Statistics

Contents

APFEL++ is often used to fit non-perturbative functions, such as parton distribution functions (PDFs) and fragmentation functions (FFs), to experimental data. As a consequence, it is necessary to make full use of the appropriate statistical tools that allow for a correct treatment of the experimental information when comparing predictions to data. This section is devoted to a detailed discussion of the concepts and tools that are often used when fitting PDFs and FFs to data.

Most of these tools are not directly implemented in APFEL++ but can be found in other codes, as for instance NangaParbat (see https://github.com/vbertone/NangaParbat).

1 The χ^2 in the presence of correlations

Suppose to have an ensamble of n measurements having the following structure:

$$m_i \pm \sigma_{i,\text{stat}} \pm \sigma_{i,\text{unc}} \pm \sigma_{i,\text{corr}}^{(1)} \pm \cdots \pm \sigma_{i,\text{corr}}^{(k)}\,, \tag{1.1}$$

where m_i, with $i = 1, \ldots, n$, is the central value of the i-th measurement, $\sigma_{i,\text{stat}}$ its (uncorrelated) statistical uncertainty, $\sigma_{i,\text{unc}}$ its uncorrelated systematic uncertainty[1], and $\sigma_{i,\text{corr}}^{(l)}$, with $l = 1, \ldots, k$, its correlated systematic uncertainties. With this information at hand, one can construct the full covariance matrix V_{ij} as follows (see for example Ref. [1]):

$$V_{ij} = \left(\sigma_{i,\text{stat}}^2 + \sigma_{i,\text{unc}}^2\right)\delta_{ij} + \sum_{l=1}^{k} \sigma_{i,\text{corr}}^{(l)} \sigma_{j,\text{corr}}^{(l)}\,. \tag{1.2}$$

This is a clearly symmetric matrix. Given a set of predictions t_i corresponding to the n measurements of the ensamble, the χ^2 takes the form:

$$\chi^2 = \sum_{i,j=1}^{n} (m_i - t_i)\, V_{ij}^{-1}\, (m_j - t_j) = \mathbf{y}^T \cdot \mathbf{V}^{-1} \cdot \mathbf{y}\,, \tag{1.3}$$

where in the second equality I have used the matricial notation and defined $y_i = m_i - t_i$. A convenient way to compute the χ^2 relies on the Cholesky decomposition of the covariance matrix $\mathbf{V}$. In particular, it can be proven that any symmetric and positive definite matrix $\mathbf{V}$ can be decomposed as:

$$\mathbf{V} = \mathbf{L} \cdot \mathbf{L}^T\,, \tag{1.4}$$

[1] There could be more than one uncorrelated systematic uncertainty. In this case, $\sigma_{i,\text{unc}}$ is just the square root of the sum in quadrature of all the uncorrelated systematic uncertainties.

where $\mathbf{L}$ is a lower triangular matrix whose entries are related recursively to those of $\mathbf{V}$ as follows:

$$\begin{aligned} L_{kk} &= \sqrt{V_{kk} - \sum_{j=1}^{k-1} L_{kj}^2}\,, \\ L_{ik} &= \frac{1}{L_{kk}}\left(V_{ik} - \sum_{j=1}^{k-1} L_{ij}L_{kj}\right)\,, \quad k < i\,, \\ L_{ik} &= 0\,, \quad k > i\,. \end{aligned} \tag{1.5}$$

It is then easy to see that the χ^2 can be written as:

$$\chi^2 = \left|\mathbf{L}^{-1}\cdot\mathbf{y}\right|^2\,. \tag{1.6}$$

But the vector $\boldsymbol{\chi} \equiv \mathbf{L}^{-1}\cdot\mathbf{y}$ is the solution of the linear system:

$$\mathbf{L}\cdot\boldsymbol{\chi} = \mathbf{y}\,, \tag{1.7}$$

that can be efficiently solved by forward substitution, so that:

$$\chi^2 = |\boldsymbol{\chi}|^2\,. \tag{1.8}$$

Following this procedure, one does not need to compute explicitly the inverse of the covariance matrix $\mathbf{V}$, simplifying significantly the computation of the χ^2.

2 Additive and multiplicative uncertainties

The correlated systematic uncertainties $\sigma^{(l)}_{i,\mathrm{corr}}$ may be either *additive* or *multiplicative*. The nature of the single uncertainties is typically provided by the experiments that release the measurements. A typical example of multiplicative uncertainty is the luminosity uncertainty but there can be others.

Now let us express all the correlated systematic uncertainties $\sigma^{(l)}_{i,\mathrm{corr}}$ as relative to the associate central value m_i, so that one defines[2]:

$$\sigma^{(l)}_{i,\mathrm{corr}} \equiv \delta^{(l)}_{i,\mathrm{corr}} m_i \tag{2.1}$$

and let us also define $s_i^2 \equiv \sigma^2_{i,\mathrm{stat}} + \sigma^2_{i,\mathrm{unc}}$ so that Eq. (1.2) can be rewritten as:

$$V_{ij} = s_i^2\delta_{ij} + \left(\sum_{l=1}^{k} \delta^{(l)}_{i,\mathrm{corr}}\delta^{(l)}_{j,\mathrm{corr}}\right) m_i m_j\,. \tag{2.2}$$

Now I split the correlated systematic uncertainties into k_a additive uncertainties and k_m multiplicative uncertainties, such that $k_a + k_m = k$. This way Eq. (2.2) takes the form:

$$V_{ij} = s_i^2\delta_{ij} + \left(\sum_{l=1}^{k_a} \delta^{(l)}_{i,\mathrm{add}}\delta^{(l)}_{j,\mathrm{add}} + \sum_{l=1}^{k_m} \delta^{(l)}_{i,\mathrm{mult}}\delta^{(l)}_{j,\mathrm{mult}}\right) m_i m_j\,. \tag{2.3}$$

It is believed that this definition of the covariance matrix is problematic in that it results in the so-called D'Agostini bias of the multiplicative uncertainties [2]. A possible solution to this problem is the so-called t_0-prescription [3], where the experimental central value m_i in the multiplicative term is replaced by a fixed theoretical predictions $t_i^{(0)}$, typically computed in a previous fit in which the "standard" definition of the covariance matrix in Eq. (1.2) (often referred to as *experimental* definition) is used. Applying the t_0 prescription, the covariance matrix takes the form:

$$V_{ij} = s_i^2\delta_{ij} + \sum_{l=1}^{k_a} \delta^{(l)}_{i,\mathrm{add}}\delta^{(l)}_{j,\mathrm{add}} m_i m_j + \sum_{l=1}^{k_m} \delta^{(l)}_{i,\mathrm{mult}}\delta^{(l)}_{j,\mathrm{mult}} t_i^{(0)} t_j^{(0)}\,. \tag{2.4}$$

[2] Note that this redefinition does not change the nature of the uncertainties, additive uncertainties remain additive as well as multiplicative uncertainties remain multiplicative.

3 Determining the systematic shifts

In order to visualise the effect of systematic uncertainties, it is instructive to compute the *systematic shift* generated by the systematic uncertainties. To do so, one needs to write the χ^2 in terms of the so-called "nuisance parameters" λ_α. One can show that the definition of the χ^2 in Eq. (1.3) is equivalent to [1]:

$$\chi^2 = \sum_{i=1}^{n} \frac{1}{s_i^2} \left(m_i - t_i - \sum_{\alpha=1}^{k} \lambda_\alpha \sigma_{i,\text{corr}}^{(\alpha)} \right)^2 + \sum_{\alpha=1}^{k} \lambda_\alpha^2 \,. \tag{3.1}$$

The optimal value of the nuisance parameters can be computed by minimising the χ^2 with respect to them, that is imposing:

$$\frac{\partial \chi^2}{\partial \lambda_\beta} = 0 \,. \tag{3.2}$$

This yields the system:

$$\sum_{\beta=1}^{k} A_{\alpha\beta} \lambda_\beta = \rho_\alpha \,, \tag{3.3}$$

with:

$$A_{\alpha\beta} = \delta_{\alpha\beta} + \sum_{i=1}^{n} \frac{\sigma_{i,\text{corr}}^{(\alpha)} \sigma_{i,\text{corr}}^{(\beta)}}{s_i^2} \quad \text{and} \quad \rho_\alpha = \sum_{i=1}^{n} \frac{m_i - t_i}{s_i^2} \sigma_{i,\text{corr}}^{(\alpha)} \,, \tag{3.4}$$

that determines the values of λ_β. The quantity:

$$d_i = \sum_{\alpha=1}^{k} \lambda_\alpha \sigma_{i,\text{corr}}^{(\alpha)} \tag{3.5}$$

in Eq. (3.1) can be interpreted as a shift caused by the correlated systematic uncertainties. Defining the shifted predictions as:

$$\bar{t}_i = t_i + d_i \,, \tag{3.6}$$

the χ^2 reads:

$$\chi^2 = \sum_{i=1}^{n} \left(\frac{m_i - \bar{t}_i}{s_i} \right)^2 + \sum_{\alpha=1}^{k} \lambda_\alpha^2 \,. \tag{3.7}$$

Therefore, up to a penalty term given by the sum of the square of the nuisance parameters, the χ^2 takes the form of the uncorrelated definition. In order to achieve a visual assessment of the agreement between data and theory, it appears natural to compare the central experimental values m_i to the shifted theoretical predictions $\bar{t}_i$ in units of the uncorrelated uncertainty s_i.

