2.2 The loss function

The χ^2 with correlated systematics, the t_0 prescription, and what a good fit quality means.

---

1. What the loss function has to do

Chapter 2.1 turned a set of parameters θ into predictions $T_i(\theta)$ for every data point. A fit needs one more ingredient: a single number that says how far those predictions are from the measurements D_i, given how uncertain the measurements are. That number is the loss function. Minimising it gives the best fit, its shape around the minimum gives the uncertainties (chapter 2.3), and its value at the minimum says whether the fit is acceptable.

This chapter builds the loss function used in parton distribution function (PDF) fits and derives each of its properties. It has five parts.

1. For Gaussian measurement errors the loss is the χ^2, a quadratic form in the residuals $D_i - T_i$ with the inverse covariance matrix C^{-1} (section 2).
2. The covariance matrix is assembled from the error sources that experiments publish: independent errors on the diagonal, and one rank-one term for each systematic source shared between points (section 3).
3. The same χ^2 can be written with one extra fit parameter per systematic source. The two forms are identical, which is proved in full (section 4).
4. Errors proportional to the measured value, such as a luminosity uncertainty, bias the fit downwards if the covariance matrix is built from the data. This is the d'Agostini bias. It is derived for any number of points and removed by the t_0 prescription (sections 5 and 6).
5. For a correct model the minimum χ^2 is close to the number of data points N_{dat}, within $\sqrt{2N_{\mathrm{dat}}}$. Global fits miss this target by a wide margin, and the reasons are specific (sections 7 and 8).

Sections 9 and 10 add the two modern extensions: theory uncertainties in the covariance matrix, and penalty terms for positivity and integrability.

The chapter follows Moore's thesis, section 1.3.2 [1], and Ubiali's thesis, sections 2.1.3 and 2.2 [2], for the definitions and the bias, and the Snowmass whitepaper [5] for correlated systematics and theory uncertainties. The form of the bias for any number of points, Eq. (2.2.14), and the statistics of section 7 are worked out here. Every formula is checked numerically by the script linked at the end.

2. The Gaussian likelihood and the chi-squared

From a probability to a loss

Collect the measurements into a vector D with N_{dat} components, the predictions into $T(\theta)$, and write $r = D - T(\theta)$ for the residuals. Assume the measurements scatter about the true values with a multivariate Gaussian distribution of covariance matrix C, where C_{ij} is the expected value of the product of the errors on points i and j. The probability density of the data, given the parameters, is the *likelihood*:

$$\mathcal{L}(D \,|\, \theta) = \frac{1}{\sqrt{(2\pi)^{N_{\mathrm{dat}}} \det C}} \, \exp\Big[-\tfrac{1}{2}\, r^{\mathsf{T}} C^{-1} r \Big]. \tag{2.2.1}$$

The parameters that make the observed data most probable are those that minimise the exponent. That exponent defines the loss:

$$\chi^2(\theta) = \big(D - T(\theta)\big)^{\mathsf{T}} C^{-1} \big(D - T(\theta)\big) = \sum_{i,j=1}^{N_{\mathrm{dat}}} r_i \, (C^{-1})_{ij} \, r_j, \qquad \mathcal{L} \propto e^{-\chi^2/2}. \tag{2.2.2}$$

Minimising χ^2 is maximising the likelihood only if C does not depend on θ. If it did, the factor $\det C$ in Eq. (2.2.1) would vary too. This condition returns in section 5.

If the errors are independent, C is diagonal with entries σ_i^2 and

$$\chi^2(\theta) = \sum_{i=1}^{N_{\mathrm{dat}}} \frac{\big(D_i - T_i(\theta)\big)^2}{\sigma_i^2}. \tag{2.2.3}$$

Each point contributes its squared distance from the prediction in units of its own standard deviation. An imprecise point has a large σ_i and constrains the fit weakly.

