Correlation of theoretical uncertainties in PDF fits and theoretical uncertainties in predictions

Richard D. Ball^a, Rosalyn L. Pearson

The Higgs Centre for Theoretical Physics, University of Edinburgh, JCMB, KB, Mayfield Rd, Edinburgh EH9 3FD, Scotland, UK

Received: 11 May 2021 / Accepted: 28 August 2021 / Published online: 18 September 2021
© The Author(s) 2021

Abstract We show how to account for correlations between theoretical uncertainties incorporated in parton distribution function (PDF) fits, and the theoretical uncertainties in the predictions made using these PDFs. We demonstrate by explicit calculations, both analytical and numerical, that these correlations can lead to corrections to the central values of the predictions, and reductions in both the PDF uncertainties and the theoretical uncertainties in the prediction. We illustrate our results with predictions for top production rapidity distributions and the Higgs total cross-section at the LHC, using the NLO NNPDF3.1 PDF set which incorporates missing higher order uncertainties. We conclude that the inclusion of correlations can increase both the accuracy and precision of predictions involving PDFs, particularly for processes with data already included in the PDF fit.

1 Introduction

Parton distribution functions (PDFs) play a crucial role in the prediction of Standard Model observables at the Large Hadron Collider (LHC) and beyond [1, 2]. PDF uncertainties are already a limiting factor in these predictions [3]. This means that a precise and accurate knowledge of PDFs is essential for full exploitation of the physics at LHC. Besides the familiar experimental uncertainties in the data that are input to a global PDF fit, it is becoming increasingly necessary to also consider theoretical uncertainties: missing higher order uncertainties (MHOUs) due to the truncation of perturbative expansions in the hard cross-sections and parton evolution, uncertainties due to the use of nuclear targets, uncertainties due to missing higher twist, uncertainties due to external parameters such as quark masses, showering and hadronization uncertainties in final state Monte Carlos, and so on. In the past the technique used to estimate the effect of theoretical uncertainties was to compare the PDFs produced

with and without various theoretical corrections, rather than to incorporate the uncertainties into the fit itself. However the limitations of such a procedure are clear.

Recently a new approach to estimating the impact of theoretical uncertainties on global PDF fits has been developed, through the construction of a ‘theory covariance matrix’ [4], analogous to the experimental covariance matrix used in global PDF fits. By adding the theory covariance matrix to the experimental covariance matrix as an additional source of uncertainty, theoretical uncertainties can be incorporated directly in the fit, where they impact not only the overall level of PDF uncertainty, but also the relative weight of different datasets. This novel approach has so far been applied to uncertainties associated with nuclear effects [5, 6], and to the estimation of MHOUs by scale variation [7–9] (though other methods [10–12] of estimating MHOUs are under development). Factorization scale variations estimate the MHOUs in parton evolution (correlated across all observables), and renormalization scale variations estimate the MHOUs in fixed order calculations of process-dependent hard cross-section (correlated only across a given class of processes). The first global PDFs including MHOUs estimated in this way were presented in [9].

When making predictions for hadronic observables there are again two sources of MHOUs: uncertainties in the PDF evolution, which can also be estimated by factorization scale variation, and uncertainties in the hard cross-section, again estimated by renormalization scale variation. These arise on top of the MHOUs incorporated in the determination of the PDFs themselves, manifested as a part of the PDF uncertainty. So there will be correlations between the MHOUs in the PDFs and the additional MHOUs in the predictions. The PDFs themselves contain a wealth of data from a wide range of processes. When making a prediction for one of these processes (for example in a new kinematic regime), there will inevitably be correlations between the renormalization scale variation in the PDF determination, and that in the predic-

^a e-mail: rdب@ph.ed.ac.uk (corresponding author)

tion. Even if the process is a new one, for example Higgs production, correlations due to factorization scale variation will still be present: all processes dependent on PDFs have a MHOU due to the need to evolve the PDFs to the scale of the process.

The potential importance of these correlations can be exposed by considering a simple situation in which a PDF is determined from data at a given scale on a single observable (such as a nonsinglet structure function), and used to make predictions of another observable [13]. In this special case the PDF can be eliminated altogether, and the predicted observable determined directly from the measured observable. Including the MHOU twice (once in the calculation of the measured observable, then again in making the predicted observable), whilst ignoring their possible correlation, would amount to “double counting”, and thus an overestimate of the uncertainty. Of course in a global PDF fit, involving many different processes, this particular formulation is no longer applicable. However the fact remains that there will still be some residual correlation between the MHOU in the PDF determination and the MHOU in the prediction, and to neglect it may lead to an overestimation of the MHOU.

In Ref. [9], it was seen that in a realistic NLO global fit, the effect of MHOU on the PDF uncertainty is quite small, the main consequence being a rebalancing of the impact of different datasets depending on their relative MHOU. The MHOU in the PDF is consequently rather smaller than the MHOU in the prediction, and if they are combined in quadrature, the effect of missing correlations will likely also be small. It was further argued that combining the PDF uncertainties and theoretical uncertainties in quadrature is conservative, since it can only lead to an overestimate of uncertainties, and thus better than neglecting MHOU altogether. Nevertheless, since the correlation between the uncertainties has not been computed explicitly, there remains the intriguing possibility that in some circumstances including the correlation may yield more precise, and perhaps even more accurate, predictions.

In this paper we show explicitly how to propagate theoretical uncertainties in the determination of global PDFs into the predictions made using these PDFs, taking account of all correlations between theoretical uncertainties. As this is a complicated problem, we proceed step by step. In Sect. 2 we show how a single source of theoretical uncertainty can be reformulated in terms of a nuisance parameter, which holds the key to the propagation of uncertainties. We explore the interplay between the theoretical uncertainties and the experimental uncertainties in the data in two idealised contexts: the first in which there are no fitted parameters, the second in which there as many parameters as data, so the experimental data can be fitted exactly. This allows us to identify three distinct effects of the correlation of theoretical uncertainties. Then in Sect. 3 we consider a more realistic, but still simplified, sit-

uation where we fit the data using a single parameter, still in a theory with a single source of theoretical uncertainty, and find that all three of these correlation effects are still present, and can be computed. In Sect. 4 these results are extended to multiple theoretical uncertainties, in a multiparameter fit, and then less trivially to a PDF fit where the PDFs are continuous functions, with a functional uncertainty. Finally in Sect. 5 we present numerical results, made in the context of the NNPDF3.1 NLO global fit with MHOU presented in Ref. [9], and make predictions with MHOU for repetitions of the experiments included in the fit (so-called ‘autopredictions’), and for genuine predictions (top and Higgs production). We are able to compute all correlations explicitly, calculating corrections to central values and changes in PDF and theoretical uncertainties for both autopredictions and genuine predictions, and confirm that the correlated predictions can be both more accurate and more precise than the conservative prescription. A summary is provided in Sect. 6.

2 Predictions with correlated theoretical uncertainties

We showed in Refs. [4,5] that when we fit N experimental data points D_i to theoretical predictions T_i , $i = 1, \dots, N$, then the uncertainties in the theoretical predictions can be incorporated into the fit simply by adding a theoretical covariance matrix S_{ij} to the experimental covariance matrix C_{ij} . The only assumptions made in deriving this result are that all uncertainties, both experimental and theoretical, can be treated as Gaussian, and that the theoretical uncertainties are independent of the experimental data. We used this result in Refs. [5,6] to include nuclear uncertainties in a PDF fit, and in Refs. [8,9] to incorporate missing higher order uncertainties in global PDF fits.

The result of Sec. 2 of Ref. [5] may be summarised in terms of the Bayesian probability

$$P(T|D) \propto \exp\left(-\frac{1}{2}(T-D)^T(C+S)^{-1}(T-D)\right) \quad (2.1)$$

where for simplicity we adopt a matrix notation: C and S are real symmetric matrices, with C strictly positive definite (thus invertible), and S positive semi-definite. In practice theoretical uncertainties are highly correlated, so S may (and generally will) have zero eigenvalues. We determine T from D by maximizing $P(T|D)$: this is equivalent to minimizing

$$\chi^2 = (T-D)^T(C+S)^{-1}(T-D) \quad (2.2)$$

with respect to free parameters characterizing the theoretical prediction.

In this section we will also assume that there is only a single source of fully correlated theoretical uncertainty, so

that the theory covariance matrix can be written as

$$S = \beta\beta^T \quad (2.3)$$

for some real nonzero vector β (i.e. that $S_{ij} = \beta_i\beta_j$). Then all eigenvalues of S except one are zero. We will develop a nuisance parameter formalism to propagate this theoretical uncertainty (Sect. 2.1), and then show how to apply it in two extreme cases: firstly to the situation in which the theory contains no free parameters, and thus where there is no fitting (Sect. 2.2), and then to the situation where there as many free parameters as data points, so that we can achieve a perfect fit (Sect. 2.3).

2.1 Nuisance parameters

It will be useful in what follows to introduce a nuisance parameter λ for the correlated theoretical uncertainty. Following the notation in Ref. [14], we model the theoretical uncertainty as a fully correlated shift in the theoretical prediction: $T \rightarrow T + \lambda\beta$. The nuisance parameter λ then gives the size of the shift. For a given shift, with Gaussian experimental uncertainties

$$P(T|D\lambda) \propto \exp\left(-\frac{1}{2}(T + \lambda\beta - D)^T C^{-1}(T + \lambda\beta - D)\right). \quad (2.4)$$

Now using Bayes' Theorem,

$$P(T|D\lambda)P(\lambda|D) = P(\lambda|TD)P(T|D). \quad (2.5)$$

To determine $P(T|D)$ we need to first fix the prior distribution of λ . Since this is a theoretical uncertainty, it is reasonable to assume that the prior is independent of the data, thus that $P(\lambda|D) = P(\lambda)$. Then, marginalizing Eq. (2.5) over λ ,

$$P(T|D) = \int d\lambda P(T|D\lambda)P(\lambda). \quad (2.6)$$

Taking $P(\lambda)$ as a Gaussian, centred on zero (so the theoretical uncertainty is unbiased), with unit width (fixing the overall normalization of the prior theoretical uncertainty), we choose

$$P(\lambda) \propto \exp\left(-\frac{1}{2}\lambda^2\right). \quad (2.7)$$

The integration over the nuisance parameter is now Gaussian,

$$P(T|D) \propto \int d\lambda \times \exp\left(-\frac{1}{2}[(T + \lambda\beta - D)^T C^{-1}(T + \lambda\beta - D) + \lambda^2]\right), \quad (2.8)$$

and can be performed in the usual way by completing the square:

$$(T + \lambda\beta - D)^T C^{-1}(T + \lambda\beta - D) + \lambda^2$$

$$= Z^{-1}\left(\lambda + Z\beta^T C^{-1}(T - D)\right)^2 + (T - D)^T C^{-1}(T - D) - Z(\beta^T C^{-1}(T - D))^2, \quad (2.9)$$

where we have defined

$$Z = (1 + \beta^T C^{-1}\beta)^{-1} = 1 - \beta^T(C + S)^{-1}\beta. \quad (2.10)$$

The second expression was obtained by noting that

$$(1 + \beta^T C^{-1}\beta)(1 - \beta^T(C + S)^{-1}\beta) = 1 \quad (2.11)$$

using Eq. (2.3).

Since

$$(\beta^T C^{-1}(T - D))^2 = (T - D)^T C^{-1}\beta\beta^T C^{-1}(T - D), \quad (2.12)$$

we can combine the two terms on the second line of Eq. (2.9) to give

$$(T - D)^T (C^{-1} - ZC^{-1}\beta\beta^T C^{-1})(T - D) = (T - D)^T (C + \beta\beta^T)^{-1}(T - D), \quad (2.13)$$

since

$$(C + \beta\beta^T)(C^{-1} - ZC^{-1}\beta\beta^T C^{-1}) = 1 + \beta\beta^T C^{-1} - Z\beta\beta^T C^{-1} - Z\beta\beta^T C^{-1}\beta\beta^T C^{-1} = 1, \quad (2.14)$$

by substituting (from Eq. (2.10)) $\beta^T C^{-1}\beta = Z^{-1} - 1$ in the last term. Using the definition Eq. (2.3), we recognise Eq. (2.13) as the χ^2 , Eq. (2.2). Defining

$$\bar{\lambda}(T, D) = Z\beta^T C^{-1}(D - T) = \beta^T(C + S)^{-1}(D - T), \quad (2.15)$$

we can thus write Eq. (2.8) as

$$P(T|D) \propto \int d\lambda \exp\left(-\frac{1}{2}Z^{-1}(\lambda - \bar{\lambda})^2 - \frac{1}{2}\chi^2\right) \propto \exp\left(-\frac{1}{2}\chi^2\right), \quad (2.16)$$

since the integration over λ yields a factor $(2\pi Z)^{1/2}$ which is independent of T and D .

The nuisance parameter formalism thus produces our original result Eq. (2.1). However it is useful because it allows us to also determine the posterior distribution of the nuisance parameter: inverting Eq. (2.5)

$$P(\lambda|TD) \propto \exp\left(-\frac{1}{2}Z^{-1}(\lambda - \bar{\lambda}(T, D))^2\right), \quad (2.17)$$

so once we use the information on T and D , the prior distribution Eq. (2.7) is modified. In particular the peak of the distribution is shifted away from zero to $\bar{\lambda}$, while the width is now given by Z . It is easy to see from the definition Eq. (2.10) that

$$0 < Z < 1, \quad (2.18)$$

so the width of the theoretical uncertainty is generally reduced by the addition of new information.

2.2 Predictions and autopredictions without fits

In order to explore the implications of the formalism described in the previous section we first consider a ‘pure’ theory $T = T_0$, one where there are no unknown parameters to be fitted, but which nevertheless has a theoretical uncertainty. Despite the fact that this theory cannot be fitted to the data, so in general $T_0 \neq D$, the data can still inform the theoretical predictions.

Computing expectation values of functions of λ using the probability distribution $P(\lambda|T_0D)$,

$$E[\lambda] = \mathcal{N}_\lambda \int d\lambda \lambda P(\lambda|T_0D) = \bar{\lambda}(T_0, D), \quad (2.19)$$

the normalization $\mathcal{N}_\lambda$ being chosen such that $E[1] = 1$, while

$$\text{Var}[\lambda] \equiv E[(\lambda - E[\lambda])^2] = Z. \quad (2.20)$$

Now in the nuisance parameter formalism, the theoretical predictions are

$$T(\lambda) = T_0 + \lambda\beta. \quad (2.21)$$

If we made these predictions before comparing to any data, we would use the prior distribution Eq. (2.7) for λ . Since this is a unit gaussian centred on zero, we would then find that $E[T(\lambda)] = T_0$, while $\text{Cov}[T(\lambda)] = \beta\beta^T = S$ as expected.