4 Effect of cuts on the χ^2

A relevant question to ask is what is the effect of possible cuts on the input data set on the form of the χ^2 in the presence of correlations. In order to address this question, I consider a simple data set made of two datapoints equipped with a single uncorrelated uncertainty and a single fully-correlated uncertainty:[3]

$$S: \quad \{m_i \pm \sigma_i^u \pm \sigma_i^c\} \,, \quad i = 1, 2 \,. \tag{4.1}$$

The resulting covariance matrix reads:

$$\mathbf{V} = \begin{pmatrix} (\sigma_1^u)^2 + (\sigma_1^c)^2 & \sigma_1^c \sigma_2^c \\ \sigma_1^c \sigma_2^c & (\sigma_2^u)^2 + (\sigma_2^c)^2 \end{pmatrix} \,, \tag{4.2}$$

with inverse:

$$\mathbf{V}^{-1} = \frac{1}{(\sigma_1^u \sigma_2^u)^2 + (\sigma_1^c \sigma_2^u)^2 + (\sigma_1^u \sigma_2^c)^2} \begin{pmatrix} (\sigma_2^u)^2 + (\sigma_2^c)^2 & -\sigma_1^c \sigma_2^c \\ -\sigma_1^c \sigma_2^c & (\sigma_1^u)^2 + (\sigma_1^c)^2 \end{pmatrix} \,. \tag{4.3}$$

[3] Note that one needs to include *also* an uncorrelated uncertainty because otherwise the covariance matrix would be singular. This a symptom of the fact that if no uncorrelated uncertainties are present, the two points would be totally dependent on each other. This observation will be relevant in the following.

In addition, the set of two theoretical predictions $\{t_i\}$ that allows one to define a column vector of residuals $\mathbf{r}^{\mathrm{T}} = (r_1,\ r_2)$ with $r_i = m_i - t_i$. The χ^2 then takes the following explicit expression:

$$\chi^2 = \mathbf{r}^{\mathrm{T}} \cdot \mathbf{V}^{-1} \cdot \mathbf{r} = \frac{r_1^2 \left[(\sigma_2^u)^2 + (\sigma_2^c)^2\right] + r_2^2 \left[(\sigma_1^u)^2 + (\sigma_1^c)^2\right] - 2 r_1 r_2 \sigma_1^c \sigma_2^c}{(\sigma_1^u \sigma_2^u)^2 + (\sigma_1^c \sigma_2^u)^2 + (\sigma_1^u \sigma_2^c)^2} \,. \tag{4.4}$$

Now, suppose one wants to exclude the datapoint with $i = 2$ from the definition of the χ^2. The possibly most natural way to proceed is to exclude the point, along with its uncertainties, directly from the set in Eq. (4.1). This straightforwardly leads to the following χ^2:

$$\chi^2_{\text{cut 1}} = \frac{r_1^2}{(\sigma_1^u)^2 + (\sigma_1^c)^2} \,. \tag{4.5}$$

However, this procedure might be questioned in that it artificially modifies the original data set by effectively defining a *new* data set where the point $i = 2$ is not present. In the absence of correlations this is a sound procedure because each single datapoint can be regarder an independent subset of the original set. However, this is not the case when correlations are present.

Therefore, it is necessary to devise a method that avoids any modification of the data set but yet allows one to prevent specific datapoints to contribute to the χ^2. In general, the purpose of excluding datapoints from a fit is to avoid that the model used to compute the predictions is forced outside its application region. One can therefore *enforce* that the model exactly agrees with the datapoints to be excluded. This effectively amounts to set $t_2 = m_2$, or equivalently $r_2 = 0$, reducing Eq. (4.4) to:

$$\chi^2_{\text{cut 2}} = \frac{r_1^2 \left[(\sigma_2^u)^2 + (\sigma_2^c)^2\right]}{(\sigma_1^u \sigma_2^u)^2 + (\sigma_1^c \sigma_2^u)^2 + (\sigma_1^u \sigma_2^c)^2} \,. \tag{4.6}$$

This has the clear advantage that the experimental information, including correlations, remains totally intact. In order to show that Eq. (4.6) has the right properties, I first take the limit for $\sigma_2^c \to 0$. This gives:

$$\lim_{\sigma_2^c \to 0} \chi^2_{\text{cut}} = \frac{r_1^2}{(\sigma_1^u)^2 + (\sigma_1^c)^2} \,, \tag{4.7}$$

which reproduces the result of Eq. (4.5). This was to be expected because no reference to the point $i = 2$ can remain after the removal of its correlation to the point $i = 1$. In addition, it is also very instructive to take the limit $\sigma_2^u \to 0$ Eq. (4.6), which gives:

$$\lim_{\sigma_2^u \to 0} \chi^2_{\text{cut}} = \frac{r_1^2}{(\sigma_1^u)^2} \,. \tag{4.8}$$

In this case, not only any reference to the point $i = 2$ drops out, but also the dependence of the χ^2 on the correlation uncertainty of the point $i = 1$ disappears. Despite at first sight this looks counterintuitive, it seems to be the correct behaviour. To see this, observe that, if the point $i = 2$ has no uncorrelated uncertainty, its fluctuations are totally driven by the fluctuations of the point $i = 1$. Therefore, the point $i = 2$ is completely dependent on the point $i = 1$ and thus cannot give any contribution to the χ^2. In addition, the correlation uncertainty of the point $i = 1$ must also be irrelevant because, being $i = 2$ totally correlated to $i = 1$, any correlation of $i = 1$ to $i = 2$ translates into a correlation to itself and this cannot contribute to the χ^2 either. Importantly, if one takes the limit $\sigma_1^u \to 0$, the χ^2 diverges no matter the value of σ_1^c. This is consistent with setting $\sigma_1^u = \sigma_2^u$ since the beginning in Eq. (4.1) which indeed gives a singular χ^2 because the covariance matrix is singular. Eq. (4.5) does not produce these features hanging in favour of Eq. (4.6).

As a further argument in favour of Eq. (4.6) and against Eq. (4.5), I consider a larger data set with, say, n data points and split it into two subsets with n_1 and n_2 data points, respectively, such that $n_1 + n_2 = n$. As I will explicitly show below, Eq. (4.6) guarantees that the χ^2's of the two subsets, χ^2_1 and χ^2_2, combine sensibly to give the total χ^2. I will also show that this is not the case for Eq. (4.5). The fulfilment of this condition is crucial to ensure that the experimental information carried by the data set as a whole is preserved also when one is either not able or does not want to make predictions for subset of data points. This property is particularly relevant in fits that adopt the cross-validation method as stopping criterion. In this context, the full data set is indeed split into a training subset and a validation subset and only the χ^2 of the training set is minimised while the validation χ^2 is monitored. Since the cross-validation method relies on the exploitation of the full experimental information to stop the fit, it is necessary that training and validation χ^2's combine meaningfully.