Correlated data are independent data in another basis

Equation (2.2.2) keeps the same meaning when C is not diagonal. A covariance matrix is symmetric and positive definite, so it has a Cholesky factorisation $C = L L^{\mathsf{T}}$ with L lower triangular. Define the whitened residuals $z = L^{-1} r$. Then $C^{-1} = L^{-\mathsf{T}} L^{-1}$ and

$$\chi^2 = r^{\mathsf{T}} L^{-\mathsf{T}} L^{-1} r = \sum_{i=1}^{N_{\mathrm{dat}}} z_i^2, \qquad \text{with covariance} \quad \langle z z^{\mathsf{T}} \rangle = L^{-1} C \, L^{-\mathsf{T}} = 1. \tag{2.2.4}$$

The z_i are independent and have unit variance. A correlated χ^2 is therefore the uncorrelated form (2.2.3) applied to N_{dat} linear combinations of the data. This is also how the χ^2 is computed in practice: C is fixed during a fit, so it is factorised once, and each evaluation costs one triangular solve.

The Gaussian assumption is good when each measurement averages many events. It fails for bins with few counts and for systematic errors with non-Gaussian tails. This is one of the approximations to keep

in mind when reading a fit quality (section 7).

---

3. Building the covariance matrix

The measurement model

Experiments do not publish C. They publish a central value for each point and a breakdown of its uncertainty by source. The covariance matrix follows from a model of how those sources act. For point i,

$$D_i = T_i + \sigma_i \, \varepsilon_i + \sum_{a=1}^{K} \beta_{ia} \, \lambda_a . \tag{2.2.5}$$

Here T_i is the true value. The ε_i and λ_a are independent random numbers with zero mean and unit variance.

- σ_i is the *uncorrelated* error of point i: the statistical error and any systematic error that fluctuates independently from point to point, added in quadrature.
- β_{ia} is the shift of point i when the *correlated* systematic source a moves by one standard deviation. A source is something like a calorimeter energy scale, a detector efficiency or the beam luminosity. One number λ_a, drawn once, moves every point of the data set together, each by its own β_{ia}.

Taking the expectation of $(D_i - T_i)(D_j - T_j)$, the cross terms vanish because the random numbers are independent, and

$$C_{ij} = \sigma_i^2 \, \delta_{ij} + \sum_{a=1}^{K} \beta_{ia} \, \beta_{ja}, \qquad \text{that is,} \qquad C = U + B B^{\mathsf{T}}, \tag{2.2.6}$$

with $U = \operatorname{diag}(\sigma_i^2)$ and B the $N_{\text{dat}} \times K$ matrix of the β_{ia}. The matrix has a diagonal part and a part of rank at most K. The second part records that the points can move together in only K specific patterns.

Correlated sources can make the correlation between points nearly complete. Two points with independent errors of 1% and a shared normalisation uncertainty of 5% have $C_{11} = C_{22} = 26$ and $C_{12} = 25$ in units of $(1\%)^2$, a correlation coefficient of $25/26 = 0.96$.

Additive and multiplicative sources

Sources divide into two kinds according to what β_{ia} is proportional to.

KIND	SIZE OF THE SHIFT	TYPICAL ORIGIN
Uncorrelated	σ_i, fixed	counting statistics, point-to-point systematics
Additive, correlated	β_{ia}, a fixed absolute amount	background subtraction, pedestal
Multiplicative, correlated	$\beta_{ia} = s_{ia} \times$ (value at point i)	luminosity, acceptance, efficiency

An additive source shifts the measurement by an amount that does not depend on what was measured. A multiplicative source rescales it: a luminosity known to 2% means every cross-section in the set is uncertain by 2% of itself, with $s_{ia} = 0.02$ for all i. Ubiali's Eq. (2.3) [2] writes the full matrix with all sources given as relative values multiplied by the measured central values.

The phrase "of itself" is ambiguous. The shift is a fraction of the true value T_i, which is unknown. The two available stand-ins are the measured value D_i and the fitted prediction. The choice looks harmless and is not; it is the subject of section 5. Until then the β_{ia} are fixed numbers.

