{P}lot$ is indeed easy. The different steps to implement the technique are the following:

1. One is dealing with a data sample in which several species of events are present.
2. A maximum Likelihood fit is performed to obtain the yields N_i of the various species. The fit relies on a discriminating variable y uncorrelated with a control variable x: the later is therefore totally absent from the fit.
3. The sWeights ${}_s\mathcal{P}$ are calculated using Eq. (14) where the covariance matrix is obtained by inverting the matrix given by Eq. (10).
4. Histograms of x are filled by weighting the events with the sWeights ${}_s\mathcal{P}$. The sum of the entries are equal to the yields N_i provided by the fit.
5. Error bars per bin are given by Eq. (22). The sum of the error bars squared are equal to the uncertainties squared $\mathbf{V}_{ii}$ provided by the fit.
6. The ${}_s\mathcal{P}lots$ reproduce the true distributions of the species in the control variable x, within the above defined statistical uncertainties.

The ${}_s\mathcal{P}lot$ method has been implemented in the ROOT framework under the class TSPlot [2].

4.4 Illustrations

To illustrate the technique, one considers in this Section an example derived from the analysis where ${}_s\mathcal{P}lots$ have been first used [3] and [4] (but see also [5]). One is dealing with a data sample in which two species are present: the first is termed signal and the second background. A maximum Likelihood fit is performed to obtain the two yields N_1 and N_2. The fit relies on two discriminating variables collectively denoted y which are chosen within three possible variables denoted (following the notations of [3]) $m_{\rm ES}$, ΔE and $\mathcal{F}$. The variable which is not incorporated in y is used as a control variable x. The six distributions of the three variables are assumed to be the ones depicted in Fig. 1.

A data sample being built through a Monte Carlo simulation based on the distributions shown in Fig. 1, one obtains the three distributions of Fig. 2. Whereas the distribution of ΔE clearly indicates the presence of the signal, the distribution of $m_{\rm ES}$ and $\mathcal{F}$ are less obviously populated by signal.

Chosing ΔE and $\mathcal{F}$ as discriminating variables to determine N_1 and N_2 through a maximum Likelihood fit, one builds, for the control variable $m_{\rm ES}$ which is unknown to the fit, the two ${}_s\mathcal{P}lots$ for signal and background shown in Fig. 3. For comparison, the PDFs of $m_{\rm ES}$ taken from Fig. 1 are superimposed on the ${}_s\mathcal{P}lots$. One observes that the ${}_s\mathcal{P}lot$ for signal reproduces correctly the PDF even where the latter vanishes, although the error bars remain sizeable. This results from the almost complete cancellation between positive and negative sWeights: the sum of sWeights is close to zero in the tails while the sum of sWeights squared is not. The occurence of negative sWeights is provided through the appearance of the covariance matrix, and its negative components, in the definition of Eq. (14).

A word of caution is in order with respect to the error bars. Whereas their sum in quadrature is identical to the statistical uncertainties of the yields determined by the

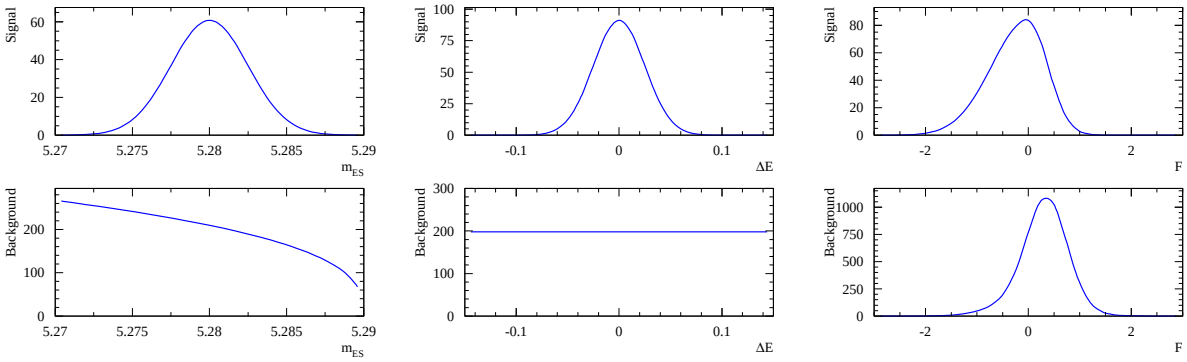

Figure 1: Distributions of the three different discriminating variables available to perform the Likelihood fit: $m_{\rm ES}$, ΔE, $\mathcal{F}$. Among the three variables, two are used to perform the fit while one is kept out of the fit to serve the purpose of a control variable. The three distributions on the top (resp. bottom) of the figure correspond to the signal (resp. background). The unit of the vertical axis is chosen such that it indicates the number of entries per bin, if one slices the histograms in 25 bins.

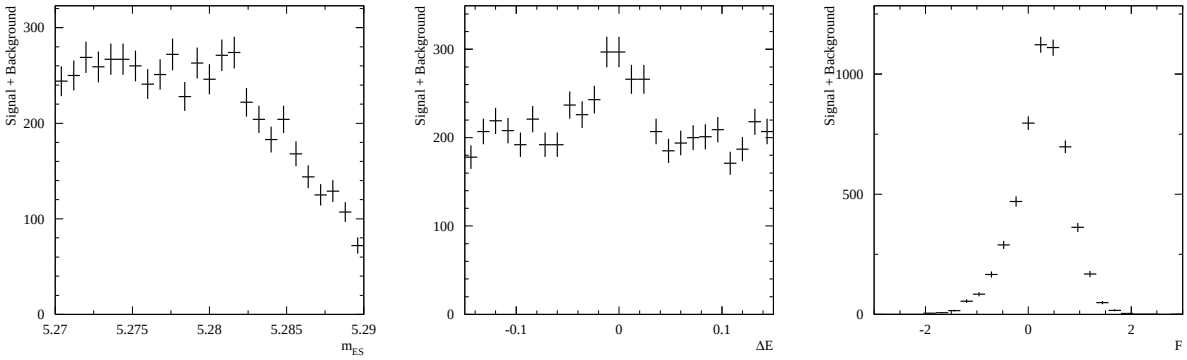

Figure 2: Distributions of the three discriminating variables for signal plus background. The three distributions are the one obtained from a data sample obtained through a Monte Carlo simulation based on the distributions shown in Fig. 1. The data sample consists of 500 signal events and 5000 background events.