However if instead we first compare T to the data D , then we can instead compute expectation values using $P(\lambda|TD)$. We then find (using Eqs. (2.19, 2.15))

$$\begin{aligned} E[T(\lambda)] &= T_0 + \bar{\lambda}(T_0, D)\beta \\ &= T_0 + \beta\beta^T(C + S)^{-1}(D - T_0), \end{aligned} \quad (2.22)$$

while (using Eq. (2.20))

$$\begin{aligned} \text{Cov}[T(\lambda)] &\equiv E[(T(\lambda) - E[T(\lambda)])(T(\lambda) - E[T(\lambda)])^T] \\ &= \text{Var}[\lambda]\beta\beta^T = ZS. \end{aligned} \quad (2.23)$$

One can consider this as an ‘autoprediction’: first the theory is compared to the data, and then using this information one makes new theoretical predictions for precise repetitions of the same experiments. The original theoretical predictions T_0 are then shifted by an amount

$$\delta T = -S(C + S)^{-1}(T_0 - D) \quad (2.24)$$

and their theoretical uncertainties are reduced by a factor of $\sqrt{Z}$, since the data add new information: the covariance matrix of the autopredictions is

$$ZS = S - S(C + S)^{-1}S = C(C + S)^{-1}S = S(C + S)^{-1}C. \quad (2.25)$$

This is a simple example of Bayesian learning: the theory ‘learns’ from the data, within the constraints imposed by the prior theoretical uncertainty.

It is interesting to compare the experimental χ^2 of the original predictions

$$\chi_{\text{exp}}^2 = (T_0 - D)^T C^{-1}(T_0 - D) \quad (2.26)$$

to that obtained with the autopredictions

$$\begin{aligned} \chi_{\text{auto}}^2 &= (T_0 + \delta T - D)^T C^{-1}(T_0 + \delta T - D) \\ &= (T_0 - D)^T (C + S)^{-1}C(C + S)^{-1}(T_0 - D), \end{aligned} \quad (2.27)$$

since

$$T_0 + \delta T - D = C(C + S)^{-1}(T_0 - D). \quad (2.28)$$

It is easy to see that $\chi_{\text{auto}}^2 \leq \chi_{\text{exp}}^2$ since C is positive definite (i.e. $C > 0$) and S semi-positive definite (i.e. $S \geq 0$), so $2S + SC^{-1}S \geq 0$, whence $(C + S)C^{-1}(C + S) \geq C$, and $(C + S)^{-1}C(C + S)^{-1} \leq C^{-1}$. So the data induced shifts are always such as to improve the quality of the fit to the data, by exploiting the theoretical uncertainty.

To make this more explicit, consider a simple model in which the experimental uncertainties are uncorrelated, and the same for each data point: then we can write

$$C = \sigma^2 1, \quad S = s^2 ee^T, \quad (2.29)$$

with e a unit vector, $e^T e = 1$, and $\beta = se$ so $S = \beta\beta^T$. Thus σ is the experimental uncertainty on each data point, and $s/\sqrt{N}$ is the size of the correlated theoretical uncertainty. Then it is easy to see that

$$(C + S)^{-1} = \frac{1}{\sigma^2} \left(1 - \frac{s^2}{\sigma^2 + s^2} ee^T \right), \quad (2.30)$$

and (using Eq. (2.10))

$$Z = (1 + s^2/\sigma^2)^{-1}. \quad (2.31)$$

So the reduction in theoretical uncertainties depends on the ratio s^2/σ^2 : when $s^2 \ll \sigma^2$ the influence of the data on the theoretical uncertainty is very small, while when $s^2 \gg \sigma^2$ the size of the theoretical uncertainty is reduced from s to σ , since in the limit $s^2/\sigma^2 \rightarrow \infty$

$$ZS = \frac{\sigma^2 s^2}{\sigma^2 + s^2} ee^T \rightarrow \sigma^2 ee^T. \quad (2.32)$$

Note that when the theoretical uncertainties are comparable to the experimental, $s^2/N \sim \sigma^2$, and $Z \sim 1/(N+1)$: if there are a large number of independent data points, the reduction of the theoretical uncertainty can be very substantial.

In this model the shift in the theoretical predictions is (using Eq. (2.24))

$$\delta T = -\frac{s^2}{\sigma^2 + s^2} (e^T (T_0 - D)) e, \quad (2.33)$$

as expected in the direction e of the theoretical uncertainty. When $s^2/\sigma^2 \rightarrow \infty$, $e^T(T_0 + \delta T) \rightarrow e^T D$, so in this direction the autopredictions coincide precisely with the data. The experimental χ^2 for the autopredictions is (using Eq. (2.27))

$$\chi_{\text{auto}}^2 = (T_0 - D)^T \frac{1}{\sigma^2} \left(1 - \frac{s^2(s^2 + 2\sigma^2)}{(s^2 + \sigma^2)^2} e e^T \right) (T_0 - D), \quad (2.34)$$

So $N - 1$ contributions to the χ^2 orthogonal to e are unchanged, while the contribution along e is reduced by a factor Z^2 . The autopredictions will in general have a χ^2 of size $N - 1$, rather than the N of the original predictions, as naively expected since the nuisance parameter is effectively fitted.

Of course we can also consider genuine predictions $\tilde{T}_I$, $I = 1, \dots, \tilde{N}$, which again have no free parameters, but have a theoretical uncertainty which is correlated to that of the observables T_i , $i = 1, \dots, N$ for which we have the data D_i . The theoretical predictions including the theoretical uncertainty may be written

$$\tilde{T}(\tilde{\lambda}) = \tilde{T} + \tilde{\lambda} \tilde{\beta}, \quad (2.35)$$

where the vector $\tilde{\beta}_I$ gives the size and direction of the theoretical uncertainty in $\tilde{T}_I$. Now if the nuisance parameters $\tilde{\lambda}$ are independent of the parameters λ , the theoretical uncertainties in $\tilde{T}$ are uncorrelated with those in T , and the theory covariance matrix for the prediction $\tilde{T}$ is given by

$$\tilde{S} = \tilde{\beta} \tilde{\beta}^T, \quad (2.36)$$

However if they are correlated, $\tilde{\lambda} = \lambda$, and then taking advantage of the data D for T , we can compute expectation values for the predictions using $P(\lambda|TD)$. We then find that

$$E[\tilde{T}(\lambda)] = \tilde{T} + \tilde{\lambda}(T, D) \tilde{\beta} = \tilde{T} + \tilde{\beta} \beta^T (C + S)^{-1} (D - T_0), \quad (2.37)$$

so the predictions are shifted by

$$\delta \tilde{T} = -\hat{S}(C + S)^{-1} (T_0 - D), \quad (2.38)$$

where

$$\hat{S} = \tilde{\beta} \beta^T, \quad (2.39)$$

is the cross-covariance matrix between observables T for which we have data D , and the predictions $\tilde{T}$. Likewise

$$\begin{aligned} \text{Cov}[\tilde{T}(\lambda)] &\equiv E[(\tilde{T}(\lambda) - E[\tilde{T}(\lambda)])(\tilde{T}(\lambda) - E[\tilde{T}(\lambda)])^T] \\ &= \text{Var}[\lambda] \tilde{\beta} \tilde{\beta}^T = Z \tilde{S}. \end{aligned} \quad (2.40)$$

so the covariance matrix of the predictions is reduced by the same factor Z as that for the autopredictions. Thus the data D can lead to more precise (and if the data is correct, also more accurate) predictions for observables that are not yet measured, through the correlation of theoretical uncertainties. The reduction in the size of the covariance matrix is through the same factor Z as for the autopredictions, while

the size of the shift is proportional to the cross-covariance between the theoretical uncertainties.

Note that if, having made the predictions $\tilde{T} + \delta \tilde{T}$, an experimentalist made measurements to produce independent data $\tilde{D}$, with experimental covariance matrix $\tilde{C}$, and we wanted to combine the datasets $\{D, \tilde{D}\}$ into a single dataset, then the combined $(N + \tilde{N}) \times (N + \tilde{N})$ theoretical covariance matrix for $\{T, \tilde{T}\}$ to be used to compare to the data would be

$$\begin{pmatrix} S & \hat{S}^T \\ \hat{S} & \tilde{S} \end{pmatrix} = \begin{pmatrix} \beta \beta^T & \beta \tilde{\beta}^T \\ \tilde{\beta} \beta^T & \tilde{\beta} \tilde{\beta}^T \end{pmatrix}. \quad (2.41)$$

While we might hope that the shifted predictions $\tilde{T} + \delta \tilde{T}$ give a better χ^2 to the new data $\tilde{D}$, this is no longer guaranteed, since the shifts are driven by the old data D , and it is possible that $\tilde{D}$ are inconsistent with them.

2.3 Autopredictions in perfect fits

In the previous section we considered the situation of a theory T which was not fitted to the data D . Now consider what is in some sense the opposite situation: a ‘perfect’ fit, where the theoretical predictions T have sufficient flexibility to fit the data D exactly. For a perfect fit, $P(T|D)$ is always maximized when $T = D$, and thus (Eq. (2.2)) $\chi^2 = 0$. We compute expectation values of functions of T using the probability distribution $P(T|D)$ in Eq. (2.1), thus

$$E[T] = \mathcal{N}_T \int dT T P(T|D) = D, \quad (2.42)$$

the normalization $\mathcal{N}_T$ being chosen such that $E[1] = 1$, while

$$\text{Cov}[T] \equiv E[(T - E[T])(T - E[T])^T] = C + S. \quad (2.43)$$

These are the expectation value and covariance matrix of the variables T fitted to the data D .

To compute the autopredictions, we need to consider $T(\lambda) = T + \lambda \beta$ (Eq. (2.21)), and this is more subtle since, as we saw in the previous section, expectation values of λ involve T . We thus need to generalise the definition Eq. (2.19) of expectation values of functions of λ using the probability distribution $P(\lambda|TD)$, to also include the subsequent integration over T : thus we define

$$E[f(T, \lambda)] \equiv \mathcal{N}_T \int dT \left(\mathcal{N}_\lambda \int d\lambda f(T, \lambda) P(\lambda|TD) \right) P(T|D), \quad (2.44)$$

for any function $f(T, \lambda)$ of λ and T . There are several things to note about this procedure:

- We always perform the integration over λ , weighted with the probability distribution $P(\lambda|TD)$, before we perform the integration over T using $P(T|D)$: this is because while $P(\lambda|TD)$ depends on T , $P(T|D)$ does not depend on λ , because it has been marginalised as in Eq. (2.16).

- The data D are always held fixed throughout: both $P(\lambda|TD)$ and $P(T|D)$ are conditional on D .
- For functions $f(T, \lambda)$ which only depend on T , the integration over λ is trivial and we recover for example the results Eqs. (2.42, 2.43).
- For the pure theory discussed in the previous section, the theory T was held fixed, so the T integration was trivial and we recover the results Eqs. (2.19, 2.20).

Thus in the case of a perfect fit, the expectation value of the nuisance parameter is

$$E[\lambda] = E[\bar{\lambda}(T, D)] = \beta^T (C + S)^{-1} E[D - T] = 0, \quad (2.45)$$

using Eq. (2.15) and then finally Eq. (2.42). The calculation of the variance requires some care, to ensure that the average over λ is separated out from the average over T . This is most easily accomplished by adding and subtracting $\bar{\lambda}(T, D)$, so

$$\begin{aligned} \text{Var}[\lambda] &= E[(\lambda - E[\lambda])^2] = E[(\lambda - \bar{\lambda}(T, D) + \bar{\lambda}(T, D))^2] \\ &= E[(\lambda - \bar{\lambda}(T, D))^2] + E[(\bar{\lambda}(T, D))^2] \\ &= Z + \beta^T (C + S)^{-1} E[(T - D)(T - D)^T] (C + S)^{-1} \beta \\ &= Z + \beta^T (C + S)^{-1} \text{Cov}[T] (C + S)^{-1} \beta \\ &= 1 - \beta^T (C + S)^{-1} \beta + \beta^T (C + S)^{-1} \beta = 1. \end{aligned} \quad (2.46)$$

where in the second line we note that the cross term vanishes, while in the last line we used Eq. (2.10) for Z , and Eqs. (2.43) to simplify the second term. We thus find that in a perfect fit, the probability distribution of the nuisance parameters after fitting the data is the same as the prior distribution Eq. (2.7): we learn nothing from the data about the theoretical uncertainty because all the information in the data is absorbed in the fitted parameters. The calculation of the variance is particularly instructive: the reduction by the factor Z found in the pure theory, Eq. (2.20) is now precisely cancelled by the fluctuation of $\bar{\lambda}(T, D)$ due to the covariance Eq. (2.43) of T .

Turning to the autopredictions Eq. (2.21), we have

$$E[T(\lambda)] = E[T + \lambda\beta] = D, \quad (2.48)$$

using Eqs. (2.42, 2.45): as expected in a perfect fit, the autopredictions simply return the original data. Their covariance is more interesting, as we now have to take account of the correlation between fitted theory T and the nuisance parameter λ : proceeding as in Eq. (2.47)

$$\begin{aligned} \text{Cov}[T(\lambda)] &= E[(T(\lambda) - E[T(\lambda)])(T(\lambda) - E[T(\lambda)])^T] \\ &= E[(T - D + \lambda\beta)(T - D + \lambda\beta)^T] \\ &= E[(T - D)(T - D)^T] + E[\lambda\beta(T - D)^T] \\ &\quad + E[(T - D)\lambda\beta^T] + E[\lambda^2]\beta\beta^T. \end{aligned} \quad (2.49)$$

Now the first term is just $\text{Cov}[T]$, and the last is just $\text{Var}[\lambda]S$, while the cross-terms can be evaluated using Eq. (2.15):

$$\begin{aligned} E[\lambda\beta(T - D)^T] &= E[\beta\bar{\lambda}(T, D)(T - D)^T] \\ &= -S(C + S)^{-1} E[(T - D)(T - D)^T] \\ &= -S(C + S)^{-1} \text{Cov}[T]. \end{aligned} \quad (2.50)$$

We thus find that

$$\begin{aligned} \text{Cov}[T(\lambda)] &= \text{Cov}[T] - S(C + S)^{-1} \text{Cov}[T] \\ &\quad - \text{Cov}[T](C + S)^{-1} S + \text{Var}[\lambda]S \end{aligned} \quad (2.51)$$

$$= (C + S) - S - S + S = C, \quad (2.52)$$

where in the last line we used Eq. (2.43) for $\text{Cov}[T]$, and Eq. (2.47) for $\text{Var}[\lambda]$. Thus the covariance of the autopredictions in a perfect fit is simply the covariance of the data. The way in which this arises is that the theory covariance arising from the fit (the first term in Eq.(2.51) and the theory covariance arising in the autoprediction (the last term in Eq.(2.51) are each cancelled by the cross-covariance between fitting and prediction, just as was argued in Ref. [13]. When we have a perfect fit, there is really no distinction between the autoprediction and the data, and the theory uncertainty thus becomes irrelevant. So in a sense this model is pure phenomenology: the only nontrivial information is in the data. Indeed in this model as it stands there is no possibility of making genuine predictions, since a prediction of the form Eq. (2.35) is useless unless we can determine $\tilde{T}$, presumably as functions of T , and for this we need a genuine theory.

3 Correlated theory uncertainties in one parameter fits

In the previous section, we considered two simple but unrealistic models: the first in which the theory T is fixed, with no free parameters to be fitted to the data (pure theory), and the second in which the theory T is so flexible that we could achieve a perfect fit, $T = D$ (pure phenomenology). These exercises were useful, in that they gave us some practice in the use of nuisance parameters to propagate theoretical uncertainties. However we now need to consider the more realistic situation in which the theory has parameters that can be constrained by data, but is still sufficiently restrictive that it can be considered a theory. The fit to the data is then not perfect, but the theory is sufficiently constraining that it can be used to predict new observables, $\tilde{T}$, for which we as yet have no data. We will find that the interesting features of the pure theory and pure phenomenology models (the shifts, the reduction in uncertainties due to Bayesian learning, and the correlations between theory uncertainties in the fitting and theory uncertainties in predictions) are also found in these more realistic theories.