Now, I move to explicitly show that the generalisation of Eq. (4.6) to an n-point data set where n_2 are cut away and n_1 remain fulfils the requirement:

$$\chi^2 = |\boldsymbol{\chi}_1 + \boldsymbol{\chi}_2|^2 \,, \tag{4.9}$$

where the vectors $\boldsymbol{\chi}_1$ and $\boldsymbol{\chi}_2$ are defined according to Eq. (1.7). To define them more precisely, without loss of generality, one can order the data points in such a way that those labelled with $i = 1, \ldots, n_1$ belong to the first subset and the points labelled with $i = n_1 + 1, \ldots, n_1 + n_2 (= n)$ belong to the second subset. The entries of the $\boldsymbol{\chi}_1$ and $\boldsymbol{\chi}_2$ are then given by:

$$\boldsymbol{\chi}_{1,k} = \sum_{j=1}^{n_1} L^{-1}_{kj} y_j \,, \quad \text{and} \quad \boldsymbol{\chi}_{2,k} = \sum_{j=n_1+1}^{n} L^{-1}_{kj} y_j \,, \tag{4.10}$$

where L^{-1}_{ij} are the entries of the inverse of the Cholesky decomposition $\mathbf{L}$ of the covariance $\mathbf{V}$. It should be clear that, according to the generalisation of the prescription of Eq. (4.6), one has:

$$\chi^2_1 = |\boldsymbol{\chi}_1|^2 \,, \quad \text{and} \quad \chi^2_2 = |\boldsymbol{\chi}_2|^2 \,. \tag{4.11}$$

The proof of Eq. (4.9) runs as follows:

$$\begin{aligned}
\chi^2 &= \sum_{i=1}^{n}\sum_{j=1}^{n} r_i V^{-1}_{ij} r_j = \sum_{i=1}^{n_1}\sum_{j=1}^{n_1} r_i V^{-1}_{ij} r_j + \sum_{i=n_1+1}^{n}\sum_{j=n_1+1}^{n} r_i V^{-1}_{ij} r_j + 2\sum_{i=1}^{n_1}\sum_{j=n_1+1}^{n} r_i V^{-1}_{ij} r_j \\
&= \sum_{k=1}^{n}\left(\sum_{i=1}^{n_1} L^{-1}_{ki} r_i\right)\left(\sum_{j=1}^{n_1} L^{-1}_{kj} r_j\right) + \sum_{k=1}^{n}\left(\sum_{i=n_1+1}^{n} L^{-1}_{ki} r_i\right)\left(\sum_{j=n_1+1}^{n} L^{-1}_{kj} r_j\right) \\
&+ 2\sum_{k=1}^{n}\left(\sum_{i=1}^{n_1} L^{-1}_{ki} r_i\right)\left(\sum_{j=n_1+1}^{n} L^{-1}_{kj} r_j\right) \\
&= \boldsymbol{\chi}_1^{\mathrm{T}} \cdot \boldsymbol{\chi}_1 + \boldsymbol{\chi}_2^{\mathrm{T}} \cdot \boldsymbol{\chi}_2 + 2\boldsymbol{\chi}_1^{\mathrm{T}} \cdot \boldsymbol{\chi}_2 = |\boldsymbol{\chi}_1 + \boldsymbol{\chi}_2|^2 \,,
\end{aligned} \tag{4.12}$$

where in the last step of the first line I have used the fact that the covariance matrix (and thus its inverse) is symmetric upon exchange of the indices while in the second and third lines I have used the Cholesky decomposition of the covariance matrix. It straightforward to see that Eq. (4.5) does not fulfils the same relation. The fact that Eq. (4.9) is consistent with how the χ^2's should combine can be better seen by giving it a geometrical interpretation.

4.1 A geometrical point of view

It is instructive to consider the question of cuts from a geometrical point of view. In fact, the whole bunch of questions related to the computation of the χ^2 can be nicely visualised as a set of geometrical operations in a multidimensional space $\mathcal{M}$ (manifold) equipped with the appropriate metric tensor. Based on the definition in Eq. (1.3), the χ^2 can be interpreted as the squared absolute value of the n-dimensional vector $\mathbf{y}$ given by the difference between the vector of experimental central values $\mathbf{m}$ and the vector of theoretical predictions $\mathbf{t}$. Importantly, the scalar product required to compute this absolute value has to be computed using the inverse of the covariance matrix $\mathbf{V}^{-1}$ as a covariant positive-definite metric tensor.[4] Borrowing the Einstein notation to represent scalar products, *i.e.* defining contravariant vectors with upper index as:

$$y^i = V^{-1}_{ij} y_j \,, \tag{4.13}$$

where the sum over j between 1 and n is understood, the χ^2 is the *norm* squared of the vector $\mathbf{y}$ in the space $\mathcal{M}$.

$$\chi^2 = \mathbf{y}^{\mathrm{T}} \cdot \mathbf{V}^{-1} \cdot \mathbf{y} = y^i y_i \equiv \|\mathbf{y}\|^2 \,. \tag{4.14}$$

It is interesting to notice that the minimisation of the χ^2 can then be regarded as the attempt to align the vector $\mathbf{t}$ to the vector $\mathbf{m}$ in the space $\mathcal{M}$ in such a way that their difference is as small as possible.

[4] The contravariant counterpart of $\mathbf{V}^{-1}$ is clearly $\mathbf{V}$, given that $\mathbf{V}^{-1} \cdot \mathbf{V} = \mathbb{I}$.

The enforcement of cuts can be accommodated in this framework by requiring that the χ^2 is the squared absolute value (or length) of the projection of $\mathbf{y}$ in the subspace of $\mathcal{M}$, $\mathcal{M}_1$, defined by the components of the vector $\mathbf{m}$ corresponding to the points that pass the cut. Introducing the operator $\mathbf{P}$ that projects any vector in $\mathcal{M}$ onto $\mathcal{M}_1$, the corresponding χ^2 will read:

$$\chi^2_1 = (\mathbf{P}\cdot\mathbf{y})^{\mathrm{T}}\cdot\mathbf{V}^{-1}\cdot(\mathbf{P}\cdot\mathbf{y}) = (\mathbf{P}\cdot\mathbf{y})^i\,(\mathbf{P}\cdot\mathbf{y})_i = \|\mathbf{P}\cdot\mathbf{y}\|^2\,. \tag{4.15}$$

The matrix corresponding to the projector $\mathbf{P}$ is simply the diagonal matrix with zero's and one's on the diagonal depending on whether the point is excluded or included, respectively.[5] The net effect of the projector $\mathbf{P}$ is to set to zero the components of the vector of residuals $\mathbf{y}$ corresponding to the points excluded by the cut. This is in agreement with Eq. (4.6) in the case of a two-dimensional space.

It is interesting to define the projector $\mathbf{Q} = \mathbb{I} - \mathbf{P}$ orthogonal to $\mathbf{P}$[6] such that the corresponding χ^2 reads:

$$\chi^2_2 = (\mathbf{Q}\cdot\mathbf{y})^{\mathrm{T}}\cdot\mathbf{V}^{-1}\cdot(\mathbf{Q}\cdot\mathbf{y}) = (\mathbf{Q}\cdot\mathbf{y})^i\,(\mathbf{Q}\cdot\mathbf{y})_i = \|\mathbf{Q}\cdot\mathbf{y}\|^2\,. \tag{4.16}$$

With these definitions at hands, one can write the total χ^2 as:

$$\chi^2 = \|\mathbf{y}\|^2 = \|\mathbf{P}\cdot\mathbf{y}+\mathbf{Q}\cdot\mathbf{y}\|^2 = \chi^2_1+\chi^2_2+2\,(\mathbf{P}\cdot\mathbf{y})^i\,(\mathbf{Q}\cdot\mathbf{y})_i\,. \tag{4.17}$$

that agrees with Eq. (4.12). In conclusion, the introduction of a cut on the full data set has a nice geometrical interpretation: the resulting χ^2 corresponds to the norm of the projection of the vector or residual $\mathbf{y}$, defined on the manifold $\mathcal{M}$ with covariant metric $\mathbf{V}^{-1}$, on to the subspace $\mathcal{M}_1$ defined by the projector $\mathbf{P}$. This is consistent with Eq. (4.9) and provides it with a nice geometrical interpretation.[7]

5 Monte Carlo replica generation

In this section, I consider the Monte Carlo generation of artificial replicas. The aim is to obtain a formula for the generation of replicas that gives rise to a figure of merit distributed correctly, *i.e.* that follows a χ^2 distribution with as many degrees of freedom as number of data points. Following Ref. [4], one requires that the χ^2, that in matricial form can be expressed as:

$$\chi^2 = \mathbf{y}^T\cdot\mathbf{V}^{-1}\cdot\mathbf{y}\,, \tag{5.1}$$

where $\mathbf{V}$ is the $n\times n$ covariance matrix, being n the number of datapoints, and $\mathbf{y}$ is a column vector with n entries that can be expressed as:

$$y_j = f_j - m_j\,, \tag{5.2}$$

with m_j a given value (to be thought as the central value) and f_j its fluctuated value, is distributed like:

$$\chi^2 \sim \sum_{i=1}^{n} z_i^2 = |\mathbf{z}|^2\,, \tag{5.3}$$

where $z_1,\ldots,z_n$ are independent, standard normal random variables, *i.e.* such that:

$$\langle z_i\rangle = 0 \quad\text{and}\quad \langle z_i z_j\rangle = \delta_{ij}\,, \tag{5.4}$$

where the symbol $\langle\ldots\rangle$ indicates the average.[8] This immediately tells that:

$$\langle\chi^2\rangle = n\,. \tag{5.7}$$

[5] Notice that $\mathbf{P}$ is indeed idempotent, *i.e.* $\mathbf{P}^2=\mathbf{P}$, as required.