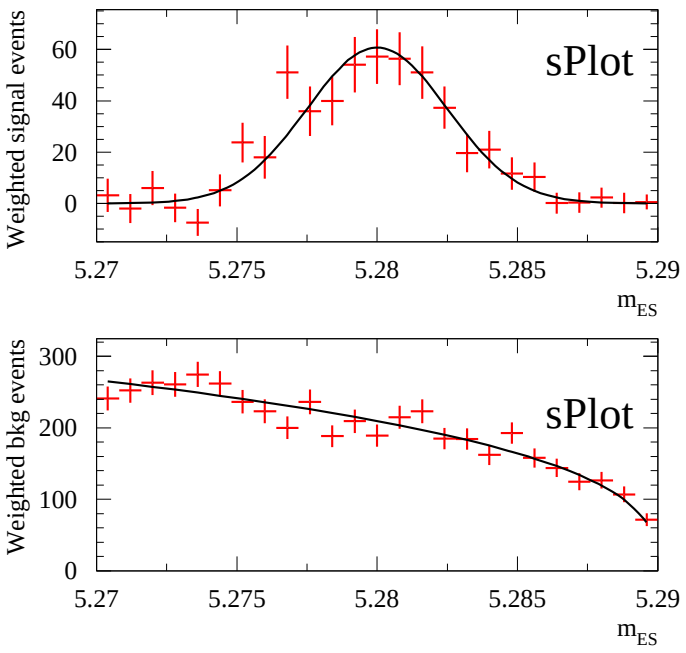

Figure 3: The ${}_s\mathcal{P}lots$ (signal on top, background on bottom) obtained for $m_{\rm ES}$ are represented as dots with error bars. They are obtained from a fit using only information from ΔE and $\mathcal{F}$. The black curves are the PDFs of $m_{\rm ES}$ of Fig. 1: these PDFs are unknown to the fit.

fit, and if, in addition, they are asymptotically correct (cf. Section 4.2.2) the error bars should be handled with care for low statistics and/or for too fine binning. This is because the error bars do not incorporate two known properties of the PDFs: PDFs are positive definite and can be non-zero in a given x-bin, even if in the particular data sample at hand, no event is observed in this bin. The latter limitation is not specific to ${}_s\mathcal{P}lots$, rather it is always present when one is willing to infer the PDF at the origin of an histogram, when, for some bins, the number of entries does not guaranty the applicability of the Gaussian regime. In such situations, a satisfactory practice is to attach allowed ranges to the histogram to indicate the upper and lower limits of the PDF value which are consistent with the actual observation, at a given confidence level. Although this is straightforward to implement, even when dealing with sWeighted events, for the sake of simplicity, this subject is not discussed further in the paper.

Chosing $m_{\rm ES}$ and ΔE as discriminating variables to determine N_1 and N_2 through a maximum Likelihood fit, one builds, for the control variable $\mathcal{F}$ which is unknown to the fit, the two ${}_s\mathcal{P}lots$ for signal and background shown in Fig. 4. For comparison, the PDFs of $\mathcal{F}$ taken from Fig. 1 are superimposed on the ${}_s\mathcal{P}lots$. In the ${}_s\mathcal{P}lot$ for signal one observes that error bars are the largest in the x regions where the background is the largest.

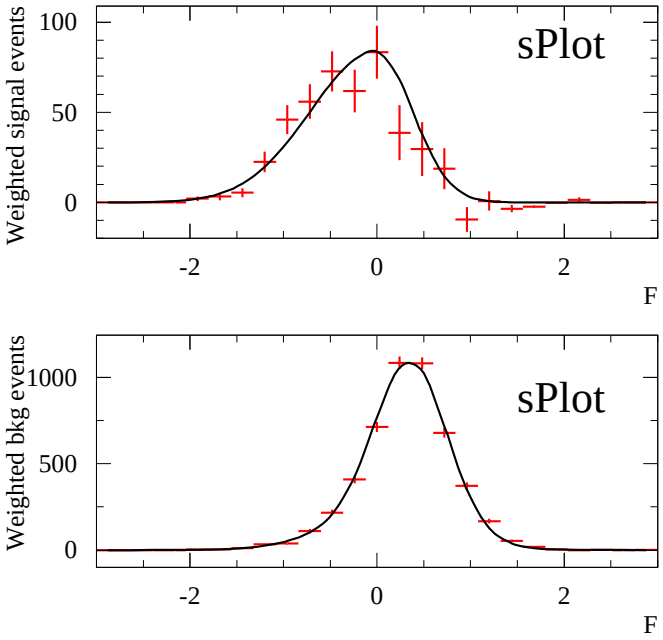

Figure 4: The ${}_s\mathcal{P}lots$ (signal on top, background on bottom) obtained for $\mathcal{F}$ are represented as dots with error bars. They are obtained from a fit using only information from $m_{\rm ES}$ and ΔE. The black curves are the PDFs of $\mathcal{F}$ of Fig. 1: these PDFs are unknown to the fit.

4.5 Application: efficiency corrected yields

Beside providing a convenient and optimal tool to cross-check the analysis by allowing distributions of control variables to be reconstructed and then compared with expectations, the ${}_s\mathcal{P}lot$ formalism can be applied also to extract physics results, which would otherwise be difficult to obtain. For example, one may be willing to explore some unknown physics involved in the distribution of a variable x. Or, one may be interested to correct a particular yield provided by the Likelihood fit from a selection efficiency which is known to depend on a variable x, for which the PDF is unknown.