In this section we consider a theory with only one fitted parameter: in Sect. 3.1 we explain how the fitting is per-

formed using replicas, in Sect. 3.2 we consider autopredictions in such a theory, and in Sect. 3.3 we consider general predictions. Generalization to many fitting parameters will be considered in the following section.

3.1 Fitting a theory with a single parameter

We can model this situation by considering theoretical predictions $T(\theta)$ which depend on a single parameter θ , so that $\chi^2(\theta)$ is minimized for some choice of this parameter, $\theta = \theta_0$, with some variance $\text{Var}[\theta]$. Other observables $\tilde{T}(\theta)$ are then predicted to be $\tilde{T}(\theta_0)$, with an associated uncertainty proportional to $\text{Var}[\theta]$. We assume as before that uncertainties are Gaussian, which means that we can linearize $T(\theta)$ about $T(\theta_0) \equiv T_0$:

$$T(\theta) = T_0 + (\theta - \theta_0)\dot{T}_0. \quad (3.1)$$

This model has the advantage that while it captures the essence of the fitting problem, it is sufficiently simple that we can solve it exactly.

In order to determine the uncertainty on θ , we will need to propagate the experimental uncertainties in the data D and the theoretical uncertainties in the predictions $T(\theta)$ into θ . This can be done most easily by generating N_{rep} pseudodata replicas $D^{(r)}$ distributed according to a Gaussian distribution centred on the actual data D , with covariance $C + S$: defining the replica average

$$\langle F(D^{(r)}) \rangle = \lim_{N_{\text{rep}} \rightarrow \infty} \frac{1}{N_{\text{rep}}} \sum_{r=1}^{N_{\text{rep}}} F(D^{(r)}) \quad (3.2)$$

for any function F of the replicas, the replicas are chosen such that

$$\langle D^{(r)} \rangle \equiv D, \quad \langle (D^{(r)} - D)(D^{(r)} - D)^T \rangle = C + S. \quad (3.3)$$

A parameter replica $\theta^{(r)}$ is then fitted to each pseudodata replica $D^{(r)}$ by maximizing $P(T(\theta)|D^{(r)})$ as given by Eq. (2.1), and thus by minimizing

$$\chi_r^2[\theta] = (T(\theta) - D^{(r)})^T (C + S)^{-1} (T(\theta) - D^{(r)}), \quad (3.4)$$

with respect to θ , replica by replica. Using Eq. (3.1), minimization of the quadratic gives

$$\theta^{(r)} - \theta_0 = \frac{\dot{T}_0^T (C + S)^{-1} (D^{(r)} - T_0)}{\dot{T}_0^T (C + S)^{-1} \dot{T}_0}. \quad (3.5)$$

Using the replica averages Eq. (3.3), and choosing $\theta_0 = \langle \theta^{(r)} \rangle$, we find for consistency

$$\dot{T}_0^T (C + S)^{-1} (D - T_0) = 0, \quad (3.6)$$

and thus we can rewrite Eq. (3.5) as

$$\theta^{(r)} - \theta_0 = \frac{\dot{T}_0^T (C + S)^{-1} (D^{(r)} - D)}{\dot{T}_0^T (C + S)^{-1} \dot{T}_0}. \quad (3.7)$$

Then since C and S are symmetric matrices,

$$\begin{aligned} \text{Var}[\theta] &= \langle (\theta^{(r)} - \theta_0)^2 \rangle \\ &= \frac{\dot{T}_0^T (C + S)^{-1} \langle (D^{(r)} - D)(D^{(r)} - D)^T \rangle (C + S)^{-1} \dot{T}_0}{(\dot{T}_0^T (C + S)^{-1} \dot{T}_0)^2} \\ &= (\dot{T}_0^T (C + S)^{-1} \dot{T}_0)^{-1}. \end{aligned} \quad (3.8)$$

Note the way the double reciprocation in this expression works: data points with a relatively large dependence on θ (i.e. large $\dot{T}_0$) contribute more than those with small dependence, however directions with large uncertainty (i.e. projections of $C + S$) contribute less than those with small uncertainty.

Now that we understand the uncertainty of the fitted parameter θ , we can use it to predict the uncertainties of $T(\theta)$:

$$E[T] \equiv \langle T(\theta^{(r)}) \rangle = T(\theta_0) = T_0, \quad (3.9)$$

so, writing $T^{(r)} = T(\theta^{(r)})$,

$$X \equiv \text{Cov}[T(\theta)] = \langle (T^{(r)} - T_0)(T^{(r)} - T_0)^T \rangle \quad (3.10)$$

$$= \dot{T}_0 \langle (\theta^{(r)} - \theta_0)^2 \rangle \dot{T}_0^T$$

$$= \dot{T}_0 (\dot{T}_0^T (C + S)^{-1} \dot{T}_0)^{-1} \dot{T}_0^T \quad (3.11)$$

$$= n(n^T (C + S)^{-1} n)^{-1} n^T, \quad (3.12)$$

where n is a unit vector in the direction of $\dot{T}_0$, $n^T n = 1$. This shows that X depends only on n , and not on $|\dot{T}_0|$.

The singular matrix X will play an important role in what follows: it is the covariance matrix of T due to the experimental and theoretical uncertainties in the fitting of the parameter θ – the ‘fitting uncertainty’. When the fitted parameter minimizes the χ^2 , and is thus given by Eq. (3.7), X satisfies the projective relation

$$X = X(C + S)^{-1} X. \quad (3.13)$$

Using Eq. (3.7) in Eq. (3.1), we see that $X(C + S)^{-1}$ projects the data replicas onto the theory replicas:

$$T^{(r)} - T_0 = X(C + S)^{-1} (D^{(r)} - D). \quad (3.14)$$

Because this relation is projective, some information is lost whenever we perform the fit: Eq. (3.14) cannot be inverted to obtain data replicas from theory replicas. This is an inevitable consequence of describing N data (assuming $N > 1$) with only a single parameter θ .

To make X more explicit, consider the simple model Eq. (2.29) for C and S . Then using Eq. (2.30),

$$n^T (C + S)^{-1} n = \frac{\sigma^2 + s^2 \sin^2 \phi}{\sigma^2 (\sigma^2 + s^2)}, \quad (3.15)$$

where $\cos \phi = n^T e$, and thus if we project X onto n (projections orthogonal to n give zero)

$$n^T X n = \frac{\sigma^2 (\sigma^2 + s^2)}{(\sigma^2 + s^2 \sin^2 \phi)}. \quad (3.16)$$

The contribution of the theory uncertainty s thus depends on how well aligned e is to the direction n of the parameter dependence: if $\phi = 0$ we have complete alignment, and the variance of T in this direction is $\sigma^2 + s^2$ as expected, while if $\phi = \frac{\pi}{2}$ we have orthogonality, and the variance of T is σ^2 – the theory uncertainty is then irrelevant to the fitting.

3.2 Autopredictions in single parameter Fits

We can now consider the evaluation of the mean and covariance of the ‘autopredictions’

$$T(\theta, \lambda) = T(\theta) + \lambda\beta \quad (3.17)$$

in this one parameter model. Just as in Sect. 2.3, we do this by first computing expectation values over λ , using $P(\lambda|TD)$, which depend on T , and then evaluate the expectation values over T , according to the probability distribution $P(T|D)$, now performed by averaging over theory replicas $T^{(r)} = T(\theta^{(r)})$. It is important to note that both these averages are performed holding the data D fixed, as both probabilities are conditional on the data: the data replicas $D^{(r)}$ employed in Sect. 3.1 are only a device to generate the theory replicas $T^{(r)}$, and are not to be averaged over when determining expectation values. Accordingly, Eq. (3.18) now becomes

$$E[f(T, \lambda)] = \left\langle \left(\mathcal{N}_\lambda \int d\lambda f(T^{(r)}, \lambda) P(\lambda|T^{(r)}D) \right) \right\rangle, \quad (3.18)$$

where the angled brackets denote the replica average Eq. (3.2).

Following the same steps as in the perfect fit in Sect. 2.3, but now using the theory replicas $T^{(r)} = T(\theta^{(r)})$ determined in the one parameter fit Sect. 3.1, we find

$$E[\lambda] \equiv \langle \bar{\lambda}(T(\theta^{(r)}), D) \rangle = \beta^T (C + S)^{-1} (D - T_0) \equiv \bar{\lambda}_0, \quad (3.19)$$

Unlike in the perfect fit Eq. (2.45), but just as in the pure theory Eq. (2.19) the nuisance parameters can now have nonzero expectation values, since the one parameter fit no longer fits the data exactly. These in turn give nontrivial shifts in the theoretical predictions:

$$\begin{aligned} E[T(\theta, \lambda)] &= \langle T^{(r)} + \bar{\lambda}(T^{(r)}, D)\beta \rangle \\ &= T_0 + \bar{\lambda}_0\beta = T_0 + \beta\beta^T (C + S)^{-1} (D - T_0). \end{aligned} \quad (3.20)$$

So again the data give us information, inducing shifts in the autopredictions:

$$\delta T = -S(C + S)^{-1} (T_0 - D). \quad (3.21)$$

Note however that since (from Eq. (3.6)) $n^T (C + S)^{-1} (T_0 - D) = 0$, these shifts will only be nonzero when n and e (the data and the theory) point in different directions: when they are parallel ($\phi = 0$), the theoretical uncertainty is simply

absorbed by the fit, just as it was in the perfect fit in Sect. 2.3. We can use the same argument as in Sect. 2.2, Eq. (2.27), to show that the shifts will always improve the fit to the experimental data.

For the uncertainties, consider first the variance of λ : following the same argument that led to Eq. (2.47) in Sect. 2.3, we now find

$$\begin{aligned} \text{Var}[\lambda] &= E[(\lambda - E[\lambda])^2] = E[(\lambda - \bar{\lambda}(T, D) \\ &\quad + \bar{\lambda}(T, D) - \bar{\lambda}_0)^2] \\ &= E[(\lambda - \bar{\lambda}(T, D))^2] + \langle (\bar{\lambda}(T^{(r)}, D) - \bar{\lambda}_0)^2 \rangle \\ &= Z + \beta^T (C + S)^{-1} \langle (T^{(r)} - T_0)(T^{(r)} - T_0)^T \rangle \\ &\quad \times (C + S)^{-1} \beta \\ &= 1 - \beta^T (C + S)^{-1} \beta \\ &\quad + \beta^T (C + S)^{-1} X (C + S)^{-1} \beta \equiv \bar{Z}. \end{aligned} \quad (3.22)$$

where in the second line we turned the expectation value over T into a replica average, and in the last line we used Eq. (2.10) for Z , and Eq. (3.10) for $\text{Cov}[T]$. We thus find that in the more restrictive environment of the one parameter fit, the last two terms no longer cancel: the information in the data can no longer be entirely absorbed in the single fitted parameter, and so it can still inform the nuisance parameter.

It is easy to see that $\bar{Z} \geq Z$ because $(C + S)^{-1} X (C + S)^{-1}$ is positive semi-definite, while $\bar{Z} \leq 1$ since $X (C + S)^{-1}$ is projective, Eq. (3.13), so its eigenvalues are either zero or one. So in place of Eq. (2.18) we now have

$$0 < Z \leq \bar{Z} \leq 1. \quad (3.23)$$

The information on theoretical uncertainties extracted from the data is thus less in the one parameter fit than it was in the pure theory of Sect. 2.2, due to the extra uncertainty arising in the fit itself, but unlike in the perfect fit Sect. 2.3, the data will still constrain the theoretical uncertainties provided the parameter and theoretical uncertainty act in different directions.

In the model Eq. (2.29) for C and S , and using Eq. (3.16) for X , Eq. (3.22) becomes

$$\bar{Z} = \frac{\sigma^2}{\sigma^2 + s^2 \sin^2 \phi}. \quad (3.24)$$

Comparing with the corresponding expression for Z , Eq. (2.32), we see that indeed $\bar{Z} = 1$ when $\phi = 0$, thus when $n = e$ and the parameter variation and theoretical uncertainty are aligned, while $\bar{Z} = Z$ only if $\phi = \pi/2$, so when n and e are orthogonal, the data have the greatest influence on the uncertainty.

For the covariance of the autopredictions Eq. (3.17) we also have to take account of the correlation between the fitted theory and the nuisance parameter: proceeding as in Eq. (2.49)

$$\begin{aligned}
\text{Cov}[T(\theta, \lambda)] &= E[(T(\theta, \lambda) - E[T(\theta, \lambda)])(T(\theta, \lambda) \\
&\quad - E[T(\theta, \lambda)])^T] \\
&= E[(T - T_0 + (\lambda - \bar{\lambda}_0)\beta) \\
&\quad (T - T_0 + (\lambda - \bar{\lambda}_0)\beta)^T] \\
&= \langle (T^{(r)} - T_0)(T^{(r)} - T_0)^T \rangle \\
&\quad + E[(\lambda - \bar{\lambda}_0)\beta(T - T_0)^T] \\
&\quad + E[(T - T_0)(\lambda - \bar{\lambda}_0)\beta^T] \\
&\quad + E[(\lambda - \bar{\lambda}_0)^2]\beta\beta^T. \quad (3.25)
\end{aligned}$$

Then again the first term is $\text{Cov}[T] = X$, Eq. (3.10), while the last is $\text{Var}[\lambda]S$, Eq. (3.22), while the cross-terms can be evaluated using Eq. (2.15):

$$\begin{aligned}
E[(\lambda - \bar{\lambda}_0)\beta(T - T_0)^T] &= \langle \beta(\bar{\lambda}(T^{(r)}, D) \\
&\quad - \bar{\lambda}_0, D)(T^{(r)} - T_0)^T \rangle \\
&= -S(C + S)^{-1} \langle (T^{(r)} - T_0) \\
&\quad (T^{(r)} - T_0)^T \rangle \\
&= -S(C + S)^{-1} \text{Cov}[T]. \quad (3.26)
\end{aligned}$$

We thus find that the result is simply

$$\text{Cov}[T(\lambda)] = X - S(C + S)^{-1}X - X(C + S)^{-1}S + \bar{Z}S. \quad (3.27)$$

The meaning of the four terms is easy to understand: the first is the ‘fitting uncertainty’ (which includes contributions from both experimental and theoretical uncertainties), the last the ‘theory uncertainty’, reduced through exposure to the data, and the middle two terms are due to the correlations between the two sources of theoretical uncertainty. We can simplify it by using Eq. (3.22) to show that

$$\bar{Z}S = S(C + S)^{-1}X(C + S)^{-1}S + ZS, \quad (3.28)$$

and then some straightforward algebra to write

$$\begin{aligned}
&X - S(C + S)^{-1}X - X(C + S)^{-1}S \\
&\quad + S(C + S)^{-1}X(C + S)^{-1}S \\
&= C(C + S)^{-1}X(C + S)^{-1}C. \quad (3.29)
\end{aligned}$$

We thus find finally

$$\text{Cov}[T(\lambda)] = C(C + S)^{-1}X(C + S)^{-1}C + ZS. \quad (3.30)$$

Note that if we write $X = C + S$ (as in the perfect fit model in Sect. 2.3), this result reduces to C , as it should. The cancellations noted in Eq. (2.52) between the cross-terms and the covariances of T and λ are no longer exact in Eq. (3.30), because $\text{Cov}[T]$ is no longer $C + S$, but rather the smaller matrix X (which is in a sense $C + S$ restricted to the space of variation of the parameter θ as in Eq. (3.10)). Thus the result is no longer the experimental covariance matrix C , but rather the sum in quadrature of the ‘fitting uncertainty’, X , and the ‘theory uncertainty’ S , each reduced to some extent

by the correlation of the theoretical uncertainties in fit and prediction.