[6] $\mathbf{Q}$ is such that $\mathbf{Q}^2=\mathbf{Q}$, $\mathbf{P}\cdot\mathbf{Q}=0$, and $\mathbf{P}+\mathbf{Q}=\mathbb{I}$.

[7] As an interesting aside, the difference between the following equalities:

$$\chi^2 = \left|\mathbf{L}^{-1}\cdot\mathbf{y}\right|^2 \quad\text{and}\quad \chi^2 = \|\mathbf{y}\|^2\,,$$

can be paralleled with the norm of a vector in special relativity in the Euclidean space in the first case (where the time component is multiplied by the imaginary unit i and the space components are left unchanged) and in Minkowski space in the second case. Indeed, the Cholesky decomposition of the metric tensor $g_{\mu\nu}=\mathrm{diag}(-1,1,1,1)$ is $L_{\mu\nu}=\mathrm{diag}(i,1,1,1)$.

[8] The relevant identities are:

$$\langle z_i^{2n-1}\rangle = 0 \quad\text{and}\quad \langle z_i^{2n}\rangle = (2n-1)!! \quad n=1,2,3,\ldots\,, \tag{5.5}$$

where $(2n-1)!! = 1\cdot3\cdot\cdots\cdot(2n-1)$ is the odd semi-factorial. When considering combinations of two random variables the result is:

$$\langle z_i^n z_j^m\rangle = \begin{cases} \langle z_i^{n+m}\rangle & i=j \\ \langle z_i^n\rangle\langle z_j^m\rangle & i\neq j \end{cases} \tag{5.6}$$

As already discussed above, Eq. (5.1) can alternatively be written as:

$$\chi^2 = \left|\mathbf{L}^{-1}\cdot\mathbf{y}\right|^2 , \tag{5.8}$$

where $\mathbf{L}$ is the Cholesky decomposition of the matrix $\mathbf{V}$, such that:

$$\mathbf{V} = \mathbf{L}\cdot\mathbf{L}^T . \tag{5.9}$$

Comparing Eqs. (5.8) and (5.3), one can infer that the following relation should hold:

$$\mathbf{L}^{-1}\cdot\mathbf{y} \sim \mathbf{z} , \tag{5.10}$$

which immediately implies:

$$\mathbf{y} \sim \mathbf{L}\cdot\mathbf{z} , \tag{5.11}$$

or equivalently:

$$f_j \sim m_j + \sum_{i=1}^{n} L_{ji} z_i . \tag{5.12}$$

In conclusion, in order for the χ^2 to have the correct distribution, the fluctuations f_j around the experimental central values m_j of a given Monte Carlo replica have to be computed according to Eq. (5.12).

I now show that Eq. (5.12) reproduces the statistical properties of the original ensamble of data. To do so, I compute average and correlations of Eq. (5.12) using Eq. (5.4). The average gives:

$$\langle f_j\rangle = m_j + \sum_{i=1}^{n} L_{ji}\langle z_i\rangle = m_j , \tag{5.13}$$

as expected. The correlation instead gives:

$$\begin{aligned}\langle f_j f_k\rangle &= \left\langle \left(m_j + \sum_{i=1}^{n} L_{ji} z_i\right)\left(m_k + \sum_{i'=1}^{n} L_{ki'} z_{i'}\right)\right\rangle \\ &= m_j m_k + \sum_{i,i'=1}^{n} L_{ji'} L_{ki'}\langle z_{i'} z_{i'}\rangle = \langle f_j\rangle\langle f_k\rangle + V_{jk} ,\end{aligned} \tag{5.14}$$

again as expected. These are fundamental requirements of a sound Monte Carlo ensamble (see also Eq. (10) of Ref. [3]).

Eq. (5.12) should be compared to the more "standard" Monte-Carlo replica generation formulas such as those given in Refs. [5, 6]. I now specifically consider Eq. (20) of Ref. [6] that I report here in terms of relative uncertainties δ's matching our notation:

$$\widetilde{f}_j \sim m_j \prod_{m=1}^{k_m}\left(1+\delta^{(m)}_{\text{mult},j} x_m\right)\left(1+\delta_{\text{unc},j} z_j + \sum_{a=1}^{k_a}\delta^{(a)}_{\text{add},j} y_a\right) , \tag{5.15}$$

where z_j, y_a, and x_m are independent random variables. I now expand the product in the front up to second order in x

$$\prod_{m=1}^{k_m}\left(1+\delta^{(m)}_{\text{mult},j} x_m\right) = 1 + \sum_{m=1}^{k_m}\delta^{(m)}_{\text{mult},j} x_m + \sum_{m=1}^{k_m}\sum_{\substack{m'=1\\ m'\neq m}}^{k_m}\delta^{(m)}_{\text{mult},j}\delta^{(m')}_{\text{mult},j} x_m x_{m'} + \text{trilinear} , \tag{5.16}$$

where the "trilinear" corrections correspond to combinations of the kind $x_i x_j x_k$, with $i\neq j\neq k\neq i$, such that their average vanishes: $\langle x_i x_j x_k\rangle = \langle x_i\rangle\langle x_j\rangle\langle x_k\rangle = 0$. This allows one to expand Eq. (5.15) as:

$$\begin{aligned}\widetilde{f}_j \sim\ & m_j\left(1+\delta_{\text{unc},j} z_j + \sum_{a=1}^{k_a}\delta^{(a)}_{\text{add},j} y_a + \sum_{m=1}^{k_m}\delta^{(m)}_{\text{mult},j} x_m\right. \\ &+ \delta_{\text{unc},j}\sum_{m=1}^{k_m}\delta^{(m)}_{\text{mult},j} z_j x_m + \sum_{a=1}^{k_a}\sum_{m=1}^{k_m}\delta^{(a)}_{\text{add},j}\delta^{(m)}_{\text{mult},j} y_a x_m \\ &\left.+ \sum_{m=1}^{k_m}\sum_{\substack{m'=1\\ m'\neq m}}^{k_m}\delta^{(m)}_{\text{mult},j}\delta^{(m')}_{\text{mult},j} x_m x_{m'}\right) + \text{trilinear} .\end{aligned} \tag{5.17}$$

Since Eq. (5.17) involves only terms that have at most one power of the same random variable, the average of $\widetilde{f}_j$ automatically yields:

$$\langle \widetilde{f}_j \rangle = m_j \,, \tag{5.18}$$

in agreement with the expectation. I now compute the correlations making use of the combination rules of random variables. This gives:

$$\begin{aligned}
\langle \widetilde{f}_j \widetilde{f}_k \rangle \;&\sim\; m_j m_k \Bigg(1 + \delta^2_{\text{unc},j}\delta_{jk} + \sum_{a=1}^{k_a} \delta^{(a)}_{\text{add},j}\delta^{(a)}_{\text{add},k} + \sum_{m=1}^{k_m} \delta^{(m)}_{\text{mult},j}\delta^{(m)}_{\text{mult},k} \\
&+\; \delta^2_{\text{unc},j} \sum_{m=1}^{k_m} \delta^{(m)}_{\text{mult},j}\delta^{(m)}_{\text{mult},k} + \sum_{a=1}^{k_a}\sum_{m=1}^{k_m} \delta^{(a)}_{\text{add},j}\delta^{(a)}_{\text{add},k}\delta^{(m)}_{\text{mult},j}\delta^{(m)}_{\text{mult},k} \\
&+\; \sum_{m=1}^{k_m}\sum_{\substack{m'=1\\ m'\neq m}}^{k_m} \delta^{(m)}_{\text{mult},j}\delta^{(m)}_{\text{mult},k}\delta^{(m')}_{\text{mult},j}\delta^{(m')}_{\text{mult},k} + \mathcal{O}(\delta^6) \Bigg) \\
&=\; \langle \widetilde{f}_j \rangle \langle \widetilde{f}_k \rangle + V_{jk} + \mathcal{O}(\sigma^2_{\text{mult}} s^2, \sigma^2_{\text{mult}}\sigma^2_{\text{add}}) \,.
\end{aligned} \tag{5.19}$$