---

4. Nuisance parameters and the shift form

Fitting the systematic shifts

The model (2.2.5) suggests a second way to write the loss. Treat each λ_a as an extra parameter of the fit, called a *nuisance parameter*. For given λ, the prediction is shifted to $T_i + \sum_a \beta_{ia}\lambda_a$ and the remaining errors are the independent σ_i. Each λ_a is itself a unit Gaussian variable, which adds a penalty λ_a^2:

$$\chi^2(\theta, \lambda) = \sum_{i=1}^{N_{\text{dat}}} \frac{\left(r_i - \sum_a \beta_{ia}\lambda_a\right)^2}{\sigma_i^2} + \sum_{a=1}^{K} \lambda_a^2. \tag{2.2.7}$$

This is the *shift form*. It is quadratic in λ, so the minimum over λ at fixed θ is found by solving a linear system (derivation D1):

$$\hat{\lambda} = A^{-1} B^{\mathsf{T}} U^{-1} r, \qquad A = 1_K + B^{\mathsf{T}} U^{-1} B. \tag{2.2.8}$$

A is a $K \times K$ matrix. Substituting $\hat{\lambda}$ back gives the *profiled* χ^2, a function of θ alone:

$$\chi^2(\theta, \hat{\lambda}) = r^{\mathsf{T}} U^{-1} r - r^{\mathsf{T}} U^{-1} B\, A^{-1} B^{\mathsf{T}} U^{-1} r. \tag{2.2.9}$$

The equivalence

The right side of Eq. (2.2.9) is $r^{\mathsf{T}} M r$ with $M = U^{-1} - U^{-1} B A^{-1} B^{\mathsf{T}} U^{-1}$. The Woodbury identity (derivation D2) states that M is the inverse of the covariance matrix (2.2.6):

$$\left(U + BB^{\mathsf{T}}\right)^{-1} = U^{-1} - U^{-1}B\left(1_K + B^{\mathsf{T}}U^{-1}B\right)^{-1}B^{\mathsf{T}}U^{-1}. \tag{2.2.10}$$

Therefore

$$\min_{\lambda} \chi^2(\theta, \lambda) = r^{\mathsf{T}}\left(U + BB^{\mathsf{T}}\right)^{-1} r.$$

The shift form, profiled over its nuisance parameters, is the covariance-matrix form (2.2.2) exactly, for every θ. For a single source with shifts β_i the identity is the Sherman–Morrison formula and Eq. (2.2.9) reads

$$\chi^2 = \sum_i \frac{r_i^2}{\sigma_i^2} - \frac{\left(\sum_i \beta_i r_i/\sigma_i^2\right)^2}{1 + \sum_i \beta_i^2/\sigma_i^2}, \qquad \hat{\lambda} = \frac{\sum_i \beta_i r_i/\sigma_i^2}{1 + \sum_i \beta_i^2/\sigma_i^2}. \tag{2.2.11}$$

Integrating over λ, instead of minimising, gives the same result, including the normalisation (derivation D2):

$$\int d^K\lambda \, \frac{e^{-\chi^2(\theta,\lambda)/2}}{\sqrt{(2\pi)^{N_{\mathrm{dat}}+K}\det U}} = \frac{e^{-\,r^{\mathsf{T}}C^{-1}r/2}}{\sqrt{(2\pi)^{N_{\mathrm{dat}}}\det C}}, \qquad \det C = \det U \, \det A. \tag{2.2.12}$$

The left side is the joint probability of data and shifts with the shifts integrated out. The right side is the likelihood (2.2.1). A Bayesian, who marginalises nuisance parameters, and a frequentist, who profiles them, therefore use the same function of θ. This holds because the shifts enter linearly and have Gaussian priors.