In the model Eq. (2.29) for C and S , and using Eq. (3.16) for X , we now find

$$\begin{aligned}
\text{Cov}[T(\lambda)] &= \frac{\sigma^2(\sigma^2 + s^2)}{(\sigma^2 + s^2 \sin^2 \phi)} \left(nn^T - \frac{s^2}{\sigma^2 + s^2} \right. \\
&\quad \times \cos \phi (en^T + ne^T) + \frac{s^2}{\sigma^2 + s^2} ee^T \left. \right). \quad (3.31)
\end{aligned}$$

The first term is just X , the off-diagonal term in the middle is the correlation term, and the last is $\bar{Z}S$. If $\phi = 0$, so $n = e$, the three terms combine to give simply $\sigma^2 nn^T$: in the direction of the fitted parameter, the uncertainty in the autoprediction is the experimental uncertainty, just as in the perfect fit described in Sect. 2.3. On the other hand, if $\phi = \pi/2$, so n and e are orthogonal, the correlation term disappears, and the result reduces to $X + ZS$: we add the uncertainties in quadrature, since they are in orthogonal directions, and the theoretical uncertainty is reduced just as in the pure theory described in Sect. 2.2. The one parameter fit thus interpolates smoothly between these two extremes. Note that we can write Eq. (3.31) as

$$\begin{aligned}
\text{Cov}[T(\lambda)] &= \frac{\sigma^2(\sigma^2 + s^2)}{(\sigma^2 + s^2 \sin^2 \phi)} \left(n - \frac{s^2 \cos \phi}{\sigma^2 + s^2} e \right) \\
&\quad \times \left(n^T - \frac{s^2 \cos \phi}{\sigma^2 + s^2} e^T \right) \\
&\quad + \frac{s^2 \sigma^2}{\sigma^2 + s^2} ee^T. \quad (3.32)
\end{aligned}$$

Here the last term is just ZS , while the first is X , but with the vector n given an additional component in the direction e due to the theory correlation: besides changing its direction, this reduces the size of the fitting uncertainty by a factor $\sqrt{\sin^2 \phi + Z^2 \cos^2 \phi}$. However it is easy to see that the size of the correlated fitting uncertainty is still larger than it would be if the theory uncertainty had not been included in the fit.

3.3 Correlated predictions in one parameter fits

We now consider predictions $\tilde{T}_I(\theta)$, $I = 1, \dots, \tilde{N}$, which depend on the same parameter θ as the fitted predictions $T_i(\theta)$, $i = 1, \dots, N$. There are two distinct sources of uncertainty in $\tilde{T}_I(\theta)$: uncertainties in the determination of θ due to the experimental uncertainties in the data D_i and theoretical uncertainties in the theory $T_i(\theta)$ used in its determination; and theoretical uncertainties in the predictions $\tilde{T}_I(\theta)$ themselves.

The first uncertainty is expressed through Eq. (3.7), which gives the variance Eq. (3.8). In analogy with Eq. (3.1) the

linearised dependence of the predictions $\tilde{T}(\theta)$ may be written

$$\tilde{T}(\theta) = \tilde{T}_0 + (\theta - \theta_0)\dot{\tilde{T}}_0, \quad (3.33)$$

with $\tilde{T}_0 = \tilde{T}(\theta_0)$. Then the covariance of $\tilde{T}_I(\theta)$ due to the uncertainty in the parameter θ is derived just as in Eqs. (3.10–3.12) for the autopredictions: writing $\tilde{T}^{(r)} \equiv \tilde{T}(\theta^{(r)})$

$$\tilde{X} \equiv \text{Cov}[\tilde{T}(\theta)] = \langle (\tilde{T}^{(r)} - \tilde{T}_0)(\tilde{T}^{(r)} - \tilde{T}_0)^T \rangle \quad (3.34)$$

$$\begin{aligned} &= \dot{\tilde{T}}_0 \langle (\theta^{(r)} - \theta_0)^2 \rangle \dot{\tilde{T}}_0^T \\ &= \dot{\tilde{T}}_0 (\dot{T}_0^T (C + S)^{-1} \dot{T}_0)^{-1} \dot{\tilde{T}}_0^T. \end{aligned} \quad (3.35)$$

The second uncertainty – the theoretical uncertainty in the predictions $\tilde{T}_I(\theta)$ – may again be either correlated or uncorrelated with the theoretical uncertainty in $T_i(\theta)$. Consider first the simpler situation when it is uncorrelated: this might be the case if, for example, the observable $\tilde{T}(\theta)$ was a different type of observable to the $T(\theta)$ used to determine θ . Then introducing a nuisance parameter $\tilde{\lambda}$, Gaussian distributed about zero with unit variance, and uncorrelated with λ , we can write (as in the pure theory model Eq. (2.35))

$$\tilde{T}(\theta, \tilde{\lambda}) = \tilde{T}(\theta) + \tilde{\lambda}\tilde{\beta}, \quad (3.36)$$

where the vector $\tilde{\beta}_I$ gives the size of the theoretical uncertainties in $\tilde{T}_I(\theta)$. Since θ and $\tilde{\lambda}$ are uncorrelated, we then have

$$E[\tilde{T}(\theta, \tilde{\lambda})] = \tilde{T}(\theta_0), \quad (3.37)$$

$$\text{Cov}[\tilde{T}(\theta, \tilde{\lambda})] = \text{Cov}[\tilde{T}(\theta)] + \text{Var}[\tilde{\lambda}]\tilde{\beta}\tilde{\beta}^T = \tilde{X} + \tilde{S}, \quad (3.38)$$

where $\tilde{S} = \tilde{\beta}\tilde{\beta}^T$ is the theory covariance matrix for the prediction $\tilde{T}(\theta)$ (compare Eq. (2.36) in the pure theory). Thus when the theoretical uncertainty is uncorrelated we simply add it in quadrature to the uncertainty due to that in the parameter θ derived from the fit.

Now consider the more interesting case in which the theoretical uncertainty in $\tilde{T}_I(\theta)$ is fully correlated to that in the $T_i(\theta)$ used in the fit to determine θ : then $\tilde{\lambda} = \lambda$, which has already been determined in the fit to have nonzero expectation value and variance Eqs. (3.19, 3.22). Then writing $\tilde{T}(\theta, \lambda) = \tilde{T}(\theta) + \lambda\tilde{\beta}$,

$$E[\tilde{T}(\theta, \lambda)] = \tilde{T}_0 + \bar{\lambda}(T_0, D)\tilde{\beta}, \quad (3.39)$$

so the correlation induces a similar shift in the predictions to that in the autopredictions Eq. (3.21): using Eq. (2.15)

$$\delta\tilde{T}(\theta_0) = \tilde{\beta}\beta^T(C + S)^{-1}(D - T_0) = -\hat{S}(C + S)^{-1}(T_0 - D). \quad (3.40)$$

where $\hat{S} = \tilde{\beta}\beta^T$, Eq. (2.39), is the matrix of cross-correlations between observables $T(\theta)$ used in the fit and the predictions $\tilde{T}(\theta)$.

Likewise using the same arguments as were used to derive $\text{Cov}[T(\theta, \lambda)]$, Eq. (3.27)

$$\text{Cov}[\tilde{T}(\theta, \lambda)] = E[(\tilde{T}(\theta, \lambda) - E[\tilde{T}(\theta, \lambda)])(\tilde{T}(\theta, \lambda) - E[\tilde{T}(\theta, \lambda)])^T]$$

$$\begin{aligned} &-E[\tilde{T}(\theta, \lambda)]^T] \\ &= \langle (\tilde{T}^{(r)} - \tilde{T}_0)(\tilde{T}^{(r)} - \tilde{T}_0)^T \rangle \\ &\quad + E[(\lambda - \bar{\lambda}_0)\tilde{\beta}(\tilde{T} - \tilde{T}_0)^T] \\ &\quad + E[(\tilde{T} - \tilde{T}_0)(\lambda - \bar{\lambda}_0)\tilde{\beta}^T] \\ &\quad + E[(\lambda - \bar{\lambda}_0)^2]\tilde{\beta}\tilde{\beta}^T. \end{aligned} \quad (3.41)$$

Then again the first term is $\text{Cov}[\tilde{T}] = \tilde{X}$, Eq. (3.34), while the last is $\text{Var}[\lambda]\tilde{S}$, Eqs. (3.22, 2.36), while the cross-terms can be evaluated using Eq. (2.15):

$$\begin{aligned} E[(\lambda - \bar{\lambda}_0)\tilde{\beta}(\tilde{T} - \tilde{T}_0)^T] &= \langle \tilde{\beta}(\bar{\lambda}(T^{(r)}, D) \\ &\quad - \bar{\lambda}(T_0, D))(\tilde{T}^{(r)} - \tilde{T}_0)^T \rangle \\ &= -\hat{S}(C + S)^{-1} \langle (T^{(r)} - T_0) \\ &\quad (\tilde{T}^{(r)} - \tilde{T}_0)^T \rangle \\ &= -\hat{S}(C + S)^{-1}\hat{X}^T, \end{aligned} \quad (3.42)$$

where $\hat{S}$ is given by Eq. (2.39), and in analogy to Eq. (3.10) and Eq. (3.34) we define the cross-covariance between $\tilde{T}$ and T

$$\hat{X} \equiv \langle (\tilde{T}^{(r)} - \tilde{T}_0)(T^{(r)} - T_0)^T \rangle \quad (3.43)$$

$$\begin{aligned} &= \dot{\tilde{T}}_0 \langle (\theta^{(r)} - \theta_0)^2 \rangle \dot{T}_0^T \\ &= \dot{\tilde{T}}_0 (\dot{T}_0^T (C + S)^{-1} \dot{T}_0)^{-1} \dot{T}_0^T. \end{aligned} \quad (3.44)$$

We thus find that

$$\begin{aligned} \text{Cov}[\tilde{T}(\theta, \lambda)] &= \tilde{X} - \hat{S}(C + S)^{-1}\hat{X}^T \\ &\quad - \hat{X}(C + S)^{-1}\hat{S}^T + \bar{Z}\bar{S}. \end{aligned} \quad (3.45)$$

Using Eq. (3.22), we can write the last term as

$$\bar{Z}\bar{S} = Z\bar{S} + \hat{S}(C + S)^{-1}X(C + S)^{-1}\hat{S}^T, \quad (3.46)$$

$$Z\bar{S} = \bar{S} - \hat{S}(C + S)^{-1}\hat{S}^T. \quad (3.47)$$

Note that since the coefficients Z and $\bar{Z}$ are the same as for the autopredictions, and thus satisfy the bounds Eq. (3.23), we must have (for positive definite C and positive semi-definite S , i.e. $C > 0$, $S \geq 0$)

$$0 \leq \hat{S}(C + S)^{-1}X(C + S)^{-1}\hat{S}^T \leq \hat{S}(C + S)^{-1}\hat{S}^T \leq \bar{S}, \quad (3.48)$$

so in particular the subtraction (the last term in Eq. (3.47)) can never be so large that it makes the entire covariance matrix negative.

In summary, comparing Eqs. (3.40, 3.45) with Eqs. (3.37, 3.38), we see that including the correlations between the theoretical uncertainties in the fit and the prediction results in three effects: a shift in the central value of the prediction, a reduction in the theoretical uncertainty, and a reduction in the fitting uncertainty due to the correlations. Performing the fit gives us information (from the data) about the theory, which results in more precise, and hopefully more accurate, predictions.

4 Correlated MHOU in PDF fits

We now repeat the above analysis, but instead of the toy model we consider the more realistic situation in which the theoretical expressions $T_i[f]$ depend on PDFs f , determined in a global fit to N data D_i , with experimental covariance matrix C_{ij} , and then used to make $\tilde{N}$ predictions $\tilde{T}_I[f]$. There are then many sources of theoretical uncertainty in the relation between the theoretical calculations and the PDFs: here we consider the most generic, the missing higher order uncertainty (MHOU), computed using scale variations according to one of the prescriptions set out in Ref. [8,9]. The theory covariance matrices S_{ij} and $\tilde{S}_{IJ}$ associated with the MHOU will then have many non-zero eigenvalues, and thus there will be n nuisance parameters λ_α , $\alpha = 1, \dots, n$ to take into account. There is in principle no limit on n , though in practice $n \ll N$. As in the toy model, the fitting of the PDFs to the data will determine the mean and covariance of the nuisance parameters, which will then translate into systematic shifts and changes in the uncertainties of the theoretical predictions.

4.1 Expectation and covariance of multiple nuisance parameters

The nuisance parameters λ_α correspond to shifts in the theoretical predictions: $T_i[f] \rightarrow T_i[f] + \lambda_\alpha \beta_{i,\alpha}[f]$, where we adopt the summation convention for the index α . The shift vectors $\beta_{i,\alpha}$ are not necessarily orthogonal to each other. We again assume Gaussian uncertainties, so that in place of Eq. (2.4) we now have

$$P(T|D\lambda) \propto \exp\left(-\frac{1}{2}(T[f] + \lambda_\alpha \beta_\alpha - D)^T C^{-1} \times (T[f] + \lambda_\alpha \beta_\alpha - D)\right), \quad (4.1)$$

and assume that each nuisance parameter has a prior which is Gaussian distributed with unit variance, centred on zero, the distributions being independent both of each other and of the data, so that

$$P(\lambda|D) = P(\lambda) \propto \exp\left(-\frac{1}{2}\lambda_\alpha \lambda_\alpha\right). \quad (4.2)$$

We now marginalize over λ_α , as in Eq. (2.6)

$$P(T|D) \propto \int d^n \lambda \exp\left(-\frac{1}{2}[(T[f] + \lambda_\alpha \beta_\alpha - D)^T C^{-1} \times (T[f] + \lambda_\alpha \beta_\alpha - D) + \delta_{\alpha\beta} \lambda_\alpha \lambda_\beta]\right), \quad (4.3)$$

by first completing the square: the details are messy, but the result is very similar to Eq. (2.16), namely

$$P(T|D) \propto \int d^n \lambda \exp\left(-\frac{1}{2}(\lambda_\alpha - \bar{\lambda}_\alpha) Z_{\alpha\beta}^{-1} (\lambda_\beta - \bar{\lambda}_\beta) - \frac{1}{2}\chi^2\right) \propto \exp\left(-\frac{1}{2}\chi^2\right), \quad (4.4)$$

where now

$$Z_{\alpha\beta} = (\delta_{\alpha\beta} + \beta_\alpha^T C^{-1} \beta_\beta)^{-1}, \quad (4.5)$$

the inverse on the right hand side being the matrix inverse with respect to the indices α and β ,

$$\bar{\lambda}_\alpha(T, D) = Z_{\alpha\beta} \beta_\beta^T C^{-1} (D - T), \quad (4.6)$$

and χ^2 is once again given by Eq. (2.2), but now with in place of Eq. (2.3)

$$S = \beta_\alpha \beta_\alpha^T, \quad (4.7)$$

as expected. The Gaussian integration in Eq. (4.4) is now trivial, taking us back again to Eq. (2.1) up to a factor $(2\pi)^{n/2}(\det Z)^{1/2}$, which we can ignore as it does not depend on T or D , while Bayes' Theorem Eq. (2.5) gives us the posterior distribution of the nuisance parameters:

$$P(\lambda|TD) \propto \exp\left(-\frac{1}{2}(\lambda_\alpha - \bar{\lambda}_\alpha) Z_{\alpha\beta}^{-1} (\lambda_\beta - \bar{\lambda}_\beta)\right), \quad (4.8)$$

whence we see that

$$E[\lambda_\alpha] = \bar{\lambda}_\alpha, \quad E[(\lambda_\alpha - \bar{\lambda}_\alpha)(\lambda_\beta - \bar{\lambda}_\beta)] = Z_{\alpha\beta}. \quad (4.9)$$