In conclusion, Eq. (5.15) reproduces exactly the central values of the original ensemble (see Eq. (5.18)) while it reproduces correlations only up to corrections proportional to the fourth power of the typical uncertainty. Interestingly, these corrections are all proportional to σ^2_{mult}. This means that, in the absence of multiplicative uncertainties, somewhat expectedly Eq. (5.15) *exactly* reproduces the correlations. On the other hand, if multiplicative uncertainties are large these correction become potentially sizeable preventing Eq. (5.15) from correctly reproducing the correlations of the original ensemble.

The case of Eq. (13) of Ref. [5] is more involved because of the presence of the square-root sign for the relative multiplicative uncertainties. This case can be handled again be expanding to second order the factor depending on multiplicative uncertainties with the only difference that, due to the expansion:

$$\sqrt{1+x} = 1 + \frac{1}{2}x - \frac{1}{8}x^2 + \mathcal{O}(x^3) \,, \tag{5.20}$$

the square root produces a factor $1/2$ in front of the linear terms and a factor $-1/8$ in front of the quadratic terms. This is expected to reduce the impact of the corrections terms in Eq. (5.19).

It is instructive to work out whether Eq. (5.12) can be directly related to the Monte-Carlo replicas generation formulas given in Refs. [5, 6]. The aim is to find a closed-form expression for the Cholesky-decomposition matrix $\mathbf{L}$ of the covariance matrix $\mathbf{V}$ in terms of the single uncertainties. Through Eq. (5.12), this will finally allow one to derive a formula of the same kind of those given in Refs. [5, 6]. The basic observation is that the covariance matrix can be written as:

$$\mathbf{V} = \mathbf{D} + \sum_{l=1}^{k} \mathbf{V}^{(l)} = \sum_{l=0}^{k} \mathbf{L}^{(l)} \cdot \mathbf{L}^{(l)T} \,, \tag{5.21}$$

with:

$$L^{(l)}_{ij} = \begin{cases} s_i\delta_{ij} & l = 0 \,, \\ \sigma^{(l)}_i \delta_{j1} & 1 \leq l \leq k \,. \end{cases} \tag{5.22}$$

On the other hand, one also knows that:

$$\mathbf{V} = \mathbf{L} \cdot \mathbf{L}^T \,, \tag{5.23}$$

therefore:

$$\mathbf{L} \cdot \mathbf{L}^T = \sum_{l=0}^{k} \mathbf{L}^{(l)} \cdot \mathbf{L}^{(l)T} \,. \tag{5.24}$$

The question is to express $\mathbf{L}$ in terms of all the $\mathbf{L}^{(l)}$. Unfortunately though this equation has no general solution. However, a step forward can be done be exploiting the fact that the matrices $\mathbf{V}^{(l)}$ are positive semidefinite (*i.e.* they have zero determinant) and thus their Cholesky decomposition is not unique. In particular, provided that $k \leq n$, one can choose:

$$\widetilde{L}^{(l)}_{ij} = \sigma^{(l)}_i \delta_{jl} \,, \quad 1 \leq l \leq k \leq n \,, \tag{5.25}$$

such that:

$$\left(\sum_{l=1}^{k}\widetilde{\mathbf{L}}^{(l)}\right)\cdot\left(\sum_{l'=1}^{k}\widetilde{\mathbf{L}}^{(l')}\right)=\sum_{l=1}^{k}\mathbf{V}^{(l)}\,. \tag{5.26}$$

However, choosing:

$$\widetilde{\mathbf{L}}=\mathbf{L}^{(0)}+\sum_{l=1}^{k}\widetilde{\mathbf{L}}^{(l)}\,, \tag{5.27}$$

and multiplying it for its transpose would give:

$$\widetilde{\mathbf{L}}\cdot\widetilde{\mathbf{L}}^{T}=\mathbf{V}+\sum_{l=1}^{k}\left(\mathbf{L}^{(0)}\cdot\widetilde{\mathbf{L}}^{(l)T}+\widetilde{\mathbf{L}}^{(l)}\cdot\mathbf{L}^{(0)T}\right)\,, \tag{5.28}$$

that gives back the covariance matrix plus a spurious term. Writing the expression above in terms of single entries, one finds:

$$\sum_{m=1}^{n}\widetilde{L}_{im}\widetilde{L}_{jm}=V_{ij}+s_i\sigma_j^{(i)}+s_j\sigma_i^{(j)}\,. \tag{5.29}$$

The peculiarity here is that the upper index of $\sigma_j^{(i)}$ and $\sigma_i^{(j)}$, since $k\leq n$ (and in fact usually $k\ll n$, *i.e.* the number of correlated uncertainties is typically much smaller than the number of data points), exceeds the allowed range. To fix this problem, I just assume $\sigma_j^{(i)}=0$ for $i>k$. This finally gives:

$$\sum_{m=1}^{n}\widetilde{L}_{im}\widetilde{L}_{jm}=V_{ij}\,,\quad\text{for } i,j>k\,, \tag{5.30}$$

that means that, by adopting the Cholesky decomposition in Eq. (5.26), part of the covariance matrix (usually the largest part) is exactly recovered. Therefore, $\widetilde{\mathbf{L}}$ Eq. (5.27) seems to be a generally good approximator to the real Cholesky decomposition of the covariance matrix $\mathbf{V}$. Plugging this equation into Eq. (5.12) immediately gives:

$$f_j\sim m_j+z_j s_j+\sum_{l=1}^{k}z_l\sigma_{j,\text{corr}}^{(l)}\,. \tag{5.31}$$

It is interesting to focus on similarities and differences with the formulas present in the literature. I will specifically consider Eq. (13) of Ref. [5] and Eq. (20) of Ref. [6].

Let us start by considering the case in which there are only uncorrelated and correlated additive uncertainties. In this case, Refs. [5, 6] are in mutual agreement and only in partial agreement with our Eq. (5.31), the difference being that the random numbers used to generate the fluctuation associated to the k correlated uncertainties are equal to the first k of the n used for the uncorrelated uncertainties (remember the requirement $k\leq n$). Of course, if $k\ll n$, as it often happens, the agreement between Eq. (5.31) and Refs. [5, 6] is asymptotically good.

When considering also correlated multiplicative uncertainties there are differences across formulas of Refs. [5, 6] and Eq. (5.31). While Eq. (5.31) treats multiplicative uncertainties on the same footing as additive uncertainties (consistently with the construction of the covariance matrix in Eq. (1.2)), Refs. [5, 6] introduce a multiplicative factor. In addition, Ref. [5] distinguishes between relative and absolute multiplicative uncertainties.