To be specific, one can take the example of a three body decay analysis of a species, the signal, polluted by background, while the signal PDF inside the two-dimensional Dalitz plot is not known, because of unknown contributions of resonances, continuum and an interference pattern. Since the x-dependence of the selection efficiency $\epsilon(x)$ can be computed without *a priori* knowledge of the x-distributions, one can build the efficiency corrected two-dimensional ${}_s\mathcal{P}lots$ (cf. Eq. (15)):

$$\frac{1}{\epsilon(\bar{x})} N_{\rm n}\ {}_s\tilde{\rm M}_{\rm n}(\bar{x})\delta x = \sum_{e\subset\delta x} \frac{1}{\epsilon(x_e)}\, {}_s\mathcal{P}_{\rm n}(y_e)\ , \tag{31}$$

and compute the efficiency corrected yields:

$$N^{\epsilon}_{\rm n} = \sum_{e=1}^{N} \frac{{}_s\mathcal{P}_{\rm n}(y_e)}{\epsilon(x_e)}\ . \tag{32}$$

Analyses can then use the ${}_s\mathcal{P}lot$ formalism for validation purposes, but also, using Eq. (31) and Eq. (32), to probe for resonance structures and to measure branching ratios.

5 Conclusion

The technique presented in this paper applies when one examines a data sample originating from different sources of events: using a set y of discriminating variables, a Likelihood fit is performed on the data sample to determine the yields of the sources. By building ${}_s\mathcal{P}lots$, one can reconstruct the distributions of variables, separately for each source present in the data sample, provided the variables are uncorrelated with the set y used in the fit. Although no cut is applied (hence, the ${}_s\mathcal{P}lot$ of a given species represents the whole statistics of this species) the distributions obtained are pure (free from the potential background arising from the other species) in a statistical sense. The more discriminating the discriminating variables y, the clearer the ${}_s\mathcal{P}lot$ is. The technique is straightforward to implement and features several nice properties: both the normalizations and the statistical uncertainties of the ${}_s\mathcal{P}lots$ reflect the fit ouputs.

References

[1] P. E. Condon and P. L. Cowell, Phys. Rev. D 9 (1974) 2558-2562

[2] http://root.cern.ch/

[3] M. Pivk, Thèse de l'Université Paris VII, BABAR-THESIS-03/012 (2003), available (in French) at http://tel.ccsd.cnrs.fr (ID 00002991)

[4] The BABAR Collaboration, Phys.Rev.Lett.93 (2004) 131801

[5] The BABAR Collaboration, Phys.Rev.Lett.93 (2004) 181805;
The BABAR Collaboration, Phys.Rev.D70 (2004) 091103;
The BABAR Collaboration, Phys.Rev.Lett.94 (2005) 181802

A Pedagogical examples

The purpose of this Appendix is to detail in simplified situations how and why ${}_s\mathcal{P}lot$ works. One begins with the simplest situation and proceed to more complex ones.

A.1 Simple cut-and-count analysis

In this Section, a very simple situation is considered where the proper way to reconstruct signal and background distributions for a control variable x is obvious from the start. The purpose is to observe the ${}_s\mathcal{P}lot$ technique at work, when one knows beforehand what the outcome should be.

One considers a data sample consisting of $\mathrm{N_s} = 2$ species: species 1 is referred to as the signal and species 2 as the background. A unique discriminating variable $y \in [0, 1]$ is used in the fit. One further assumes that:

- the signal distribution is the step-function:

$$\begin{aligned} \mathrm{f}_1(y < y_0) &= 0 & (33) \\ \mathrm{f}_1(y \geq y_0) &= (1 - y_0)^{-1} \ , & (34) \end{aligned}$$

- the background distribution is uniform in the full range:

$$\mathrm{f}_2(y) = 1 \ . \tag{35}$$

Therefore, one is dealing with a cut-and-count analysis: there is a pure background sideband for $y < y_0$, and the shapes of the signal and background distributions offer no discriminating power in the region where the signal is present, for $y \geq y_0$. Denoting N the total number of events present in the data sample, $N_<$ the number of events located below y_0, and $N_>$ the number of events located above y_0:

1. the expected number of background and signal events can be deduced without any fit, from the sideband:

$$\begin{aligned} N_2 &= \frac{1}{y_0} N_< & (36) \\ N_1 &= -\frac{1 - y_0}{y_0} N_< + N_> \ , & (37) \end{aligned}$$

2. $N_<$ and $N_>$ being two independent numbers of events, the covariance matrix can be deduced directly from Eqs. (36)-(37):

$$\mathbf{V} = \begin{pmatrix} N_> + \left(\dfrac{1 - y_0}{y_0}\right)^2 N_< & -\dfrac{1 - y_0}{y_0^2} N_< \\ -\dfrac{1 - y_0}{y_0^2} N_< & \dfrac{1}{y_0^2} N_< \end{pmatrix} \ , \tag{38}$$

3. denoting $\delta N^x_<$ the number of events in a given x-bin, with $y \leq y_0$, the background distribution $\mathrm{M}_2(x)$ can also be deduced by a mere rescaling of $\delta N^x_<$, as in Eq. (36):

$$N_2 \ \mathrm{M}_2(x)\delta x = \frac{\delta N^x_<}{y_0} \ . \tag{39}$$

Similarly to Eq. (37), the signal distribution is given by:

$$N_1 \ \mathrm{M}_1(x)\delta x = -(1-y_0)N_2 \ \mathrm{M}_2(x) + \delta N^x_> \ , \tag{40}$$

that is to say, one can obtain the signal distribution from the (mixed) events populating the domain $y \geq y_0$, if one subtracts the contribution of background events, which is known from Eq. (39). Stated differently, one is lead to assign the negative weight $-(1-y_0)/y_0$ to those events in the x-bin which satisfy $y \leq y_0$.