It is easy to see from the definition Eq. (4.5) that if e_α is a unit eigenvector of $Z_{\alpha\beta}$, and $\beta = e_\alpha \beta_\alpha$, then the corresponding eigenvalue of $Z_{\alpha\beta}$ is $z = (1 + \beta^T C^{-1} \beta)^{-1}$, so $0 < z < 1$, and (in analogy to the bounds Eq. (2.18)) $Z_{\alpha\beta}$ is positive definite (thus invertible) and $\delta_{\alpha\beta} - Z_{\alpha\beta}$ is also positive definite (because the eigenvalues z are all less than one). We can summarise this by writing, in place of Eq. (2.18)

$$0 < Z_{\alpha\beta} < \delta_{\alpha\beta}. \quad (4.10)$$

We can express $Z_{\alpha\beta}$ in terms of the inverse of $C + S$:

$$Z_{\alpha\beta} = \delta_{\alpha\beta} - \beta_\alpha^T (C + S)^{-1} \beta_\beta, \quad (4.11)$$

since

$$\begin{aligned} & (\delta_{\alpha\gamma} + \beta_\alpha^T C^{-1} \beta_\gamma)(\delta_{\gamma\beta} - \beta_\gamma^T (C + S)^{-1} \beta_\beta) \\ &= \delta_{\alpha\beta} + \beta_\alpha^T (C^{-1} - (C + S)^{-1} - C^{-1} S (C + S)^{-1}) \beta_\beta \\ &= \delta_{\alpha\beta}. \end{aligned} \quad (4.12)$$

Combining Eq. (4.11) with Eq. (4.6), we then have

$$\bar{\lambda}_\alpha = \beta_\alpha^T (C + S)^{-1} (D - T), \quad (4.13)$$

since $(1 - (C + S)^{-1} S) C^{-1} = (C + S)^{-1}$.

4.2 Fitting the PDFs

We now proceed to apply the above results in the context of a PDF fit incorporating MHOu [8,9]. We consider first a fixed parametrization: the PDFs $f(\theta)$ will then depend on m parameters θ_p , $p = 1, \dots, m$, with $m < N$ so the data D are sufficient to determine all the parameters through minimization of the χ^2 Eq. (2.2).

We can then follow the same procedure as in Sect. 3, the only difference being that now we fit the m parameters θ_p rather than just the single parameter θ . Writing the theoretical predictions $T[f(\theta)] \equiv T(\theta)$, the linearization relation Eq. (3.33) becomes

$$T(\theta) = T_0 + (\theta_p - \theta_p^0)T_p, \quad (4.14)$$

where $f(\theta^0) \equiv f_0$ is the PDF that minimizes the χ^2 , $T_0 \equiv T(\theta^0)$, $T_p \equiv \partial T(\theta^0)/\partial \theta_p^0$, and we use the summation convention for indices p . Using the data replicas Eq. (3.2, 3.3), minimizing Eq. (3.4) with respect to θ_p rather than θ , we find in place of Eq. (3.7) that the fluctuations of the PDF replica parameters are given by

$$\theta_p^{(r)} - \theta_p^0 = (T_p^T (C + S)^{-1} T_q)^{-1} T_q^T (C + S)^{-1} (D^{(r)} - D), \quad (4.15)$$

where the matrix inverse in the first factor on the right hand side is with respect to the p, q indices. It follows that instead of Eq. (3.8) we now have the covariance matrix

$$\begin{aligned} \text{Cov}_{pq}[\theta] &= \langle (\theta_p^{(r)} - \theta_p^0)(\theta_q^{(r)} - \theta_q^0) \rangle \\ &= (T_p^T (C + S)^{-1} T_q)^{-1}, \end{aligned} \quad (4.16)$$

while the expression Eq. (3.12) for the covariance of the predictions $T[f]$ becomes, on writing $T_p = |T_p|n_p$, where n_p are unit vectors (which are however not necessarily orthogonal)

$$X = n_p (n_p^T (C + S)^{-1} n_q)^{-1} n_q^T. \quad (4.17)$$

so the projective relation Eq. (3.13) still holds, and $X(C + S)^{-1}$ projects data replicas onto theory replicas as in Eq. (3.14).

It is now easy to see that the results for the autopredictions in Sect. 3.2 continue to hold, and that in particular that since the central values of the nuisance parameters $\bar{\lambda}_\alpha$ are given by

$$E[\lambda_\alpha] = -\beta_\alpha^T (C + S)^{-1} (\langle T^{(r)} \rangle - D), \quad (4.18)$$

the shifts Eq. (3.21) are now

$$\begin{aligned} \delta T[f] &= \beta_\alpha \beta_\alpha^T (C + S)^{-1} (D - T[f_0]) \\ &= -S(C + S)^{-1} (T[f_0] - D). \end{aligned} \quad (4.19)$$

These shifts will improve the χ^2 to the experimental data, in just the same way as in Sect. 2.2.

Likewise for the uncertainties: Eq. (3.22) for the variance of the nuisance parameter becomes an equation for the covariance matrix of the nuisance parameters in the context of the PDF fit,

$$\begin{aligned} \text{Cov}_{\alpha\beta}[\lambda] &\equiv E[(\lambda_\alpha - E[\lambda_\alpha])(\lambda_\beta - E[\lambda_\beta])] \\ &= \delta_{\alpha\beta} - \beta_\alpha^T (C + S)^{-1} \beta_\beta - \beta_\alpha^T (C + S)^{-1} X (C + S)^{-1} \beta_\beta \\ &\equiv \bar{Z}_{\alpha\beta}, \end{aligned} \quad (4.20)$$

using the projective relation Eq. (3.13). Again, both $\bar{Z}_{\alpha\beta} - Z_{\alpha\beta}$ and $\delta_{\alpha\beta} - \bar{Z}_{\alpha\beta}$ are positive semi-definite, so

$$0 < Z_{\alpha\beta} \leq \bar{Z}_{\alpha\beta} \leq \delta_{\alpha\beta}. \quad (4.21)$$

The covariance matrix of the theoretical autopredictions $T(f, \lambda) \equiv T[f] + \lambda_\alpha \beta_\alpha$, Eq. (3.27, 3.30) then become

$$\begin{aligned} \text{Cov}[T(f, \lambda)] &= X - S(C + S)^{-1} X - X(C + S)^{-1} S \\ &\quad + \beta_\alpha \bar{Z}_{\alpha\beta} \beta_\beta^T \end{aligned} \quad (4.22)$$

$$\begin{aligned} &= C(C + S)^{-1} X (C + S)^{-1} C \\ &\quad + S - S(C + S)^{-1} S, \end{aligned} \quad (4.23)$$

the second expression being identical to the one we found in Sect. 3.2.

The same holds true of course for correlated predictions: the shifts Eq. (3.40) are now

$$\begin{aligned} \delta \tilde{T}[f] &= \tilde{\beta}_\alpha \beta_\alpha^T (C + S)^{-1} (D - T[f_0]) \\ &= \hat{S}(C + S)^{-1} (D - T[f_0]), \end{aligned} \quad (4.24)$$

where $\hat{S} = \tilde{\beta}_\alpha \beta_\alpha^T$, while if $\tilde{T}(f, \lambda) = \tilde{T}[f] + \lambda_\alpha \tilde{\beta}_\alpha$, Eq. (3.45) becomes

$$\begin{aligned} \text{Cov}[\tilde{T}(f, \lambda)] &= \tilde{X} - \hat{S}(C + S)^{-1} \hat{X}^T - \hat{X}(C + S)^{-1} \hat{S}^T \\ &\quad + \tilde{\beta}_\alpha \bar{Z}_{\alpha\beta} \tilde{\beta}_\beta^T, \end{aligned} \quad (4.25)$$

where $\tilde{S} = \tilde{\beta}_\alpha \tilde{\beta}_\alpha^T$, and

$$\tilde{X} = \tilde{T}_p (T_p^T (C + S)^{-1} T_q)^{-1} \tilde{T}_q^T, \quad (4.26)$$

$$\hat{X} = \tilde{T}_p (T_p^T (C + S)^{-1} T_q)^{-1} T_q^T. \quad (4.27)$$

Using Eq. (4.20), we can write the last term as

$$\tilde{\beta}_\alpha \bar{Z}_{\alpha\beta} \tilde{\beta}_\beta^T = \tilde{\beta}_\alpha Z_{\alpha\beta} \tilde{\beta}_\beta^T + \hat{S}(C + S)^{-1} X (C + S)^{-1} \hat{S}^T, \quad (4.28)$$

$$\tilde{\beta}_\alpha Z_{\alpha\beta} \tilde{\beta}_\beta^T = \tilde{S} - \hat{S}(C + S)^{-1} \hat{S}^T. \quad (4.29)$$

The final result thus again has exactly the same form as that found in Sect. 3.3: once the nuisance parameters are all eliminated, the only changes are in the expressions for the covariances X , $\tilde{X}$ and $\hat{X}$, generalizing the previous one parameter expressions to many parameters.

4.3 Fitting NNPDFs

PDFs in an NNPDF fit are parametrized by a neural network, with a very large number of parameters. The fitting procedure differs from that using a fixed parametrization, since we want to avoid fitting noise. In practice this is achieved using a cross-validation procedure. It follows that when we fit to each data replica $D^{(r)}$, the neural net parameters, and thus the PDF replicas $f^{(r)}$, are not precisely determined through exact minimization of the χ^2 , but rather include some random noise, which is responsible for the ‘functional uncertainty’ inherent in the fit [15]. It is not easy to describe this analytically: all we can say is that while all the general results in Sect. 4.1 remain valid, relations such as Eq. (4.15) for the fitted parameters, and thus the subsequent results Eqs. (4.16, 4.17), no longer hold. However we can still use Eq. (4.13) to compute the expectation and covariance of the nuisance parameters, and obtain the same results Eqs. (4.16, 4.20), provided we define $T^{(r)} \equiv T[f^{(r)}]$, and $T^{(0)} \equiv \langle T^{(r)} \rangle$

$$X \equiv \text{Cov}[T[f]] = \langle (T^{(r)} - T^{(0)})(T^{(r)} - T^{(0)})^T \rangle \quad (4.30)$$

as averages over the PDF replicas. This matrix gives the PDF uncertainties (and correlations) for the observables $T[f]$, which includes both the experimental uncertainties in the data and the theoretical uncertainties in extracting the PDFs from the data.

Note that in an NNPDF fit X no longer satisfies the projective relation Eq. (3.13), and indeed $X(C + S)^{-1}$ no longer projects data replicas directly onto theory replicas as in Eq. (3.14). We can confirm this for a given set of PDF replicas by computing the cross-covariance matrix

$$Y \equiv \text{Cov}[T, D] = \langle (T^{(r)} - T^{(0)})(D^{(r)} - D)^T \rangle. \quad (4.31)$$

For a fixed parametrization, we can use Eq. (3.14) and Eq. (3.3) to show that then $Y = X = Y^T$. However it is easy to check by explicit computation that in an NNPDF fit Y is generally considerably smaller than X : the fluctuations of the theory replicas are not very well correlated to the fluctuations of the data replicas due to the functional uncertainty. So although many of the eigenvalues of $X(C + S)^{-1}$ will still be zero (because $m < N$), the nonzero eigenvalues will differ from one, and many will be somewhat larger than one due to the functional uncertainty. This means that while Eq. (4.10) still holds, the upper bound on $\bar{Z}_{\alpha\beta}$, Eq.(4.21), does not: the covariance of the nuisance parameters can be larger than the prior when the functional uncertainty is large.

Note that the fact that X is not invertible is not in any sense a technical limitation: the mapping of a global dataset into a set of PDFs cannot in principle be invertible (except possibly in certain special cases, such as data from a single process taken at a single scale [13]), since it is impossible to recover the data solely from the PDFs. This is in part because the PDFs are only functions of x , while the data also depend on

a scale: when we determine PDFs, all the data are effectively projected onto a common scale. But it is also because PDFs are by definition universal, i.e. process independent, so given a set of PDFs it is impossible in principle to say even which processes were used to determine them.

We can now derive general results for the expectation and covariance of autopredictions from the three matrices C , S and X , following the procedure set out in Sect. 3.2. The shifts in the autopredictions are given by a similar expression to Eq. (4.19),

$$\delta T[f] = -S(C + S)^{-1}(T^{(0)} - D), \quad (4.32)$$

and will reduce the experimental χ^2 as explained already in Sect. 2.2. The covariance matrix is still given by Eq. (4.23):

$$P \equiv \text{Cov}[T(f, \lambda)] = C(C + S)^{-1}X(C + S)^{-1}C + (S - S(C + S)^{-1}S). \quad (4.33)$$

If the theory uncertainty S is much smaller than the experimental uncertainty C , P approaches the result

$$P_{\text{con}} = X + S; \quad (4.34)$$

the fitting uncertainty and theoretical uncertainty can be combined in quadrature. So when the experimental uncertainties dominate there is almost complete decorrelation of the theoretical uncertainties, and the ‘conservative’ prescription recommended in Ref. [9] is a useful approximation.

Returning to the more general correlated result Eq. (4.33), both contributions to P (which we may call the correlated PDF uncertainty and the correlated theory uncertainty) are also positive semi-definite, and combine in quadrature to give the total uncertainty. Moreover the correlated theory uncertainty bounded above by the corresponding uncorrelated theory uncertainty:

$$0 \leq S - S(C + S)^{-1}S = C(C + S)^{-1}S \leq S, \quad (4.35)$$

It is tempting to also think that the correlated PDF uncertainty will also be bounded above by the uncorrelated PDF uncertainty X , because since C is positive definite, and S positive semi-definite, $C \leq C + S$, so $C(C + S)^{-1} \leq 1$, and $C(C + S)^{-1}X(C + S)^{-1}C \leq X$. This argument is wrong, however, and the correlated PDF uncertainty can sometimes exceed the uncorrelated. Writing

$$C(C + S)^{-1}X(C + S)^{-1}C = X - S(C + S)^{-1}X - X(C + S)^{-1}S + S(C + S)^{-1}X(C + S)^{-1}S, \quad (4.36)$$

in some circumstances the sum of the last three terms can be positive. For this reason it seems impossible to prove in general that $P \leq P_{\text{con}}$, though in all practical applications we have tested so far this seems to be the case.