6 The D'Agostini "bias"

The main topic of this section is the so-called D'Agostini "bias" (DAB) originally presented in Ref. [2]. Quoting from the abstract of that paper: "Best fits to data which are affected by systematic uncertainties on the normalisation factor have the tendency to produce curves lower than expected [...]. This bias can become unacceptable if the normalisation error is large, or a large number of data points are used". In the following, I will summarise the arguments behind this conclusion. I will then move to giving my own view on this question arguing that the DAB is in fact not a bias but rather a feature of correlation uncertainties in general and not of multiplicative uncertainties in particular. This implies that the distinction between additive and normalisation uncertainties is effectively immaterial. Specifically, what really triggers the DAB is *not* the fact the uncertainty is on the multiplication factor but rather that it typically gives raise to correlated uncertainties that are different in *absolute* size. To show this, below I will show that the DAB takes place also in the presence of additive

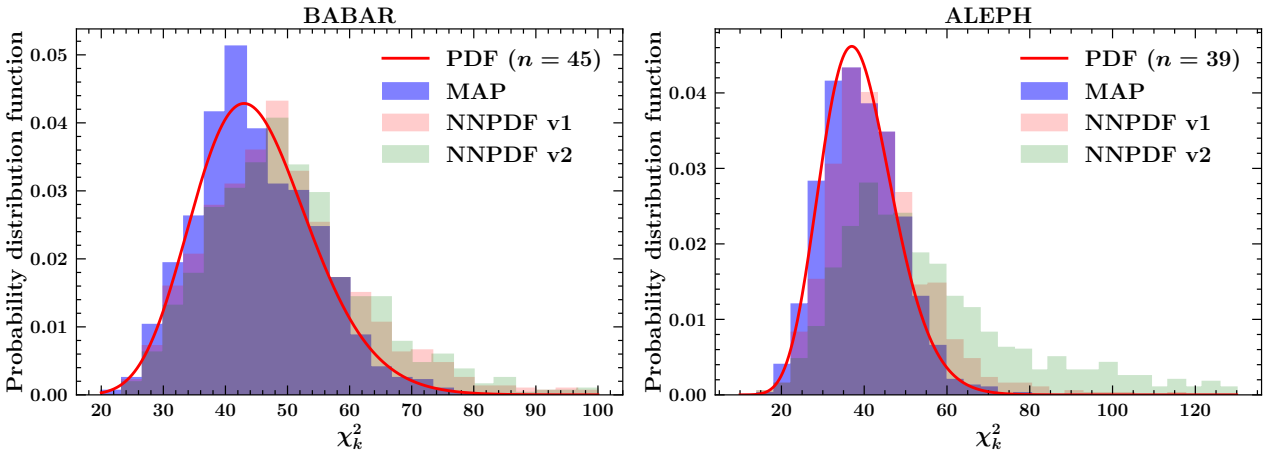

Fig. 5.1: Distribution of the χ^2's of the Monte Carlo replica w.r.t. to the corresponding central values for BABAR (left) and ALEPH (right) experiments compared to the probability distribution function with to the appropriate number of degrees of freedom given by the number of data points. The blue histograms are produced using Eq. (5.12), the red histogram is produced using Eq. (13) of Ref. [5] where multiplicative uncertainties are treated as relative, while the green histogram is produced using Eq. (20) of Ref. [6].

uncertainties that are different in absolute size. Conversely, I will show that the DAB does not takes place when considering multiplicative uncertainties and the central values are identical. The bottom line is that any difference in absolute size between correlated uncertainties causes a deviation from the expectation according to which the best fit value should be given by the weighted average of the experimental central values weighted by the inverse of the squared uncorrelated uncertainties. This deviation from the expectation is the origin of the DAB.

In order to present the argument, Ref. [2] considers a data set made of two data points with the respective uncorrelated uncertainty, $x_i \pm \sigma_i$ with $i = 1, 2$. In addition, it is assumed that x_i measure the same quantity and thus they can be fitted with a single constant μ. Two cases are treated. In the first case the two data points are also affected by a single offset (additive) correlated uncertainty σ_c. In the second case, instead, the data points are affected by a normalisation correlated uncertainty whose size is proportional to the central values through the factor δ_f. Given the simplicity of the problem, it is possible in both cases to analytically compute the best value of μ and its uncertainty $\delta\mu$ by minimising the χ^2 and computing its second derivative at the minimum. As usual the χ^2 is defined as:

$$\chi^2 = \sum_{i,j=1}^{2} y_i V_{ij}^{-1} y_j = \mathbf{y}^T \cdot \mathbf{V}^{-1} \cdot \mathbf{y}\,, \tag{6.1}$$

where $y_i = \mu - x_i$ and V_{ij}^{-1} are the entries of the inverse of the covariance matrix.

Let us start with the offset case in which the covariance matrix takes the form:

$$\mathbf{V} = \begin{pmatrix} \sigma_1^2 + \sigma_c^2 & \sigma_c^2 \\ \sigma_c^2 & \sigma_2^2 + \sigma_c^2 \end{pmatrix}, \tag{6.2}$$

whose inverse is:

$$\mathbf{V}^{-1} = \frac{1}{D} \begin{pmatrix} \sigma_2^2 + \sigma_c^2 & -\sigma_c^2 \\ -\sigma_c^2 & \sigma_1^2 + \sigma_c^2 \end{pmatrix}, \tag{6.3}$$

with $D = \sigma_1^2\sigma_2^2 + (\sigma_1^2 + \sigma_2^2)\sigma_c^2$. With this I can straightforwardly compute the χ^2:

$$\chi^2 = \frac{\sigma_2^2 y_1^2 + \sigma_1^2 y_2^2 + (y_1 - y_2)^2 \sigma_c^2}{D}\,, \tag{6.4}$$

that can be minimised w.r.t. to μ by requiring that:

$$\left.\frac{d\chi^2}{d\mu}\right|_{\mu=\overline{\mu}} = 2\left.\frac{\sigma_2^2 y_1 + \sigma_1^2 y_2}{D}\right|_{\mu=\overline{\mu}} = 0\,, \tag{6.5}$$

where $\overline{\mu}$ is the best fit value of μ. Notice the cancellation in the numerator of the correlated uncertainty σ_c. This "accidental" cancellation in the offset case, caused by the fact that σ_c is the same for both x_1 and x_2, can

be regarded as the (misleading) evidence that the DAB only takes place in the normalisation case. The solution to Eq. (6.5) is:

$$\overline{\mu} = \frac{\sigma_2^2 x_1 + \sigma_1^2 x_2}{\sigma_1^2 + \sigma_2^2} = \frac{\frac{x_1}{\sigma_1^2} + \frac{x_2}{\sigma_2^2}}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}}\,, \tag{6.6}$$

that fulfils the intuitive expectation that the best value of μ should be a weighted average of the central values weighted by the inverse of the uncorrelated uncertainties squared. One can also compute the uncertainty of $\overline{\mu}$, $\delta\mu$, by simply computing the inverse of the second derivative of the χ^2 at $\mu = \overline{\mu}$:[9]

$$\delta\mu^2 = \left[\frac{1}{2}\frac{d^2\chi^2}{d\mu^2}\right]^{-1}_{\mu=\overline{\mu}} = \frac{\sigma_1^2\sigma_2^2 + (\sigma_1^2 + \sigma_2^2)\sigma_c^2}{\sigma_2^2 + \sigma_1^2} = \frac{1 + \frac{\sigma_c^2}{\sigma_1^2} + \frac{\sigma_c^2}{\sigma_2^2}}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}}\,. \tag{6.7}$$

I now move to considering the normalisation case in which the covariance matrix reads:

$$\mathbf{V} = \begin{pmatrix} \sigma_1^2 + \delta_f^2 x_1^2 & \delta_f^2 x_1 x_2 \\ \delta_f^2 x_1 x_2 & \sigma_2^2 + \delta_f^2 x_2^2 \end{pmatrix}\,, \tag{6.8}$$

whose inverse is:

$$\mathbf{V}^{-1} = \frac{1}{D}\begin{pmatrix} \sigma_2^2 + \delta_f^2 x_2^2 & -\delta_f^2 x_1 x_2 \\ -\delta_f^2 x_1 x_2 & \sigma_1^2 + \delta_f^2 x_1^2 \end{pmatrix}\,, \tag{6.9}$$

with $D = \sigma_1^2\sigma_2^2 + \delta_f^2(\sigma_1^2 x_2^2 + \sigma_2^2 x_1^2)$. The resulting χ^2 is:

$$\chi^2 = \frac{1}{D}\left[\sigma_2^2 y_1^2 + \sigma_1^2 y_2^2 + \delta_f^2 (x_2 y_1 - x_1 y_2)^2\right]\,. \tag{6.10}$$

The minimum is computed as:

$$\left.\frac{d\chi^2}{d\mu}\right|_{\mu=\overline{\mu}} = \left. 2\frac{\sigma_2^2 y_1 + \sigma_1^2 y_2 + \delta_f^2 (x_2 - x_1)(x_2 y_1 - x_1 y_2)}{D}\right|_{\mu=\overline{\mu}} = 0\,. \tag{6.11}$$