Whereas in such a simple situation the use of the ${}_s\mathcal{P}lot$ formalism would be awkward, the latter should reproduce the above obvious results, and indeed it does. The proof goes as follows:

1. denoting $\mathrm{f}_i(0)$ (resp. $\mathrm{f}_i(1)$) the value taken by the PDF of species i for $y \leq y_0$ (resp. $y > y_0$), Eq. (17) reads:

$$\begin{aligned} 1 &= \sum_{e=1}^{N} \frac{\mathrm{f}_1(y_e)}{\sum_{k=1}^{\mathrm{N_s}} N_k \mathrm{f}_k(y_e)} = N_< \frac{\mathrm{f}_1(0)}{N_1\mathrm{f}_1(0)+N_2\mathrm{f}_2(0)} + N_> \frac{\mathrm{f}_1(1)}{N_1\mathrm{f}_1(1)+N_2\mathrm{f}_2(1)} \\ &= \frac{N_>(1-y_0)^{-1}}{N_1(1-y_0)^{-1}+N_2} \end{aligned} \tag{41}$$

$$\begin{aligned} 1 &= \sum_{e=1}^{N} \frac{\mathrm{f}_2(y_e)}{\sum_{k=1}^{\mathrm{N_s}} N_k \mathrm{f}_k(y_e)} = N_< \frac{\mathrm{f}_2(0)}{N_1\mathrm{f}_1(0)+N_2\mathrm{f}_2(0)} + N_> \frac{\mathrm{f}_2(1)}{N_1\mathrm{f}_1(1)+N_2\mathrm{f}_2(1)} \\ &= \frac{N_<}{N_2} + \frac{N_>}{N_1(1-y_0)^{-1}+N_2} \ . \end{aligned} \tag{42}$$

The first equation yields:

$$N_1(1-y_0)^{-1} + N_2 = N_>(1-y_0)^{-1} \tag{43}$$

and thus, for the second equation:

$$1 = \frac{N_<}{N_2} + 1 - y_0 \ , \tag{44}$$

which leads to Eqs. (36)-(37).

2. similarly, Eq. (10) yields

$$\mathbf{V}^{-1} = \begin{pmatrix} \dfrac{1}{N_>} & \dfrac{1-y_0}{N_>} \\ \dfrac{1-y_0}{N_>} & \dfrac{(1-y_0)^2}{N_>} + \dfrac{y_0^2}{N_<} \end{pmatrix} \ . \tag{45}$$

For example, using Eq. (43), the $\mathbf{V}_{11}$ component is computed as follows:

$$\mathbf{V}_{11}^{-1} \;=\; \sum_{e=1}^{N} \frac{\mathrm{f}_1(y_e)\mathrm{f}_1(y_e)}{(\sum_{k=1}^{\mathrm{N_s}} N_k \mathrm{f}_k(y_e))^2} \;=\; N_> \frac{(1-y_0)^{-2}}{(N_1(1-y_0)^{-1}+N_2)^2} = \frac{1}{N_>}\;. \qquad (46)$$

And similarly for the other components. Inverting $\mathbf{V}^{-1}$ one gets Eq. (38).

3. Eq. (15) then reproduces Eqs. (39)-(40). Namely:

$$\begin{aligned}
N_1\ \mathrm{M}_1(x)\delta x &= \sum_{e\subset\delta x} \frac{\mathbf{V}_{11}\mathrm{f}_1(y_e)+\mathbf{V}_{12}\mathrm{f}_2(y_e)}{\sum_{k=1}^{\mathrm{N_s}} N_k \mathrm{f}_k(y_e)} \\
&= \delta N_<^x \frac{\mathbf{V}_{11}\mathrm{f}_1(0)+\mathbf{V}_{12}\mathrm{f}_2(0)}{N_1\mathrm{f}_1(0)+N_2\mathrm{f}_2(0)} + \delta N_>^x \frac{\mathbf{V}_{11}\mathrm{f}_1(1)+\mathbf{V}_{12}\mathrm{f}_2(1)}{N_1\mathrm{f}_1(1)+N_2\mathrm{f}_2(1)} \\
&= \delta N_<^x \frac{\mathbf{V}_{12}}{N_2} + \delta N_>^x \frac{\mathbf{V}_{11}(1-y_0)^{-1}+\mathbf{V}_{12}}{N_1(1-y_0)^{-1}+N_2} \\
&= \delta N_<^x \frac{-\frac{1-y_0}{y_0^2}N_<}{N_< y_0^{-1}} \\
&\quad +\delta N_>^x \frac{(N_> + (\frac{1-y_0}{y_0})^2 N_<)(1-y_0)^{-1} - \frac{1-y_0}{y_0^2}N_<}{N_>(1-y_0)^{-1}} \\
&= -\frac{1-y_0}{y_0}\delta N_<^x + \delta N_>^x \qquad (47)
\end{aligned}$$

and:

$$\begin{aligned}
N_2\ \mathrm{M}_2(x)\delta x &= \sum_{e\subset\delta x} \frac{\mathbf{V}_{21}\mathrm{f}_1(y_e)+\mathbf{V}_{22}\mathrm{f}_2(y_e)}{\sum_{k=1}^{\mathrm{N_s}} N_k \mathrm{f}_k(y_e)} \\
&= \delta N_<^x \frac{\mathbf{V}_{21}\mathrm{f}_1(0)+\mathbf{V}_{22}\mathrm{f}_2(0)}{N_1\mathrm{f}_1(0)+N_2\mathrm{f}_2(0)} + \delta N_>^x \frac{\mathbf{V}_{21}\mathrm{f}_1(1)+\mathbf{V}_{22}\mathrm{f}_2(1)}{N_1\mathrm{f}_1(1)+N_2\mathrm{f}_2(1)} \\
&= \delta N_<^x \frac{\mathbf{V}_{22}}{N_2} + \delta N_>^x \frac{\mathbf{V}_{21}(1-y_0)^{-1}+\mathbf{V}_{22}}{N_1(1-y_0)^{-1}+N_2} \\
&= \delta N_<^x \frac{\frac{1}{y_0^2}N_<}{N_< y_0^{-1}} \\
&\quad +\delta N_>^x \frac{-\frac{1-y_0}{y_0^2}N_<(1-y_0)^{-1} + \frac{1}{y_0^2}N_<}{N_>(1-y_0)^{-1}} \\
&= \frac{\delta N_<^x}{y_0}\;. \qquad (48)
\end{aligned}$$

4. it can be shown as well that Eqs. (18)-(19)-(20)-(21)-(25) hold.