For genuine predictions, with theory uncertainties correlated to those in the fit, shifts are given by Eq. (4.24),

$$\delta\tilde{T}[f] = -\hat{S}(C+S)^{-1}(T^{(0)} - D), \quad (4.37)$$

while Eq. (4.25) is most usefully written in the form

$$\begin{aligned} \tilde{P} &\equiv \text{Cov}[\tilde{T}(f, \lambda)] \\ &= \tilde{X} - \hat{X}(C+S)^{-1}\hat{S}^T - \hat{S}(C+S)^{-1}\hat{X}^T \\ &\quad + \hat{S}(C+S)^{-1}X(C+S)^{-1}\hat{S}^T \\ &\quad + (\tilde{S} - \hat{S}(C+S)^{-1}\hat{S}^T), \end{aligned} \quad (4.38)$$

where now besides the matrix X Eq. (4.30) we must now also evaluate

$$\tilde{X} \equiv \text{Cov}[\tilde{T}[f, \lambda]] = \langle (\tilde{T}^{(r)} - \tilde{T}^{(0)})(\tilde{T}^{(r)} - \tilde{T}^{(0)})^T \rangle, \quad (4.39)$$

$$\hat{X} \equiv \text{Cov}[\tilde{T}[f, \lambda], T[f, \lambda]] = \langle (\tilde{T}^{(r)} - \tilde{T}^{(0)})(T^{(r)} - T^{(0)})^T \rangle. \quad (4.40)$$

Again the covariance Eq. (4.38) separates into the sum in quadrature of a correlated PDF uncertainty (the first two lines) and a correlated theory uncertainty (the third line). When the cross-covariance $\hat{S}$ is very small, we obtain the conservative result

$$\tilde{P}_{\text{con}} = \tilde{X} + \tilde{S}, \quad (4.41)$$

proposed in Ref. [9]. This will typically be the case for predictions of new processes where the dominant MHOU is in the hard cross-section. However for processes already included in the fit, the situation is more complex, since $\tilde{S}$ and $\hat{S}$ may be large even if S is small.

Note that since the inclusion of MHOU in the PDF determination leads to only a small increase in the uncertainties of the PDFs [9], the conservative results Eq. (4.34, 4.41) give uncertainties very close to the conventional prescription, in which the PDF uncertainty (without MHOU) is combined in quadrature with the MHOU in the prediction. This will be particularly true when the MHOU in the prediction is larger than the PDF uncertainty.

5 Numerical results

In Sect. 4.3 we saw that in a realistic global PDF fit we can still use the same analytic expressions Eqs. (4.32, 4.33, 4.37, 4.38) for the shifts and reduction in uncertainties induced by the correlations between the theoretical uncertainties in fit and prediction as we would use in a fit with a fixed parametrization Sect. 4.2. This is despite the fact that PDFs are smooth functions, which cannot be determined uniquely from a finite set of discrete data, but necessarily have an additional ‘functional uncertainty’, so the PDF parameters are not fixed uniquely by the fit. All that is necessary is to evaluate the matrices X , $\tilde{X}$ and $\hat{X}$ Eqs. (4.30, 4.39, 4.40) as ensemble averages over the PDF replicas determined in the fit. In this

section we will compute these matrices in a realistic global PDF fit with theory uncertainties, and use them to evaluate autopredictions and genuine predictions including the effect of the correlated theoretical uncertainties.

We will carry out these studies in the context of the NNPDF3.1 NLO global fit with MHOU presented in Ref. [9]. This in turn employed the same experimental data and theory calculations in NNPDF3.1 [16] with two minor differences: the value of the lower kinematic cut was increased from $Q_{\text{min}}^2 = 2.69 \text{ GeV}^2$ to 13.96 GeV^2 , and the HERA F_2^b , fixed-target Drell–Yan cross-sections, and some LHC inclusive jet data were removed, for technical reasons. This left a total of $N_{\text{dat}} = 2819$ data points. The complete list of data included in the fit may be found in Tab. 6.3 of Ref. [9]. These data were divided into five classes, depending on the type of process involved, as summarized in Table 1. The MHOU covariance matrix S_{ij} was constructed using renormalization and factorization scale variations, by a factor of two either side. The factorization scale variations (estimating the MHOU in the NLO parton evolution) are correlated across all processes, but the renormalization scale variations (estimating the MHOU in the NLO hard cross-sections peculiar to each process) while correlated within data belonging to the same process, are uncorrelated between different processes. These variations were then combined to give S_{ij} using a 9pt scheme, as explained in Ref. [9]. The matrices C_{ij} and S_{ij} computed in Ref. [9] are reproduced in Fig. 1 as heat maps.

5.1 Covariance of PDF uncertainties X

We begin by computing the covariance matrix X_{ij} , Eq. (4.30), shown in Fig. 2 as a heat map alongside the corresponding correlation matrix. It can be seen that the off-diagonal elements of X_{ij} are almost as large as the diagonal elements: this is confirmed by examination of the correlation matrix. This is because theoretical predictions are often very strongly correlated, not only for nearby bins within the same experiment, but also for different processes at nearby scales, due primarily to the smoothness of the underlying PDFs, both in x and in Q^2 , but also due to the highly correlated theoretical uncertainties included in the fit.

We compare the PDF uncertainties to the experimental and theoretical uncertainties by looking at the per-point uncertainty (Fig. 3). Recall (Eq. (3.3)) that $C+S$ is the covariance of the data replicas to which the PDF replicas are fitted. It can be seen that at NLO the relative size of the experimental uncertainties C_{ii} and the theoretical uncertainties S_{ii} varies considerably between different datasets: for the fixed target DIS data S_{ii} is generally below C_{ii} , except at large x , whereas for the HERA NC data S_{ii} is much less than C_{ii} at large x , but the other way around at small x where the theoretical uncertainty dominates. For CHORUS, the experimental uncertainty also dominates, while for most DY datasets the

Table 1 Classification of datasets into process types

Process type	Datasets
DIS NC	SLAC, BCDMS, NMC, HERA NC
DIS CC	NuTeV, CHORUS, HERA CC
DY	CDF, D0, ATLAS, CMS, LHCb (dileptons, W and Z diff xsecs)
JET	ATLAS, CMS inclusive jets
TOP	ATLAS, CMS total and diff xsecs

Fig. 1 The experimental covariance matrix, C_{ij} , normalized to the theoretical predictions $T_i^{(0)}$ (left), and the corresponding theory covariance matrix for MHO, S_{ij} (right). The datasets are

arranged in the order given in Fig. 7 below: so SLAC data are in the top left corner, and LHC top data in the lower right corner

Fig. 2 The covariance matrix of PDF uncertainties, X_{ij} , normalized to the theoretical predictions $T_i^{(0)}$ (left), and the corresponding correlation matrix $X_{ij}/\sqrt{X_{ii}X_{jj}}$ (right). The datasets

are arranged in the order given in Fig. 7 below: so SLAC data are in the top left corner, and LHC top data in the lower right corner

Fig. 3 The square root of the diagonal elements of the matrices X (in orange), C (in green) and S (in purple) normalized to the theoretical predictions $T_i^{(0)}$, with those for C and S the same as in Ref. [9]. The datasets are arranged in the order given in Fig. 7 below

theoretical uncertainty dominates. In contrast the PDF uncertainties X_{ii} are generally less than either C_{ii} or S_{ii} , because the combination of data within given datasets and information from other datasets in the fit conspire to reduce the uncertainty. This is especially evident for DIS CC, DY and JETS. However for some data sets, particularly cross-section ratios with very small theory uncertainty (such as NMC d/p , the asymmetries, and the differential top data), X_{ii} lies above S_{ii} , though still below C_{ii} .

5.2 Nuisance parameters

Having computed X , we next calculate the nuisance parameters λ_α of the theory covariance matrix S_{ij} for the MHOU in the NNPDF3.1 NLO global fit. The posterior distributions of these nuisance parameters give us information on which directions of MHOU are constrained by the fit. We showed in Ref. [9] that when there are five different processes, there are 28 nonzero eigenvalues. Thus we have 28 nuisance parameters, in one-to-one correspondence with the 28 eigenvectors with nonzero eigenvalues. The eigenvalues are shown in descending order in Fig. 4, with their nuisance parameters below them. The expectation values of the nuisance parameters are computed using Eq. (4.18), and their uncertainties using Eq. (4.20). The nuisance parameters are normalized so that their prior is a unit gaussian centred on zero, as in Eq. (4.2). It can be seen that after fitting, the uncertainty in the nuisance parameters associated with the largest nine or so

Fig. 4 The 28 positive eigenvalues s^α of the theory uncertainty matrix S_{ij} (above), shown in descending order, and 28 nuisance parameters λ_α corresponding to the 28 eigenvectors β_α (below), as given by Eq. (4.18). The uncertainties in the nuisance parameters are shown in total (square roots of the diagonal entries of Eq. (4.20)), and broken down into the contribution from scale uncertainties alone (square roots of the diagonal entries of Eq. (4.11)) and from PDF uncertainties (square roots of the diagonal entries of the last term in Eq. (4.20)). The yellow bands highlight the region between ± 1

Fig. 5 Nuisance parameters λ for directions in the space of scale variations corresponding to up/down changes in factorization scale, and in renormalization scale for the five types of processes in the determination of the 9-pt theory covariance matrix for MHOUs. The uncertainties in the nuisance parameters are shown in total, and broken down into the contribution from scale uncertainties alone and from PDF uncertainties, just as in Fig. 4. The yellow bands highlight the region between ± 1

eigenvalues has been substantially reduced from one, indicating that exposure to the data has reduced the MHOUs. For those corresponding to the smaller eigenvalues there is very little reduction, showing that the data do not much constrain these directions in the space of MHOUs. The central values for the three largest eigenvalue nuisance parameters remain close to zero within uncertainties, showing that the prior choices (mainly overall normalizations) were reasonable, while the next three or four show significant deviations from zero: for these the data seem to carry significant information about the MHOUs. For the remaining (smaller) eigenvalues the central values of nuisance parameters are all consistent with zero, and clearly for the very small ones the data have no effect at all, the posterior distributions being the same as the prior. This shows that only the eigenvectors corresponding to the larger eigenvalues are actually relevant for the PDF determination: the remainder correspond to such small changes in theoretical uncertainty that the fit ignores them.

We can understand these features better by separating out the two contributions to the total uncertainty in the nuisance parameters: that due purely to the impact of the fit of a given replica on the MHOUs (given by Eq. (4.11)), and that due to the additional PDF uncertainty when the fits to all the replicas are averaged over PDF replicas, given by the last term in Eq. (4.20), also shown in Fig. 4. We see that when fitting

a single replica, the uncertainties in the nuisance parameters corresponding to the larger eigenvalues are indeed very substantially reduced: the MHOUs along the eigenvectors corresponding to these larger eigenvalues is learnt in the fit to the data, just as we saw in the simple models in Sect. 2.2. Again, very little information is retained about the smaller eigenvalues. The uncertainty contributed by averaging over the PDF replicas is also small for the largest eigenvalue nuisance parameters, but becomes the dominant contribution after the first three. For the smallest it is very small again: for these the data have no effect.

We can learn a little more about which MHOUs are learnt most by choosing different directions for the shift vectors β_α in Eq. (4.1) than the eigenvectors of S_{ij} . Specifically, we can choose the β_α to correspond to factorization scale variations (up or down), or renormalization scale variations (up or down, but now separately for each process). The results are shown in Fig. 5, where we again show the total uncertainty, scale uncertainty, and PDF uncertainty. The central values fluctuate about zero, but all remain within the band ± 1 , showing that the effect of fitting the experimental data on the nuisance parameters is rather mild: this is reassuring, as it confirms the choice of central scales used to make the predictions, and the choice of the range of the scale variations (implicit in the choice of the prior for the nuisance parameters, Eq. (4.2)). We also see that the uncertainties in the nuisance parameters corresponding to factorization scale variations (estimating the MHOUs in parton evolution), are reduced the most when fitting to any given replica, as we would expect since MHOUs in parton evolution are common to all data included in the fit. A little is also learnt about the renormalization scales for DIS. However, the PDF uncertainties partially wash out these effects. This suggests that the significant shifts in the nuisance parameters seen in Fig. 4 are due to global tensions between different processes, rather than problems with the choice of scales for particular processes.

Already we see that information from the data in the fit significantly updates the priors for the nuisance parameter distribution. From this it is likely that there will be an effect at the level of autopredictions, which is the subject of the next section.

5.3 Autopredictions

We now present results for the ‘autopredictions’. As we explained already in Sect. 2.2, these are the theoretical predictions we make for all the datasets included in the PDF fit, including theoretical uncertainties, after the fitting of the PDFs (with these same theoretical uncertainties). They can thus be thought of as ‘postdictions’, or predictions for the results of experiments run in exactly the same way, with the same equipment, as the original equipment but taking

account of the original global dataset. They thus form an ideal theoretical laboratory for testing the extent of the decorrelation between the theoretical uncertainties in the PDF fit, and those in the (auto)predictions.

We begin by computing the shifts δT_i , Eq. (4.32), in the autopredictions, due to the correlation in theoretical uncertainty: these are shown in Fig. 6, normalized to the original theoretical prediction $T_i^{(0)}$. We also show for comparison the differences $D_i - T_i^{(0)}$. It can be seen from the plot that these shifts are generally much smaller than the difference between data and theory, particularly for DIS NC and DY. However for some datasets (in particular CHORUS and inclusive jets), there seems to be an overall shift in central value of the same order as the difference between experiment and theory. However it is difficult to draw any further conclusions from these observations, since the shifts are very correlated within datasets.

To see whether the shifts actually improve the autopredictions, in Fig. 7 we show the experimental χ^2 for the original data, computed using the autopredictions for the fit with no theory uncertainties, those when the theory uncertainties are included in the PDF fit, and then when the autoprediction includes the shift. Needless to say all three results are generally very close, and including the theory uncertainties in the fit has mixed results, some predictions getting better, but at the expense of others getting worse, since the main effect of the theory uncertainties is to rebalance the datasets in the fit [9]. Nevertheless when the correlated shifts are included, the fit to most datasets improves, sometimes quite substantially, just as anticipated in the very simple ‘pure theory’ model in Sect. 2.2, and confirmed for the more realistic models in Sect. 3.2 and Sect. 4.2. The numbers broken down by process are shown in Table 2. When including theory uncertainties the total χ^2 increases just a little, from 1.17 to 1.19, but when the correlated shift is added to the theoretical predictions, we see a significant improvement to 1.10. This improvement is seen across all the processes.

Although these autopredictions are in some sense artificial – in practice experiments are never repeated using exactly the same equipment and settings – the implications of this exercise for the learning of theoretical uncertainties are nevertheless rather general. This is because in a global fit of the size of that performed here, with 2819 data points from 35 datasets, involving five different processes, removing any one of the smaller datasets has very little impact on the PDFs, and removing any dataset has the effect of increasing PDF uncertainties, while theoretical uncertainties for the remaining data remain unchanged. Consequently if we were to perform the PDF fit without a given dataset, and repeat the analysis, so that the autoprediction becomes a genuine prediction (or more properly ‘postdiction’), the result for this genuine prediction would be very close to the autoprediction. So we

expect the shifts in central values that we see in the autopredictions to also give improvements in such predictions: the shifts should improve the accuracy of the predictions.