This time, differently from the offset case, the correlated uncertainties in the denominator do not cancel and the result of the equation above is:

$$\overline{\mu} = \frac{\sigma_2^2 x_1 + \sigma_1^2 x_2}{\sigma_1^2 + \sigma_2^2 + (\delta_f x_1 - \delta_f x_2)^2} = \frac{\frac{x_1}{\sigma_1^2} + \frac{x_2}{\sigma_2^2}}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2} + \left(\frac{\delta_f x_1 - \delta_f x_2}{\sigma_1\sigma_2}\right)^2}\,, \tag{6.12}$$

while:

$$\delta\mu^2 = \left[\frac{1}{2}\frac{d^2\chi^2}{d\mu^2}\right]^{-1}_{\mu=\overline{\mu}} = \frac{\sigma_1^2\sigma_2^2 + \delta_f^2(\sigma_1^2 x_2^2 + \sigma_2^2 x_1^2)}{\sigma_2^2 + \sigma_1^2 + \delta_f^2 (x_1 - x_2)^2} = \frac{1 + \frac{\delta_f^2 x_1^2}{\sigma_1^2} + \frac{\delta_f^2 x_2^2}{\sigma_2^2}}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2} + \left(\frac{\delta_f x_1 - \delta_f x_2}{\sigma_1\sigma_2}\right)^2}\,. \tag{6.13}$$

According to Ref. [2], the term proportional to δ_f in the denominator of the r.h.s. of Eq. (6.12) is a source of bias in that it deviates from the intuitive expectation of Eq. (6.6). Being this additional term positive, it tends to diminish the best fit value.

I will now show that the additional term in the denominator of Eq. (6.12) is not a specific feature of normalisation uncertainties. To show this, I consider the same data set of two data points but where the data point x_1 is affected by the additive uncertainty $\sigma_{c,1}$ and the point x_2 by the additive uncertainty $\sigma_{c,2}$. According to the same procedure followed above the final result for best-fit value and uncertainty of μ is:

$$\overline{\mu} = \frac{\frac{x_1}{\sigma_1^2} + \frac{x_2}{\sigma_2^2}}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2} + \left(\frac{\sigma_{c,1} - \sigma_{c,2}}{\sigma_1\sigma_2}\right)^2}\,, \tag{6.14}$$

and:

$$\delta\mu^2 = \frac{1 + \frac{\sigma_{c,1}^2}{\sigma_1^2} + \frac{\sigma_{c,2}^2}{\sigma_2^2}}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2} + \left(\frac{\sigma_{c,1} - \sigma_{c,2}}{\sigma_1\sigma_2}\right)^2}\,. \tag{6.15}$$

[9] Actually, the second derivative of the χ^2 does not depend on μ and thus it is not necessary to set $\mu = \overline{\mu}$.

These equations are (somewhat expectedly) equal to Eqs. (6.12) and (6.13) with the only difference that $\delta_f x_i$ is replaced by $\sigma_{c,i}$. Therefore, the DAB shows up also when additive uncertainties are considered as long as they are different in absolute value. Therefore, the DAB is not a prerogative of normalisation uncertainties only. As expected, the equations above reduce to Eqs. (6.6) and (6.7) when $\sigma_{c,1} = \sigma_{c,2} = \sigma_c$.

Another important aspect of Eqs. (6.12) and (6.13) is that the term that causes the "bias" is proportional to the difference $x_1 - x_2$. Therefore, if the central values happen to be the same that term vanishes and thus no bias is present. This observation, along with the fact that the DAB is potentially present also when additive uncertainties are considered, leads to a simple conclusion: the DAB is present in all cases whenever there are correlated uncertainties different in absolute value. This is trivially the case when one introduces different correlated additive uncertainties, $\sigma_{c,1}$ and $\sigma_{c,2}$ with $\sigma_{c,1} \neq \sigma_{c,2}$. But this also the case in the presence of a normalisation uncertainties when central values are different ($x_1 \neq x_2$). On the contrary, if $\sigma_{c,1} = \sigma_{c,2}$ in the additive case or $x_1 = x_2$ in the multiplicative case, no "bias" is present. The conclusion is that the very distinction between additive and multiplicative uncertainties is immaterial.

Now that I have established that the DAB is not a prerogative of normalisation uncertainties only, I will argue that this is in fact not a bias but a manifestation of large correlated uncertainties in general that indicates an intrinsic inconsistency of the data set being considered. I start by observing that, in the presence of correlations, a visual comparison between data and best fit value is meaningful only when including the systematic shifts introduces in Sect. 3. The task is particularly simple due to the small number of points and the fact that there one single correlated uncertainty. In particular, there is a single nuisance parameter that in the normalisation case is given by:

$$\lambda = \frac{\frac{\delta_f x_1^2}{\sigma_1^2} + \frac{\delta_f x_2^2}{\sigma_2^2}}{1 + \frac{\delta_f^2 x_1^2}{\sigma_1^2} + \frac{\delta_f^2 x_2^2}{\sigma_2^2}} - \overline{\mu}\frac{\frac{\delta_f x_1}{\sigma_1^2} + \frac{\delta_f x_2}{\sigma_2^2}}{1 + \frac{\delta_f^2 x_1^2}{\sigma_1^2} + \frac{\delta_f^2 x_2^2}{\sigma_2^2}} \,. \tag{6.16}$$

It is the quantity:

$$\hat{\mu}_i = \overline{\mu} + \lambda \delta_f x_i \,, \tag{6.17}$$

that can be directly compared to the data-point central values. Unfortunately, the analytic form of this quantity is not very informative but to see the effect of the systematic shifts I consider the very same example given in the introduction of Ref. [2], *i.e.* a data set of two measurements: $8.0 \pm 2\%$ and $8.5 \pm 2\%$ having a common 10% normalisation uncertainty.

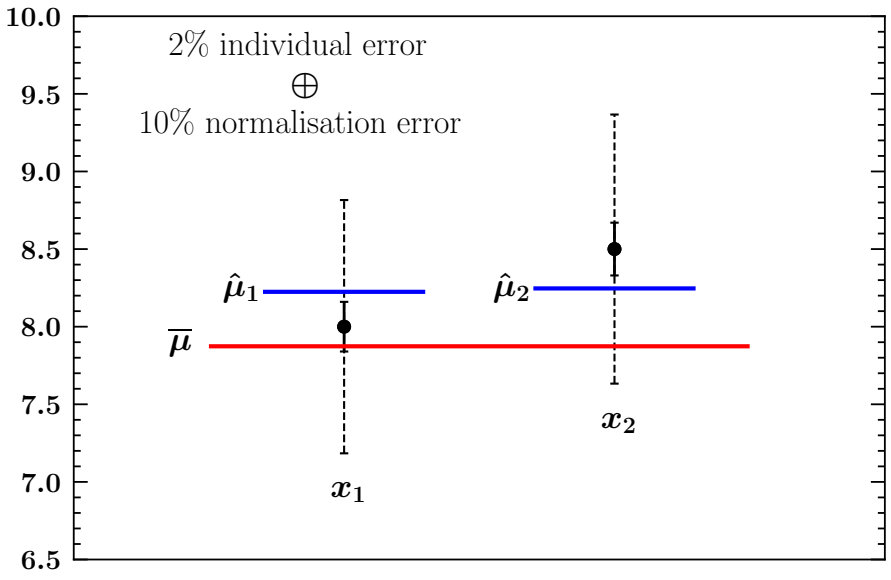

Fig. 6.1: Numerical example taken from Ref. [2] of a data set of two points with central value 8.0 and 8.5 (black points) each affected by a 2% uncorrelated uncertainty (solid error bars) and a common 10% normalisation uncertainty (the dashed error bars indicate the sum in quadrature of correlated and uncorrelated uncertainties). The horizontal red line corresponds to the best fit value given in Eq. (6.6) while the horizontal blue lines correspond to the shifted values according to Eq. (6.17).