Therefore, in this very simple situation where the problem of reconstructing the distributions of signal and background events is glaringly obvious, the ${}_s\mathcal{P}lot$ formalism reproduces the expected results.

A.2 Extended cut-and-count analysis

The above example of the previous Section A.1 is a very particular case of a more general situation where the y-range is split into n_y slices inside which one disregards the shape of the distributions of the species, whether these distributions are the same or not. Using greek letters to index the y-slices, this amounts to replacing the $\mathrm{f}_i(y)$ PDFs by step functions with constant values. For each y-bin F_i^α, these constant values are defined by the integral over the y-bin α:

$$\mathrm{f}_i(y) \to \mathrm{F}_i^\alpha = \int_\alpha \mathrm{f}_i(y)\mathrm{d}y \tag{49}$$

$$\sum_{\alpha=1}^{n_y} \mathrm{F}_i^\alpha = 1 \ . \tag{50}$$

With this notation, the number of events $\bar{N}_\alpha$ expected in the slice α is given by:

$$\bar{N}_\alpha = \sum_{i=1}^{\mathrm{N_s}} N_i \mathrm{F}_i^\alpha \ . \tag{51}$$

To make particularly obvious what must be the outcome of the ${}_s\mathcal{P}lot$ technique, in the previous Section it was assumed that $n_y = \mathrm{N_s} = 2$, and that the signal was utterly absent in one of the two y-slices: $\mathrm{F}_1^1 = 0$, $\mathrm{F}_1^2 = 1$, $\mathrm{F}_2^1 = y_0$ and $\mathrm{F}_2^2 = 1 - y_0$.

Below one proceeds in two steps, first considering the more general case where only $n_y = \mathrm{N_s}$ is assumed (Section A.2.1), then considering the extended cut-and-count analysis where $n_y > \mathrm{N_s}$ (Section A.2.2). Since the general case discussed in the presentation of the ${}_s\mathcal{P}lot$ formalism corresponds to the limit $n_y \to \infty$, what follows amounts to a step-by-step new derivation of the technique.

A.2.1 Generalized cut-and-count analysis: $n_y = \mathrm{N_s}$

When the number of y-slices equals the number of species, the solution remains obvious, if the $\mathrm{N_s} \times \mathrm{N_s}$ matrix F_i^α is invertible (if not, the N_i cannot be determined). In that case, one can identify the expected numbers of events $\bar{N}_\alpha$ with the observed number of events N_α, and thus:

1. one recovers the expected number of events N_i from the numbers of events N_α observed in the n_y slice, by inverting Eq. (51):

$$N_i = \sum_{\alpha=1}^{\mathrm{N_s}} N_\alpha (\mathrm{F}^{-1})_i^\alpha \ , \tag{52}$$

2. the number N_α being statistically independent, one obtains directly the covariance matrix:

$$\mathbf{V}_{ij} = \sum_{\alpha=1}^{\mathrm{N_s}} N_\alpha (\mathrm{F}^{-1})_i^\alpha (\mathrm{F}^{-1})_j^\alpha \ , \tag{53}$$

3. similarly to Eq. (51), the number of events δN_α^x observed in the y-slice α and in the bin x of width δx is given by:

$$\delta N_\alpha^x = \sum_{i=1}^{\mathrm{N_s}} N_i \mathbf{M}_i(x) \delta x \mathrm{F}_i^\alpha \tag{54}$$

and thus, the x-distribution of species i is:

$$\delta N_i^x \equiv N_i \mathbf{M}_i(x) \delta x = \sum_{\alpha=1}^{\mathrm{N_s}} \delta N_\alpha^x (\mathrm{F}^{-1})_i^\alpha \ . \tag{55}$$

It remains to be shown that Eq. (55) is reproduced using the ${}_s\mathcal{P}lot$ formalism. First, using Eq. (49) and Eq. (51), one observes that:

$$\sum_{e=1}^{N} \frac{\mathrm{f}_i(y_e)}{\sum_{k=1}^{\mathrm{N_s}} N_k \mathrm{f}_k(y_e)} \ \rightarrow \ \sum_{\alpha=1}^{\mathrm{N_s}} N_\alpha \frac{\mathrm{F}_i^\alpha}{\sum_{k=1}^{\mathrm{N_s}} N_k \mathrm{F}_k^\alpha} \ = \ \sum_{\alpha=1}^{\mathrm{N_s}} N_\alpha \frac{\mathrm{F}_i^\alpha}{N_\alpha} \ = \ \sum_{\alpha=1}^{\mathrm{N_s}} \mathrm{F}_i^\alpha = 1 \ , \tag{56}$$

which shows that the obvious solution Eq. (52) is the one which maximizes the extended log-Likelihood. Similarly:

$$\mathbf{V}_{ij}^{-1} = \sum_{e=1}^{N} \frac{\mathrm{f}_i(y_e)\mathrm{f}_j(y_e)}{(\sum_{k=1}^{\mathrm{N_s}} N_k \mathrm{f}_k(y_e))^2} \rightarrow \sum_{\alpha=1}^{\mathrm{N_s}} N_\alpha \frac{\mathrm{F}_i^\alpha \mathrm{F}_j^\alpha}{N_\alpha^2} \ = \ \sum_{\alpha=1}^{\mathrm{N_s}} \frac{1}{N_\alpha} \mathrm{F}_i^\alpha \mathrm{F}_j^\alpha \ , \tag{57}$$

which inverse is given by Eq. (53), and thus:

$$N_i \ {}_s\tilde{\mathrm{M}}_i(x) \delta x \rightarrow \sum_{\alpha=1}^{\mathrm{N_s}} \delta N_\alpha^x \frac{\sum_j \mathbf{V}_{ij} \mathrm{F}_j^\alpha}{N_\alpha} \ = \ \sum_{\alpha=1}^{\mathrm{N_s}} \delta N_\alpha^x (\mathrm{F}^{-1})_i^\alpha \ . \tag{58}$$

The ${}_s\mathcal{P}lot$ formalism reproduces Eq. (55).