To see whether we can also increase the precision, we consider the uncertainties in the autopredictions. In Fig. 8 we show the full covariance matrix P_{ij} , Eq. (4.33), again normalized to the theoretical predictions, and corresponding correlation matrix. The matrix P_{ij} is the sum of the PDF uncertainty (derived from data uncertainties and theoretical uncertainties combined) and the theoretical uncertainty in the autoprediction, each reduced in size to account for the learning of the theoretical uncertainties and the correlation between the two sources of theoretical uncertainty. As might be expected there are very large correlations in the autopredictions within datasets, particularly for nearby kinematic points, but there are also smaller correlations, and anticorrelations, between datasets. They are due not only to the correlations of experimental uncertainties within datasets, but also to the use of a common set of smooth underlying PDFs, and the correlations of the theory uncertainties. The correlations within each process are generally larger than those between processes. This suggests that the combined effects of the correlations due to the use of a common factorization scale and the correlations induced by the smoothness of the PDFs is small compared to the correlations from the renormalization scale.

In Fig. 9 we show the percentage uncertainties of the autopredictions: $\sqrt{P_{ii}}$, compared to the purely PDF uncertainties $\sqrt{X_{ii}}$ to aid comparison. The correlated autoprediction uncertainties are generally of similar size to the PDF uncertainties; they are rather larger for some of the DY datasets and JETS, but are actually smaller for most of the DIS NC data (most remarkably for the HERA data at small x), and some DY data. So the full autopredictions are not only more accurate: they are also more precise. This increase in precision must be taken with a pinch of salt, since it depends to some extent on the assumptions made in modelling the prior MHO [9], in particular the choice of independent scales, and the scheme through which they are combined into the theory covariance matrix S . In particular the aggressive reduction in the small x uncertainties for the HERA NC autopredictions seen in Fig. 9 may be due to the adoption of the same factorization scale for both singlet and nonsinglet, which overconstrains the singlet evolution at small x [13]. We leave the relaxation of these kinds of assumptions for future work.

As expected all of the autopredictions uncertainties are smaller than would be obtained by the standard prescription of adding PDF uncertainties and theory uncertainties in quadrature [9], which ignores both learning and correlation. However the conservative approach overestimates the correlated uncertainty for almost all datapoints, typically by a factor of two or more, particularly those for which the the-

Fig. 6 The shifts δT_i , Eq. (4.32) (in blue) compared to the differences between theory and data, $D_i - T_i^{(0)}$ (in green), both normalized to $T_i^{(0)}$

Fig. 7 The experimental χ^2 for each data set, comparing the original result of the NLO fit with no theory uncertainties to the fit with theory uncertainties, and then including the correlated shift in the autopredictions

Table 2 The experimental χ^2 per data point for each process, comparing the original result of the NLO fit with no theory uncertainties to the fit with theory uncertainties, and then including the shift in the autopredictions

	DIS NC	DIS CC	DY	JETS	TOP	Total
No th uncs	1.13	0.98	1.56	0.88	1.20	1.17
Uncorr th uncs	1.15	1.06	1.53	0.90	1.27	1.19
Correlated th uncs	1.09	0.91	1.47	0.83	0.97	1.10

Fig. 8 The autoprediction covariance matrix P_{ij} Eq. (4.33), normalized to the theoretical predictions $T_i^{(0)}$ (left), and the corresponding correlation matrix $P_{ij}/\sqrt{P_{ii}P_{jj}}$ (right)

Fig. 9 The percentage uncertainties of the autopredictions $\sqrt{P_{ii}}$ Eq. (4.33) (blue) compared to the PDF uncertainty $\sqrt{X_{ii}}$ (orange), and the conservative result, $\sqrt{P_{ii}^{con}}$ Eq. (4.34) (cyan), all normalised to the theoretical predictions $T_i^{(0)}$

oretical uncertainty is larger than the PDF uncertainty. The only data for which the conservative prescription works well are ratio data (for example the NMC d/p data), for which theoretical uncertainties are very small.

To understand better how these changes in uncertainty arise, we show in Fig. 10 the contributions to the diagonal elements of the correlated theory uncertainty (the second term in Eq. (4.33)), normalised to the uncorrelated elements S_{ii} .

The ‘learning’ of the theoretical uncertainty, given by the contribution $-S(C+S)^{-1}S$ (note that $S - S(C+S)^{-1}S$ is equal in the one parameter example described in Sect. 3.2 to ZS) is very significant, reducing the prior uncertainty S almost to zero for NC DIS and DY (where there is considerable data), and by an order of magnitude for DIS CC, JETS and TOP. Probably more flexibility is required in the modelling of the prior for this uncertainty. However

Fig. 10 The contributions to the diagonal elements of the correlated theory uncertainty normalised to diagonal elements of S : $(S - S(C + S)^{-1}S)_{ii}/S_{ii}$ (pink), and $(S - S(C + S)^{-1}S + S(C + S)^{-1}X(C + S)^{-1}S)_{ii}/S_{ii}$ (black)

Fig. 11 The contributions to the diagonal elements of the correlated PDF uncertainty normalised to diagonal elements of X : $(X - S(C + S)^{-1}X - X(C + S)^{-1}S)_{ii}/X_{ii}$ (lilac), and $(C(C + S)^{-1}X(C + S)^{-1}C)_{ii}/X_{ii}$ (see Eq. (4.36)) (green)

the PDF fluctuations $S(C + S)^{-1}X(C + S)^{-1}S$ (note that $S - S(C + S)^{-1}S + S(C + S)^{-1}X(C + S)^{-1}S$ is equal in the one parameter example described in Sect. 3.2 to $\bar{Z}S$) undo much of this learning, though for all data points this effect is insufficient to take the ratio to S above one.

A similar breakdown of the contributions to the diagonal elements of the correlated PDF uncertainty (the first term in Eq. (4.33), which can be expanded as in Eq. (4.36)), normalised to the uncorrelated elements X_{ii} , is shown in Fig. 11. The correlation terms $-S(C + S)^{-1}X - X(C + S)^{-1}S$ are

indeed very large, as anticipated in Ref. [13], in particular for data with relatively large theoretical uncertainty (such as HERA NC at small x , or JETS), sufficient there to overwhelm X and give a negative result. However they can also be positive for some data (such as JETS). In any event, the addition of the PDF fluctuation term $S(C + S)^{-1}X(C + S)^{-1}S$ (remember the decomposition Eq. (4.36) of the total correlated PDF uncertainty) is always sufficient to restore positivity of the correlated PDF uncertainty, and can in some situations (where S is large, in particular for JETS, but also some DIS CC and DY data) take the total correlated PDF uncertainty above the uncorrelated result X . Thus the correlations, while generally reducing uncertainties, can in some circumstances increase them, in contrast to learning which always reduces them.

For autopredictions we expect high levels of learning and correlation because we are making predictions for exact repeats of experiments already in the fit. As we noted for the shifts however, removing a smaller dataset from the fit has little effect in the PDFs, so we might expect similar effects for genuine predictions of processes already included in the fit, particular if they are in a similar kinematic region.

5.4 Predictions for top

We now consider genuine predictions, for experiments not used in the PDF fit. These are of two kinds: those for datasets obtained through processes already contained in the fit, and those for completely new processes. For the former we consider the $t\bar{t}$ production rapidity distributions (dilepton and lepton+jets) measured by CMS at 13 TeV [17, 18]. We chose these datasets for two reasons: firstly, the MHOU is large compared to the experimental uncertainty; secondly, the fit in Ref. [9] contains the total $t\bar{t}$ cross-sections at 7, 8 and 13 TeV, and the normalized rapidity distributions at 8 TeV, all from both ATLAS and CMS. Both these factors mean that we expect to see significant effects due to correlations between the theoretical uncertainties of the data in the PDFs and the theoretical uncertainties of the 13 TeV rapidity distributions.

To make genuine predictions, including all correlations, we need to first calculate the covariance matrix for the MHOU of the predictions, $\tilde{S}_{IJ}$, and its cross-covariance with the MHOU of the theoretical predictions for the data used in the PDF fit, $\hat{S}_{ij}$ (the indices I, J running over the predicted data points, while i, j run over the data included in the PDF fit): these are in fact the same as if we were planning to include data for the new process in a PDF fit, since the complete covariance matrix for the MHOU would be then of the form Eq. (2.41). Similarly we need to compute, using the fitted PDF replicas the covariance matrix $\tilde{X}_{IJ}$ of the PDF uncertainty for the new predictions, and the cross-covariance $\hat{X}_{ij}$ with the PDF uncertainty of the observables used in the fit. All of these matrices are required for an exact calculation

of the correlated shifts in the theoretical predictions $\delta\tilde{T}_I$, Eq. (4.37) and the covariance matrix $\tilde{P}_{IJ}$ of their combined PDF and correlated theoretical uncertainties Eq. (4.38).

Predictions for the CMS 13 TeV $t\bar{t}$ production rapidity distributions were computed using the same tool chain as in Ref. [16] for the 8 TeV distributions: NLO theoretical predictions were generated with Sherpa [19], in a format compliant with APPLgrid [20], using the MCgrid code [21] and the Rivet [22] analysis package, with OpenLoops [23] for the NLO matrix elements. Renormalization and factorization scales have been chosen based on the recommendation of Ref. [24] as $H_T/4$.

The results of these calculations are shown in Fig. 12. The prior theoretical uncertainty in the original prediction is around 10%, considerably greater than the PDF uncertainty, as expected since the hard cross-sections are only computed at NLO. The correlated shift is sizeable, around 5%, and almost fully correlated across all the rapidity distributions. This is because these are unnormalized distributions, and thus have an overall theoretical normalization uncertainty which is strongly correlated to the measurements of the total $t\bar{t}$ cross-sections at 7, 8 and 13 TeV by ATLAS and CMS included in the PDF fit. This is confirmed by breaking down the contributions to the shift Eq. (4.24) from the various data points included in the fit: the results are shown in Table 3. All of the shift comes from the six total cross-section measurements, while the 8 TeV normalized rapidity distributions push it down again by around 25%. The remaining data make almost no contribution. Nevertheless, the shift is still rather less than the theoretical uncertainty in the original prediction, as expected from the shifts in the nuisance parameters for the renormalization scale variation for top processes shown in Fig. 5.

We can compare the shift due to correlations to that from going from NLO to NNLO. We thus show in Fig. 12 the results of a complete NNLO calculation (without theory uncertainties): the NNLO corrections also increase predictions by 5–8%. It is very interesting that the shift, driven by the data for the $t\bar{t}$ total cross-sections at 7, 8 and 13 TeV, largely accounts for the NNLO correction: the data know that the NLO theoretical predictions are on the low side, and this information is carried over into the prediction for the 13 TeV rapidity distributions. Indeed, we compare the experimental χ^2 for these data in the various calculations in Table 4: while the theory uncertainties increase the χ^2 (due presumably to the other top data being deweighted in the fit), the shift gives a significant improvement both for dileptons and lepton+jets, comparable to that obtained with the complete NNLO corrections. So the shifts provide a new method for using experimental data to make improved theoretical predictions through the learning of theoretical uncertainties. The method should be particularly effective when there is sub-

Fig. 12 The upper two panels show predictions for $t\bar{t}$ unnormalized rapidity distribution data taken at 13 TeV by CMS, the dilepton rapidity distribution [17] (left) and the lepton+jets distribution [18] (right). The four predictions show: the NLO fit with no MHOU, PDF error only; the combined PDF and MHOU fit, ignoring correlations (thus $\sqrt{P_{II}^{cons}}$); the correlated result including the shift, and uncertainty computed using

the simplified result (thus $\sqrt{P_{II}^{sim}}$); the result with the same shift, but with the correlations included exactly (thus P_{II}), and the NNLO result with no MHOU. In the middle panels the same is shown, but normalized to the uncorrelated result. In the lower panels we show the fractional reduction in the PDF uncertainty and the theory uncertainty due to the inclusion of the correlations

Table 3 The fractional contributions of different data sets included in the fit to the shifts in the top rapidity distributions, averaged over all 21 data points

ATLAS				CMS				Other
Tot		Diff		Tot		Diff		
7 TeV	8 TeV	13 TeV	8 TeV	7 TeV	8 TeV	13 TeV	8 TeV	
0.37	0.11	0.24	− 0.21	0.26	0.21	0.07	− 0.04	− 0.01

stantial data on the process to be predicted already included in the PDF fit, as is the case here.

The middle panels of Fig. 12 shows the same points as the top panel, but as a ratio to the conservative result, making the uncertainties more visible. Comparing the uncertainties, the difference between the uncorrelated and correlated uncertainty is striking; the correlated uncertainties are much smaller than the uncorrelated. The correlated uncertainties

are however still larger than the purely PDF uncertainties, but the very large theoretical uncertainty has been substantially reduced. So not only are the correlated predictions more accurate, they are also more precise. Despite this significant shrinking of uncertainties, the correlated predictions are still compatible with the NNLO result, thanks to the shift in central values. While the conservative prescription is also com-

Table 4 The experimental χ^2 per data point for the CMS 13 TeV top dilepton and lepton+jet rapidity distributions, comparing the original result of the NLO fit with no theory uncertainties to the fit with theory uncertainties, and then including the correlated shift in the autopredictions. Also shown for comparison is the result in a NNLO fit with no theory uncertainties, only PDF uncertainties

	Dilepton	Lepton+jet	Combined
No th uncs	0.55	0.37	0.46
Uncorr th uncs	0.57	0.42	0.49
Correlated th uncs	0.49	0.37	0.43
NNLO, no th uncs	0.45	0.39	0.42

patible with the NNLO result, it is immediately clear from the plot that it is inferior to the correlated prediction.

The breakdown of the reduction in uncertainties due to the correlations is shown in the lower panels of Fig. 12. The correlated theory uncertainty is substantially reduced (uniformly across the rapidities), due to the learning of the normalization from the data already included in the fit, while the correlated PDF uncertainty is reduced rather less: as much as a factor of two when the differential cross-section is small, but hardly at all when it is large. So here the dominant effect is clearly the learning of the theoretical uncertainty in the overall normalization.

The theoretical uncertainties in the theoretical predictions are strongly correlated amongst themselves, and between the two rapidity distributions: in Fig. 13 we show the correlation matrices for the PDF uncertainties, $\tilde{X}_{IJ}$, and that of the correlated prediction $\tilde{P}_{IJ}$. We see that while the predictions are more than 50% correlated across the range of rapidities by the PDF, when the correlated theoretical uncertainties are also included all points are rather more correlated, to more than 70%. The pattern of correlations reflects the symmetry in the dilepton distribution and asymmetry in the lepton+jet distribution: the least correlated points are those with the greatest rapidity separation.

5.5 Predictions for Higgs

As an example of a process not included in the PDF fit we consider the prediction for the total cross-section for Higgs production in gluon fusion at 14 TeV. For the calculation of the cross-section we performed calculations using `ggHiggs` [25–27]. Renormalization and factorization scales are set to $m_h/2$, and the computation is performed using rescaled effective theory.

Our results are shown in in Fig. 14. The PDFs are still the NLO PDFs with MHOU from Ref. [9], but the Higgs total cross-sections are computed at NLO, NNLO and N3LO. At NLO the MHOU in the Higgs cross-section, estimated by varying the renormalization scale, completely dominates

all other uncertainties (the PDF uncertainty, the theoretical uncertainty in the PDFs, and the scale uncertainty in the PDF evolution, estimated by factorization scale variation), so the effect of correlations in the MHOU is completely negligible. However when the cross-section is computed at NNLO, the renormalization scale uncertainty is rather smaller, and at N3LO it becomes more comparable to the other sources of uncertainty. The shift due to the correlation between MHOU is always small compared to the overall uncertainty, and becomes smaller still as the perturbative order of the cross-section is increased, as expected. This is because, unlike in the top predictions, data for this process are not included in the fit, so the renormalization scale uncertainty is completely uncorrelated. It is interesting to note however that the small shift due to the correlation in factorization scale takes the NNLO prediction very close to the result computed with NNLO PDFs (though the coincidence is surely accidental).