> **Two forms, one function.** The covariance-matrix χ^2 and the shift-form χ^2 with fitted systematic shifts are the same function of the PDF parameters. The covariance form needs no extra parameters. The shift form shows what the fit is doing: it moves each systematic source by $\hat{\lambda}_a$ standard deviations, at a cost $\hat{\lambda}_a^2$, to bring theory and data together.

What a correlated source does to a fit

Take five points with independent errors of 1% and one shared source of 5%. Suppose the prediction lies 4% below all five. In units of 1%, $r_i = 4$, $\sigma_i = 1$, $\beta_i = 5$. Without the correlation the χ^2 would be $5 \times 4^2 = 80$. With it, Eq. (2.2.11) gives $\sum \beta_i r_i/\sigma_i^2 = 100$ and $1 + \sum \beta_i^2/\sigma_i^2 = 126$, so

$$\chi^2 = 80 - \frac{100^2}{126} = 0.63, \qquad \hat{\lambda} = \frac{100}{126} = 0.79.$$

The fit moves the systematic source by 0.79 of a standard deviation, which shifts every prediction up by 3.97%, and pays 0.63 in total. Now let the residuals alternate in sign, $r = (4, -4, 4, -4, 4)$. Then

$\sum \beta_i r_i/\sigma_i^2 = 20$, the source can absorb only the small net offset, and $\chi^2 = 80 - 400/126 = 76.8$.

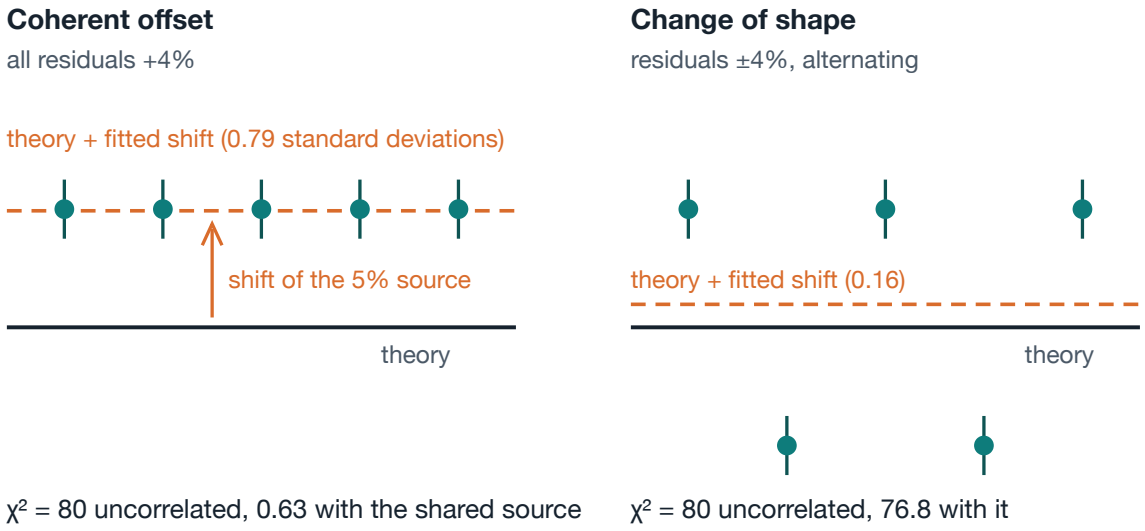

Figure 1. The worked example of the text: five points with independent errors of 1% (bars) and one shared systematic source of 5%. Left: a common offset of four independent standard deviations is absorbed by moving the shared source, at a total cost of 0.63. Right: the same residuals with alternating signs cannot be absorbed, and the cost stays at 76.8. A correlated χ^2 is tolerant of offsets along the directions β_{ia} and strict about everything else.

Three practical consequences follow.