A.2.2 Extended cut-and-count analysis: $n_y > \mathrm{N_s}$

In the more general situation where the number of y-slices is larger than the number of species, there is no blatant solution neither for determining the N_i, nor for reconstructing the x-distribution of each species (in particular, Eq. (52) is lost). Because of this lack of an obvious solution, what follows is a rephrasing of the derivation of the ${}_s\mathcal{P}lots$, but taking a different point of view, and in the case where the y-distributions are binned.

The best determination of the N_i (here as well as in the previous simpler situations) is provided by the Likelihood method which yields (cf. Eq. (17)):

$$\sum_{\alpha=1}^{n_y} \frac{N_\alpha \mathrm{F}_i^\alpha}{\sum_{k=1}^{\mathrm{N_s}} N_k \mathrm{F}_k^\alpha} = 1 \quad , \forall i \tag{59}$$

with a variance matrix (cf. Eq. (10)):

$$\mathbf{V}_{ij}^{-1} = \sum_{\alpha=1}^{n_y} N_\alpha \frac{\mathrm{F}_i^\alpha \mathrm{F}_j^\alpha}{(\sum_{k=1}^{\mathrm{N_s}} N_k \mathrm{F}_k^\alpha)^2} \ , \tag{60}$$

from which one computes the covariance matrix $\mathbf{V}_{ij}$. Instead of Eq. (52) the number of events N_i provided by Eq. (59) is shown below to satisfy the equality (cf. Eq. (20)):

$$N_i = \sum_{\alpha=1}^{n_y} N_\alpha \ ({}_s\mathcal{P})_i^\alpha \ , \tag{61}$$

where the matrix element $({}_s\mathcal{P})_i^\alpha$ is the sWeight (Eq. (14)) for species i of events with y_e lying in the y-slice α, namely:

$$({}_s\mathcal{P})_i^\alpha \ = \ \frac{\sum_{j=1}^{\mathrm{N_s}} \mathbf{V}_{ij} \mathrm{F}_j^\alpha}{\sum_{k=1}^{\mathrm{N_s}} N_k \mathrm{F}_k^\alpha} \ . \tag{62}$$

The identity of Eq. (61) is not asymptotic, it holds even for finite statistics, since the contractions with $\mathbf{V}_{li}^{-1}$ of both the left- and right-hand sides yield the same result. Indeed (Eq. (18)):

$$\sum_{i=1}^{\mathrm{N_s}} N_i \mathbf{V}_{li}^{-1} \ = \ \sum_{i=1}^{\mathrm{N_s}} \sum_{\alpha=1}^{n_y} \frac{N_\alpha \mathrm{F}_l^\alpha N_i \mathrm{F}_i^\alpha}{(\sum_{k=1}^{\mathrm{N_s}} N_k \mathrm{F}_k^\alpha)^2} \ = \ \sum_{\alpha=1}^{n_y} \frac{N_\alpha \mathrm{F}_l^\alpha (\sum_{i=1}^{\mathrm{N_s}} N_i \mathrm{F}_i^\alpha)}{(\sum_{k=1}^{\mathrm{N_s}} N_k \mathrm{F}_k^\alpha)^2} \ = \ \sum_{\alpha=1}^{n_y} \frac{N_\alpha \mathrm{F}_l^\alpha}{\sum_{k=1}^{\mathrm{N_s}} N_k \mathrm{F}_k^\alpha} \ = \ 1 \ , \tag{63}$$

which is identical to:

$$\sum_{i=1}^{\mathrm{N_s}} \mathbf{V}_{li}^{-1} \sum_{\alpha=1}^{n_y} N_\alpha \frac{\sum_{j=1}^{\mathrm{N_s}} \mathbf{V}_{ij} \mathrm{F}_j^\alpha}{\sum_{k=1}^{\mathrm{N_s}} N_k \mathrm{F}_k^\alpha} \ = \ \sum_{\alpha=1}^{n_y} N_\alpha \frac{\sum_{j=1}^{\mathrm{N_s}} (\sum_{i=1}^{\mathrm{N_s}} \mathbf{V}_{li}^{-1} \mathbf{V}_{ij}) \mathrm{F}_j^\alpha}{\sum_{k=1}^{\mathrm{N_s}} N_k \mathrm{F}_k^\alpha} \ = \ \sum_{\alpha=1}^{n_y} \frac{N_\alpha \mathrm{F}_l^\alpha}{\sum_{k=1}^{\mathrm{N_s}} N_k \mathrm{F}_k^\alpha} \ = \ 1 \ . \tag{64}$$

Since Eq. (61) holds for the complete sample of events, it must hold as well for any sub-sample, provided the splitting into sub-samples is not correlated with the variable y. Namely, for all x-bin, one is guaranteed to observe, on average, the same relationship between the numbers of events δN_i^x and δN_α^x. The ${}_s\mathcal{P}lot$ obtained from the weighted sum

$$\delta N_i^x = \sum_{\alpha=1}^{n_y} \delta N_\alpha^x \ ({}_s\mathcal{P})_i^\alpha \ , \tag{65}$$

is an unbiased estimator of the true distribution of x for species i. One can provide a direct proof that the above ${}_s\mathcal{P}lot$ of Eq. (65) reproduces the true distribution by following the same line which leads to Eq. (12). On average, using successively:

$$\langle \delta N_\alpha^x \rangle = \sum_{l=1}^{\mathrm{N_s}} N_l \mathbf{M}_l(x) \mathrm{F}_l^\alpha \delta x \tag{66}$$

and hence:

$$\langle N_\alpha \rangle \ = \ \sum_x \langle \delta N_\alpha^x \rangle \ = \ \sum_x \mathbf{M}_l(x) \sum_{k=1}^{\mathrm{N_s}} N_k \mathrm{F}_k^\alpha \ = \ \sum_{k=1}^{\mathrm{N_s}} N_k \mathrm{F}_k^\alpha \ , \tag{67}$$

one gets:

$$\left\langle \sum_{\alpha=1}^{n_y} \delta N_\alpha^x \ ({}_s\mathcal{P})_i^\alpha \right\rangle \ = \ \sum_{\alpha=1}^{n_y} \left(\sum_{l=1}^{\mathrm{N_s}} N_l \mathbf{M}_l(x) \mathrm{F}_l^\alpha \delta x \right) \frac{\sum_{j=1}^{\mathrm{N_s}} \mathbf{V}_{ij} \mathrm{F}_j^\alpha}{\sum_{k=1}^{\mathrm{N_s}} N_k \mathrm{F}_k^\alpha}$$

$$
\begin{aligned}
&= \delta x \sum_{l=1}^{\mathrm{N_s}} N_l \mathbf{M}_l(x) \left(\sum_{j=1}^{\mathrm{N_s}} \mathbf{V}_{ij} \sum_{\alpha=1}^{n_y} \frac{\mathrm{F}_j^\alpha \mathrm{F}_l^\alpha}{\sum_{k=1}^{\mathrm{N_s}} N_k \mathrm{F}_k^\alpha} \right) \\
&= \delta x \sum_{l=1}^{\mathrm{N_s}} N_l \mathbf{M}_l(x) \left(\sum_{j=1}^{\mathrm{N_s}} \mathbf{V}_{ij} \sum_{\alpha=1}^{n_y} N_\alpha \frac{\mathrm{F}_j^\alpha \mathrm{F}_l^\alpha}{(\sum_{k=1}^{\mathrm{N_s}} N_k \mathrm{F}_k^\alpha)^2} \right) \\
&= \delta x \sum_{l=1}^{\mathrm{N_s}} N_l \mathbf{M}_l(x) \left(\sum_{j=1}^{\mathrm{N_s}} \mathbf{V}_{ij} \mathbf{V}_{jl}^{-1} \right) \\
&= N_i \mathbf{M}_i(x) \delta x \equiv \langle \delta N_i^x \rangle \ , \qquad (68)
\end{aligned}
$$

which concludes the discussion of the situation where the y-distributions are step functions.

B Extended ${}_s\mathcal{P}lots$: a species is known (fixed)

It may happen that the yields of some species are not derived from the data sample at hand, but are taken to be known from other sources of information. Here, one denotes collectively as species '0' the overall component of such species. The number of expected events for species '0', N_0, being assumed to be known, is held fixed in the fit. In this Section, the indices $i, j...$ run over the $\mathrm{N_s}$ species for which the yields N_i are fitted, the fixed species '0' being excepted ($i, j... \neq 0$).

One can meet various instances of such a situation. Two extreme cases are:

1. the species '0' is very well known, such that the information on it contained by the data sample at hand is irrelevant. Not only is N_0 already pinned down by other means, but $\mathbf{M}_0(x)$, the marginal distribution of the fixed species, is available,
2. the species '0' is poorly known, and the data sample at hand is unable to resolve its contribution. This is the case if the y variables cannot discriminate between species '0' against any one of the other $\mathrm{N_s}$ species. Stated differently, if N_0 is left free to vary in the fit, the covariance matrix blows up for certain species and the measurement is lost. To avoid that, one is lead to accept an *a priori* value for N_0, and to compute systematics associated to the choice made for it. In that case, the worst case scenario is met if $\mathbf{M}_0(x)$ is unknown as well.

It is shown below that the ${}_s\mathcal{P}lot$ formalism can be extended to deal with this situation, whether or not $\mathbf{M}_0(x)$ is known, although in the latter case the statistical price to pay can be prohibitive.

B.1 Assuming $\mathbf{M}_0$ to be known

Here, it is assumed that $\mathbf{M}_0(x)$, is taken for granted. Then, it is not difficult to show that the Extended ${}_s\mathcal{P}lot$, which reproduces the marginal distribution of species n, is now given by:

$$
N_{\mathrm{n}} \ {}_s\tilde{\mathrm{M}}_{\mathrm{n}}(\bar{x}) \delta x = c_{\mathrm{n}} \mathbf{M}_0(x) \delta x + \sum_{e \subset \delta x} {}_s\mathcal{P}_{\mathrm{n}} \ , \qquad (69)
$$

where:

- ${}_s\mathcal{P}_{\rm n}$ is the previously defined sWeight of Eq. (14):

$$ {}_s\mathcal{P}_{\rm n} = \frac{\sum_j \mathbf{V}_{{\rm n}j} {\rm f}_j}{\sum_k N_k {\rm f}_k + N_0 {\rm f}_0} \ , \tag{70} $$

  where the covariance matrix $\mathbf{V}_{ij}$ is the one resulting from the fit of the $N_{i\neq 0}$ expected number of events, that is to say the inverse of the matrix:

$$ \mathbf{V}^{-1}_{ij} = \sum_{e=1}^{N} \frac{{\rm f}_i {\rm f}_j}{(\sum_k N_k {\rm f}_k + N_0 {\rm f}_0)^2} \ , \tag{71} $$

- $c_{\rm n}$ is the species dependent coefficient:

$$ c_{\rm n} = N_{\rm n} - \sum_j \mathbf{V}_{{\rm n}j} \ . \tag{72} $$

Some remarks deserve to be made:

- The Likelihood is now written:

$$ \mathcal{L} = \sum_{e=1}^{N} \ln \Big\{ \sum_{i=1}^{\rm N_s} N_i {\rm f}_i(y_e) + N_0 {\rm f}_0(y_e) \Big\} - \Big\{ \sum_{i=1}^{\rm N_s} N_i + N_0 \Big\} \ . \tag{73} $$