Fig. 6.1 shows the result of minimising the χ^2 to determine the best fit value of the parameter μ. The solid error bars display the uncorrelated uncertainties while the dashed error bars display the sum in quadrature of

correlated and uncorrelated uncertainties. The best fit value $\overline{\mu}$, Eq. (6.6), is represented by the horizontal red line. As argued in Fig. 6.1, this line lies below both the central experimental values ($\overline{\mu} \simeq 7.87$) contradicting the common expectation that best fit value should be somewhere between the experimental central values. However, as explained above, a visual comparison in the presence of (large) correlations is meaningful only when including the systematic shifts. The inclusion of the shifts yields the two horizontal blue lines in Fig. 6.1. The shifted values are $\hat{\mu}_1 \simeq 8.22$ and $\hat{\mu}_2 \simeq 8.25$ that indeed fulfil the common expectation. It is worth noticing that the best fit χ^2 per degree of freedom is around 4.4, signifying that the measurements, that are assumed to refer to the same quantity, are in strong tension. This is a first suggestion that that there is no problem with the standard inclusion of correlated normalisation uncertainties in the covariance matrix while the problem seems to be in the data set.

One may ask what is the meaning of a value of $\overline{\mu}$ that can be smaller than both x_1 and x_2 or in general far from both of them. To answer this question I first observe that, according to Eq. (6.6), $\overline{\mu}$ tends to zero as the uncorrelated uncertainties σ_1 and σ_2 tend to zero. This can be seen more explicitly by setting $\sigma_1 = \sigma_2 = \varepsilon$ and expanding the χ^2 around $\varepsilon = 0$:

$$\chi^2 \to \frac{\mu^2(x_1 - x_2)^2}{\varepsilon^2(x_1^2 + x_2^2)} + \frac{\left[\mu(x_1 + x_2) - (x_1^2 + x_2^2)\right]^2}{\delta_f^2(x_1^2 + x_2^2)^2} + \mathcal{O}(\varepsilon^2)\,. \tag{6.18}$$

Therefore, if $x_1 \neq x_2$, for $\varepsilon \to 0$ the dominant term is the first one and the minimum of the χ^2 in this limit is clearly at $\mu = 0$. If $x_1 = x_2 = x$, instead, the first term drops and the best fit value of the parameter μ is determined by the second term that has its minimum at $\mu = x$.

This can be interpreted as follows: as σ_1 and σ_2 tend to zero, the covariance matrix tends to become singular signifying that x_1 and x_2 are no longer independent. However, if $x_1 = x_2 = x$ the singularity cancels. This is due to the fact that, without uncorrelated uncertainties, the number of points of the data set effectively reduces to one. As a matter of fact, in this case ε can safely be set to zero and the χ^2 becomes:

$$\chi^2 = \left(\frac{\mu - x}{\delta_f x}\right)^2\,, \tag{6.19}$$

that is evidently the χ^2 associated to one single data point with central value x and uncertainty $\delta_f x$.[10] As a consequence, the best fit value of the parameters μ is simply given by x, as one would expect.

If $x_1 \neq x_2$, instead, the singularity of the covariance matrix does not cancel and the first term in Eq. (6.18) dominates in the limit $\varepsilon \to 0$. This is a situation in which two totally dependent measurements of the same quantity have different central values. It is thus to be expected that the χ^2 becomes arbitrarily large because the two data points are inconsistent and thus impossible to be simultaneously fitted. In this situation, the parameter μ attempts to minimise the χ^2 by tending to zero. More specifically, when correlations dominate over uncorrelated uncertainties, the χ^2 is proportional to (the square of) the cross difference $\delta_f x_1 y_2 - \delta_f x_2 y_1$,[11] with $y_i = \mu - x_i$. Therefore, if $x_1 \neq x_2$, this difference is minimal when $y_i \simeq x_i$ which is equivalent to $\mu \simeq 0$. It is also interesting to observe that the first term in Eq. (6.18) does not depend of δ_f. Therefore, the size of the χ^2 is entirely driven by the difference $x_1 - x_2$, no matter the precise value of normalisation uncertainty.

In conclusion, it appears that data sets dominated by normalisation (or equivalently widely spread correlated) uncertainties carry an intrinsic degree of inconsistency that, when attempting to fit them, manifests into the DAB. Therefore, from this point of view, the DAB is not a flaw of the fitting procedure but rather an indication that the data set is intrinsically inconsistent. It thus appears counterproductive, if not wrong, to adjust the definition of the χ^2 to accommodate an inconsistency of the data set.

6.1 The effect of the t_0 prescription

In this section, I consider the effect of the iterative t_0 prescription on the best fit value $\overline{\mu}$. The t_0 prescription essentially replaces the normalisation uncertainties $\delta_f x_1$ and $\delta_f x_2$ with one single correlated uncertainties $\delta_f t_0$, where t_0 is a "guess" for μ. The net effect of this replacement is that the terms $\delta_f x_1$ and $\delta_f x_2$ in the r.h.s. of both Eqs. (6.12) and (6.13) are replaced by $\delta_f t_0$ and thus cancel, reducing to Eqs.(6.6) and (6.7). But this should not be surprising because the t_0 prescription is essentially replacing two correlated uncertainties that are different in value with two correlated uncertainties that are instead *equal*. This makes the normalisation case artificially equal to the offset case in which the DAB is not present.

[10] Note that in the case of one-point data set there is no meaning in talking about correlated uncertainties. Therefore, $\delta_f x$ plays the role of an uncorrelated uncertainty.

[11] In general, this can be seen as $\sigma_{c,1} y_2 - \sigma_{c,2} y_1$.

Now, what happens if one iterates the t_0? Let us indicate with $\overline{\mu}_i$ and $\delta\mu_i^2$ best fit value and variance at the i-th iteration. As explained above, the zero-th iteration gives:

$$\overline{\mu}_0 = \frac{\frac{x_1}{\sigma_1^2} + \frac{x_2}{\sigma_2^2}}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}}\,, \tag{6.20}$$

and:

$$\delta\mu_0^2 = \frac{1 + \frac{\delta_f^2 t_0^2}{\sigma_1^2} + \frac{\delta_f^2 t_0^2}{\sigma_2^2}}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}}\,. \tag{6.21}$$

Since $\overline{\mu}_0$ does not depend on t_0, all following iterations will give the exact same result, *i.e.* $\overline{\mu}_i = \overline{\mu}_0\ \forall i > 0$. The stabilisation happens only one step further for the uncertainty for which one finds $\delta\mu_i^2 = \delta\mu_1^2\ \forall i > 1$, with:

$$\delta\mu_1^2 = \frac{1 + \frac{\delta_f^2 \overline{\mu}_0^2}{\sigma_1^2} + \frac{\delta_f^2 \overline{\mu}_0^2}{\sigma_2^2}}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}} = \frac{1}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}} + \delta_f^2 \overline{\mu}_0^2\,. \tag{6.22}$$

As above, I set $\sigma_1 = \sigma_2 = \varepsilon$ getting:

$$\overline{\mu}_0 = \frac{x_1 + x_2}{2}\,, \tag{6.23}$$

and:

$$\delta\mu_1^2 = \frac{\varepsilon^2}{2} + \frac{\delta_f^2 (x_1 + x_2)^2}{4}\,. \tag{6.24}$$

Now, I plug $\overline{\mu}_0$ into the original definition of the χ^2. The result is:

$$\chi^2 = \left(\frac{x_1 - x_2}{2}\right)^2 \left[\frac{1}{\varepsilon^2 + \delta_f^2 (x_1^2 + x_2^2)} + \frac{1}{\varepsilon^2}\right]\,. \tag{6.25}$$

References

[1] R. D. Ball *et al.*, "Parton Distribution Benchmarking with LHC Data," *JHEP*, vol. 04, p. 125, 2013.

[2] G. D'Agostini, "On the use of the covariance matrix to fit correlated data," *Nucl. Instrum. Meth. A*, vol. 346, pp. 306–311, 1994.

[3] R. D. Ball, L. Del Debbio, S. Forte, A. Guffanti, J. I. Latorre, J. Rojo, and M. Ubiali, "Fitting Parton Distribution Data with Multiplicative Normalization Uncertainties," *JHEP*, vol. 05, p. 075, 2010.

[4] W. contributors, "Chi-square distribution — Wikipedia, The Free Encyclopedia," 2020.

[5] R. D. Ball, L. Del Debbio, S. Forte, A. Guffanti, J. I. Latorre, A. Piccione, J. Rojo, and M. Ubiali, "A Determination of parton distributions with faithful uncertainty estimation," *Nucl. Phys. B*, vol. 809, pp. 1–63, 2009. [Erratum: Nucl.Phys.B 816, 293 (2009)].

[6] R. D. Ball *et al.*, "Parton distributions for the LHC Run II," *JHEP*, vol. 04, p. 040, 2015.