Unlike for the top predictions, the effect of the correlation on the size of the overall uncertainty is small. Again, this is because the PDF contains far less information about Higgs production than it does about top production. We saw in Fig. 8 that the information propagated through to the correlated uncertainties was primarily through the renormalization scales, and that correlation due to the factorization scale was a lot weaker. Higgs production, being a new process, is therefore only impacted through the weak factorization scale correlations, and so the reduction in uncertainties is small. Overall, the uncorrelated conservative prescription [9] is comparable in this case to the fully correlated one. Note that if we were to perform these studies with NNLO (or indeed N3LO) PDFs which include MHOU, the MHOU in the PDF determination would have presumably been rather smaller than at NLO, and thus the effect of correlations in the MHOU, in particular the shift, but also the effect on the size of the uncertainty would be even smaller than the small corrections we see here.

From these examples of autopredictions, and genuine predictions for top and Higgs, we have seen that the extent of the shift and correlation can vary quite significantly, depending on the type of prediction being made and what information is already contained in the PDFs. The conservative prescription recommended in Ref. [9] is not always optimal, as the full inclusion of correlations can be quite substantially reduce uncertainties, as we saw both for the autopredictions and top predictions. However, when predicting a new process for which the PDF contains little information about correlated theoretical uncertainties, unsurprisingly the impact of correlations is small and the conservative prescription is quite sufficient.

Fig. 13 The left hand plot shows the correlation matrix $\tilde{X}_{IJ}/\sqrt{\tilde{X}_{II}\tilde{X}_{JJ}}$ of the contribution of the PDF uncertainties to the predictions for the 13 TeV rapidity distributions by CMS: the right

hand plot shows the correlation matrix $\tilde{P}_{IJ}/\sqrt{\tilde{P}_{II}\tilde{P}_{JJ}}$ of the total uncertainties including the correlated theoretical uncertainties. Note the expanded scales on the heat maps, different in each plot

Fig. 14 Predictions for the Higgs total cross-section at 14 TeV, made using a variety of approximations. All results use NLO PDFs, while the Higgs total cross-section is computed at NLO (left panel), NNLO (centre panel) and N³LO (right panel). In each panel, we then have, from left to right: MHOU included only in the PDF determination in the 9pt scheme; the same but with the factorization scale uncertainty (MHOU in PDF evolution) included in quadrature; the same but with instead the renormalization scale uncertainty (MHOU in the Higgs cross-section); the total PDF uncertainty and 9pt MHOU combined in quadrature, as recommended in Ref. [9]; the total PDF plus 9pt MHOU, but now including also the shift and the correlation between theoretical uncertainties. In the centre panel we also show the NNLO prediction with NNLO PDFs (but no theoretical uncertainties), as a dashed line

6 Summary

In this paper we studied in detail the correlation between theoretical uncertainties in the calculations used in the determination of PDFs in a global fit, as formulated in Ref. [9], and the theoretical uncertainties in the predictions made using these PDFs. We began by recasting the theoretical uncertainties using nuisance parameters, determined replica by replica, which carry all the information about the effect of the experimental data on the theoretical uncertainties. Using increasingly realistic models of the fitting procedure, we produced analytical formulae for computing fully correlated predictions. In the process we identified three distinct but related effects, each of which has a significant impact on the final theoretical predictions:

- **Shifts in central values.** These are an effect of Bayesian learning: just as we can use experimental data to determine PDFs, so we can also use it to identify theoretical corrections that improve the agreement between data and theory, while remaining within theoretical uncertainties. The correlations between theoretical uncertainties in the fit and those in predictions then lead to more accurate predictions. This effect was first identified in Sect. 2.2.
- **Learning of theoretical uncertainties.** A second consequence of Bayesian learning is a reduction in theoretical uncertainty, due to the information provided by the data in the PDF fit, which through correlation of theoretical uncertainties can lead to a corresponding reduction in the theoretical uncertainties in predictions. This effect was

also identified in Sect. 2.2, and is complementary to the shift in central values.

- **Correlation in theoretical uncertainties.** The third effect is that the correlation between the theoretical uncertainties in the fit and the theoretical uncertainties in the predictions lead to a change in the PDF uncertainties in the prediction, even in situations where there is no shift, thus avoiding any ‘double counting’ of the theoretical uncertainty. The existence of this effect was first noted in Ref. [13], and identified as an effect distinct from Bayesian learning in Sect. 2.3.

While these three effects were first identified in the simple models of Sect. 2, we showed that they are all present in the one parameter fits of Sect. 3 and the more realistic fits with multiple parameters in Sect. 4. Using the NNPDF3.1 NLO global fits with MHOU [9], we demonstrated in Sect. 5 that the shifts can give sensible estimates of NNLO corrections, and thereby reduce the χ^2 to the experimental data. We also showed that while the uncertainty in NLO predictions is still a sum in quadrature of the theoretical uncertainty and the PDF uncertainty (which also includes a theory uncertainty), this sum can be significantly reduced, depending on the relative size of the theoretical and experimental uncertainties. Consequently the ‘conservative’ prescription of Ref. [9], where the theory uncertainty in the prediction is combined in quadrature with the PDF uncertainty, is indeed conservative. We expect these conclusions to also hold in global PDF fits with fixed parametrization and tolerance [28,29], if these were to include MHOU in the PDF fit.

The degree of correlation is highly dependent on the type of prediction being made. For the autopredictions (predictions for new measurements of the same data points as those included in the fit), Sect. 5.3, where there is maximal correspondence between the data in the fit and the predictions being made, the correlation is very high, leading to shifts that improve the quality of the fit to the data, together with a significant reduction in uncertainties, in some cases down to a small fraction of the uncorrelated values. For genuine predictions for new measurements of processes already included in the PDF fit, such as the new measurements of differential top production discussed in Sect. 5.4, we observe that the shift takes the correlated NLO predictions very close to the NNLO prediction, with a significant reduction in uncertainties: the prediction is both more accurate and more precise. For Higgs production, discussed in Sect. 5.5, a process not included in the PDF fit, the level of correlation is much smaller, since the dominant uncertainty (the MHOU in the hard cross-section) is uncorrelated with the MHOU of the fitted processes. In this case the shift is well within uncertainties, and the reduction in uncertainty very modest, so here the use of the conservative prescription Ref. [9] is entirely appropriate. We expect this to be true of predictions for any new process.

Thus our main conclusion is that when using PDFs which include MHOUs, taking account of the correlations between the MHOU included in the determination of the PDFs and the MHOU in the prediction can result in a significant improvement in both accuracy and precision. This is especially true in the case where the predicted process is among those included in the fit. However the correlated predictions must be treated with care, since their reliability relies to some extent on the generality of the prior estimation of the MHOU: if unjustified assumptions are made in the choice of prior, the uncertainty estimates in the correlated predictions may be too aggressive. For these reasons the conservative prescription, as an upper bound on the overall uncertainty, may sometimes be preferable, especially for predictions of new processes.

In order to calculate fully correlated predictions and uncertainties, one requires besides the PDF replicas some additional information: the cross-correlations between the theoretical uncertainties in the prediction and those in the theoretical calculations used to determine the PDFs, $\hat{S}_{Ij}$; and the cross-correlations between the PDF uncertainties in the prediction and all the calculations included in the fit, $\hat{X}_{Ij}$. In the future, it may be possible to present this information in separate NNPDF deliverables to facilitate the calculation of the correlation effects.

Although we presented our numerical study of correlations in the context of MHOUs, we would expect similar results for other kinds of theoretical uncertainty, such as nuclear uncertainties, higher twist uncertainties, or indeed parametric uncertainties: once the theory covariance matrix has been computed, the linear algebra has no concern for the type of theoretical uncertainty it contains. This suggests a new technique for determining external parameters in PDF fits, such as quark masses or electroweak parameters, taking full account of all correlations with the PDFs and theoretical uncertainties. We hope to explore this possibility in the near future.

Acknowledgements We would like to offer our warmest thanks to Emanuele Nocera for his assistance in the numerical computations. We would also like to thank an anonymous referee for comments which significantly improved the final draft of the paper. R.D.B. and R.L.P. are supported by the UK Science and Technology Facility Council through grants ST/P000630/1 and ST/R504737/1.

Data Availability Statement This manuscript has no associated data or the data will not be deposited. [Authors’ comment: The data associated with this manuscript are not deposited as they are not intended for public use, however they are available from the authors upon request.]

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not

included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.
Funded by SCOAP³.

References

1. S. Forte, S. Carrazza, Parton distribution functions. [arXiv:2008.12305](#)
2. J.J. Ethier, E.R. Nocera, Parton distributions in nucleons and nuclei. *Annu. Rev. Nucl. Part. Sci.* **70**, 1–34 (2020). [arXiv:2001.07722](#)
3. P. Azzi et al., Report from Working Group 1: standard model physics at the HL-LHC and HE-LHC. CERN Yellow Rep. Monogr. **7**, 1–220 (2019). [arXiv:1902.04070](#)
4. R. D. Ball, A. Deshpande, The proton spin, semi-inclusive processes, and measurements at a future Electron Ion Collider. in *From My Vast Repertoire: Guido Altarelli's Legacy* ed. by A. Levy, S. Forte, G. Ridolfi (2019). [arXiv:1801.04842](#)
5. R.D. Ball, E.R. Nocera, R.L. Pearson, Nuclear uncertainties in the determination of proton PDFs. *Eur. Phys. J. C* **79**(3), 282 (2019). [arXiv:1812.09074](#)
6. R.D. Ball, E.R. Nocera, R.L. Pearson, Deuteron uncertainties in the determination of proton PDFs. *Eur. Phys. J. C* **81**(1), 37 (2021). [arXiv:2011.00009](#)
7. R.L. Pearson, C. Voisey, Towards parton distribution functions with theoretical uncertainties. *Nucl. Part. Phys. Proc.* **300–302**, 24–29 (2018). [arXiv:1810.01996](#)
8. NNPDF Collaboration, R. Abdul Khalek et al., A first determination of parton distributions with theoretical uncertainties. *Eur. Phys. J. C* **79**, 838 (2019). [arXiv:1905.04311](#)
9. NNPDF Collaboration, R. Abdul Khalek et al., Parton distributions with theory uncertainties: general formalism and first phenomenological studies. *Eur. Phys. J. C* **79**(11), 931 (2019). [arXiv:1906.10698](#)
10. M. Cacciari, N. Houdeau, Meaningful characterisation of perturbative theoretical uncertainties. *JHEP* **09**, 039 (2011). [arXiv:1105.5152](#)
11. E. Bagnaschi, M. Cacciari, A. Guffanti, L. Jenniches, An extensive survey of the estimation of uncertainties from missing higher orders in perturbative calculations. *JHEP* **02**, 133 (2015). [arXiv:1409.5036](#)
12. M. Bonvini, Probabilistic definition of the perturbative theoretical uncertainty from missing higher orders. *Eur. Phys. J. C* **80**(10), 989 (2020). [arXiv:2006.16293](#)
13. L.A. Harland-Lang, R.S. Thorne, On the consistent use of scale variations in PDF fits and predictions. *Eur. Phys. J. C* **79**(3), 225 (2019). [arXiv:1811.08434](#)
14. R.D. Ball et al., Parton distribution benchmarking with LHC data. *JHEP* **04**, 125 (2013). [arXiv:1211.5142](#)
15. NNPDF Collaboration, R.D. Ball et al., Parton distributions for the LHC Run II. *JHEP* **04**, 040 (2015). [arXiv:1410.8849](#)
16. NNPDF Collaboration, R.D. Ball et al., Parton distributions from high-precision collider data. *Eur. Phys. J. C* **77**(10), 663 (2017). [arXiv:1706.00428](#)
17. CMS Collaboration, A.M. Sirunyan et al., Measurements of $t\bar{t}$ differential cross sections in proton–proton collisions at $\sqrt{s} = 13$ TeV using events containing two leptons. *JHEP* **02**, 149 (2019). [arXiv:1811.06625](#)
18. CMS Collaboration, A.M. Sirunyan et al., Measurement of differential cross sections for the production of top quark pairs and of additional jets in lepton+jets events from pp collisions at $\sqrt{s} = 13$ TeV. *Phys. Rev. D* **97**(11), 112003 (2018). [arXiv:1803.08856](#)
19. T. Gleisberg, S. Hoeche, F. Krauss, M. Schonherr, S. Schumann, F. Siegert, J. Winter, Event generation with SHERPA 1.1. *JHEP* **02**, 007 (2009). [arXiv:0811.4622](#)
20. T. Carli, D. Clements, A. Cooper-Sarkar, C. Gwenlan, G.P. Salam, F. Siegert, P. Starovoitov, M. Sutton, A posteriori inclusion of parton density functions in NLO QCD final-state calculations at hadron colliders: the APPLGRID Project. *Eur. Phys. J. C* **66**, 503–524 (2010). [arXiv:0911.2985](#)
21. L. Del Debbio, N.P. Hartland, S. Schumann, MCgrid: projecting cross section calculations on grids. *Comput. Phys. Commun.* **185**, 2115–2126 (2014). [arXiv:1312.4460](#)
22. A. Buckley, J. Butterworth, D. Grellscheid, H. Hoeth, L. Lonnblad, J. Monk, H. Schulz, F. Siegert, Rivet user manual. *Comput. Phys. Commun.* **184**, 2803–2819 (2013). [arXiv:1003.0694](#)
23. F. Cascioli, P. Maierhofer, S. Pozzorini, Scattering amplitudes with open loops. *Phys. Rev. Lett.* **108**, 111601 (2012). [arXiv:1111.5206](#)
24. M. Czakon, D. Heymes, A. Mitov, Dynamical scales for multi-TeV top-pair production at the LHC. *JHEP* **04**, 071 (2017). [arXiv:1606.03350](#)
25. R.D. Ball, M. Bonvini, S. Forte, S. Marzani, G. Ridolfi, Higgs production in gluon fusion beyond NNLO. *Nucl. Phys. B* **874**, 746–772 (2013). [arXiv:1303.3590](#)
26. M. Bonvini, R.D. Ball, S. Forte, S. Marzani, G. Ridolfi, Updated Higgs cross section at approximate N³LO. *J. Phys. G* **41**, 095002 (2014). [arXiv:1404.3204](#)
27. M. Bonvini, S. Marzani, C. Muselli, L. Rottoli, On the Higgs cross section at N³LO+N³LL and its uncertainty. *JHEP* **08**, 105 (2016). [arXiv:1603.08000](#)
28. S. Bailey, T. Cridge, L.A. Harland-Lang, A.D. Martin, R.S. Thorne, Parton distributions from LHC, HERA, Tevatron and fixed target data: MSHT20 PDFs. *Eur. Phys. J. C* **81**(4), 341 (2021). [arXiv:2012.04684](#)
29. T.-J. Hou et al., New CTEQ global analysis of quantum chromodynamics with high-precision data from the LHC. *Phys. Rev. D* **103**(1), 014013 (2021). [arXiv:1912.10053](#)