  Because N_0 is held fixed, in general, its assumed value combined with the fitted values for the N_i, does not maximize it:

$$ \frac{\partial \mathcal{L}}{\partial N_0} = \sum_{e=1}^{N} \frac{{\rm f}_0}{\sum_k N_k {\rm f}_k + N_0 {\rm f}_0} - 1 \neq 0 \ . \tag{74} $$

- It follows that the sum over the number of events per species does not equal the total number of events in the sample:

$$ \sum_i N_i = N - N_0 \left(\sum_{e=1}^{N} \frac{{\rm f}_0}{\sum_k N_k {\rm f}_k + N_0 {\rm f}_0} \right) \neq N - N_0 \ . \tag{75} $$

- Similarly, the Variance Matrix Sum Rule Eq. (18) holds only for $N_0 = 0$:

$$ \sum_i N_i \mathbf{V}^{-1}_{ij} = 1 - N_0 v_j \ , \tag{76} $$

  where the vector v_j is defined by:

$$ v_j \equiv \sum_{e=1}^{N} \frac{{\rm f}_0 {\rm f}_j}{(\sum_k N_k {\rm f}_k + N_0 {\rm f}_0)^2} \ . \tag{77} $$

- Accordingly, Eq. (19) becomes:

$$ \sum_j \mathbf{V}_{jl} = N_l + N_0 \sum_j \mathbf{V}_{lj} v_j \ . \tag{78} $$

- Thus, as they should, the $c_{\rm n}$ coefficients vanish only for $N_0 = 0$:

$$c_{\rm n} = -N_0 \sum_j \mathbf{V}_{nj} v_j \ . \tag{79}$$

- The above defined Extended ${}_s\mathcal{P}lots$ share the same properties as the ${}_s\mathcal{P}lots$:
  1. They reproduce the true marginal distributions, as in Eq. (16).
  2. In particular, they are properly normalized, as in Eq. (20).
  3. The sum of ${}_s\mathcal{P}^2_{\rm n}$ reproduces $\sigma^2[N_{\rm n}]$, as in Eq. (24).

B.2 Assuming $\mathbf{M}_0$ to be unknown

In the above treatment, because one assumes that a special species '0' enters in the sample composition, the sWeights per event do not add up to unity, as in Eq. (21). Instead one may define the sWeights for species '0' as:

$${}_s\mathcal{P}_0 \equiv 1 - \sum_i {}_s\mathcal{P}_i \tag{80}$$

and introduce the reconstructed ${}_s\tilde{\rm M}_0$ distribution (normalized to unity):

$${}_s\tilde{\rm M}_0(x)\delta x \ = \ \left(N - \sum_{i,j} \mathbf{V}_{ij} \right)^{-1} \sum_{e\subset\delta x} {}_s\mathcal{P}_0 \ , \tag{81}$$

which reproduces the true distribution $\mathbf{M}_0(x)$ if (by chance) the value assumed for N_0 is the one which maximizes the Likelihood.

Taking advantage of ${}_s\tilde{\rm M}_0(x)$, one may redefine the Extended ${}_s\mathcal{P}lots$ by:

$$N_{\rm n}\ {}_s\tilde{\rm M}_{\rm n}(\bar{x})\delta x = c_{\rm n}\ {}_s\tilde{\rm M}_0(x)\delta x + \sum_{e\subset\delta x} {}_s\mathcal{P}_{\rm n} = \sum_{e\subset\delta x} {}_{es}\mathcal{P}_{\rm n} \ , \tag{82}$$

where the redefined sWeight which appears on the right hand side is given by:

$${}_{es}\mathcal{P}_{\rm n} \ \equiv \ {}_s\mathcal{P}_{\rm n} + \frac{N_i - \sum_j \mathbf{V}_{ij}}{N - \sum_{i,j} \mathbf{V}_{ij}} {}_s\mathcal{P}_0 \ . \tag{83}$$

It does not rely on *a priori* knowledge on the true distribution $\mathbf{M}_0(x)$. With this redefinition, the following properties hold:

- The set of reconstructed x-distributions $N_i\tilde{\rm M}_i$ of Eq. (82) completed by $(N - \sum_i N_i)\tilde{\rm M}_0$ of Eq. (81) are such that they add up in each x-bin to the number of events observed.
- The normalization constant of the $\tilde{\rm M}_0$ distribution vanishes quadratically with N_0. It can be rewritten in the form:

$$N - \sum_{i,j} \mathbf{V}_{ij} = N_0^2 \ \left(v_0 - \sum_{i,j} \mathbf{V}_{ij} v_i v_j \right) \ , \tag{84}$$

  where v_0 is defined as v_j (cf. Eq. (77)) and where the last term is regular when $N_0 \to 0$.

- Whereas the normalization of the redefined extended ${}_s\mathcal{P}lots$ remains correct, the sum of the redefined sWeights Eq. (83) squared is no longer equal to $\mathbf{V}_{\rm nn}$. Instead:

$$\sum({}_{es}\mathcal{P}_{\rm n})^2 = \mathbf{V}_{\rm nn} + \frac{(N_{\rm n} - \sum_j \mathbf{V}_{{\rm n}j})^2}{N - \sum_{i,j} \mathbf{V}_{ij}} = \mathbf{V}_{\rm nn} + \frac{\sum_{ij} \mathbf{V}_{{\rm n}i}\mathbf{V}_{{\rm n}j} v_i v_j}{v_0 - \sum_{ij} \mathbf{V}_{ij} v_i v_j} \ . \tag{85}$$

  Since the expression on the right hand side is regular when $N_0 \to 0$, it follows that there is a price to pay to drop the knowledge of $\mathbf{M}_0(x)$, even though one expects a vanishing N_0. Technically, this feature stems from

$$\sum({}_s\mathcal{P}_0)^2 = \sum {}_s\mathcal{P}_0 = N - \sum_{i,j} \mathbf{V}_{ij} \ . \tag{86}$$

  Hence, the sum in quadrature of the ${}_s\tilde{\rm M}_0$ uncertainties per bin diverges with $N_0 \to 0$. This just expresses the obvious fact that no information can be extracted on species '0' from a sample which contains no such events.