- **A fit can lie visibly off a data set and still describe it well.** A plot of data against theory with uncorrelated error bars cannot show this. The comparison that reflects the χ^2 is data against shifted theory, $T_i + \sum_a \beta_{ia}\hat{\lambda}_a$.
- **The fitted shifts are a diagnostic.** For a correct model the $\hat{\lambda}_a$ scatter about zero with variance below one. Shifts of two or three standard deviations in several sources say that the fit is using the systematics to absorb a real disagreement.
- **The correlation model matters as much as the sizes of the errors.** Equation (2.2.5) assumes each source is fully correlated across all points. If a source in fact acts differently in different kinematic regions, the fit is denied a freedom it should have, and a good description becomes a poor one (section 8).

---

5. Multiplicative errors and the d'Agostini bias

The set-up

Consider the simplest fit: one quantity T, measured n times with results D_i, independent errors σ_i, and one common normalisation uncertainty of fractional size s. Taking the multiplicative shift proportional to the measured values, $\beta_i = s\,D_i$, the covariance matrix (2.2.6) is

$$C_{ij} = \sigma_i^2\,\delta_{ij} + s^2 D_i D_j. \tag{2.2.13}$$

This is the natural reading of a published table with a column “normalisation: 2%”. The fit minimises $\chi^2(T) = \sum_{ij}(T - D_i)(C^{-1})_{ij}(T - D_j)$.

The result

The minimum can be found in closed form for any n (derivation D3). Let w be the average of the data weighted by the independent errors, and X the scatter of the data about it:

$$w = \frac{\sum_i D_i/\sigma_i^2}{\sum_i 1/\sigma_i^2}, \qquad X = \sum_{i=1}^{n} \frac{(D_i - w)^2}{\sigma_i^2}.$$

Then the fitted value is

$$\hat{T} = \frac{w}{1 + s^2 X}. \tag{2.2.14}$$

For two points, $X = (D_1 - D_2)^2/(\sigma_1^2 + \sigma_2^2)$ and

$$\hat{T} = \frac{w}{1 + s^2\,(D_1 - D_2)^2/(\sigma_1^2 + \sigma_2^2)}, \tag{2.2.15}$$

which is Ubiali’s Eq. (2.17) [2]. The weighted mean w is the correct answer: a common rescaling of all the data cannot favour one point over another. Equation (2.2.14) says the fit returns less than w whenever the data scatter at all, and by an amount set by the normalisation uncertainty.

The standard illustration is due to d’Agostini [3]. Two measurements, 8.0 and 8.5, each with a 2% independent error, share a 10% normalisation uncertainty. Then $w = 8.235$, $X = 4.59$, $s^2 X = 0.046$, and the fit returns $\hat{T} = 7.87$. The best estimate lies below both measurements.

Why the data are shrunk

The origin of the bias is clear in the shift form. With $\beta_i = s D_i$, Eq. (2.2.7) becomes

$$\chi^2(T, \lambda) = \sum_{i=1}^{n} \frac{\left(f\,D_i - T\right)^2}{\sigma_i^2} + \lambda^2, \qquad f = 1 - s\lambda. \tag{2.2.16}$$

The nuisance parameter rescales the data by a factor f, but leaves the error bars σ_i where they were. Shrinking the data by $f < 1$ shrinks the distances between them by the same factor, so the scatter term falls from X to $f^2 X$. The fit can always lower its χ^2 by shrinking the data towards zero, and the penalty λ^2 only partly resists. At the minimum $f = 1/(1 + s^2 X)$. The model has rewarded a downward rescaling because smaller values come with smaller absolute normalisation errors, $s D_i$, and so look more precise.

It grows with the number of points

If the model is right, the scatter X of n points about their mean is on average $n-1$. Replacing X by its expectation in Eq. (2.2.14),

$$\frac{\langle \hat{T} \rangle}{T} \approx \frac{1}{1+(n-1)\,s^2}. \tag{2.2.17}$$

The independent errors σ_i have dropped out. The bias depends only on the normalisation uncertainty and on the number of points that share it, and it grows linearly with that number. For 100 points with a 2% normalisation the fit is low by 3.8%. For 300 points with 3% it is low by 21%. More data make the result worse, which is the opposite of a statistical error.

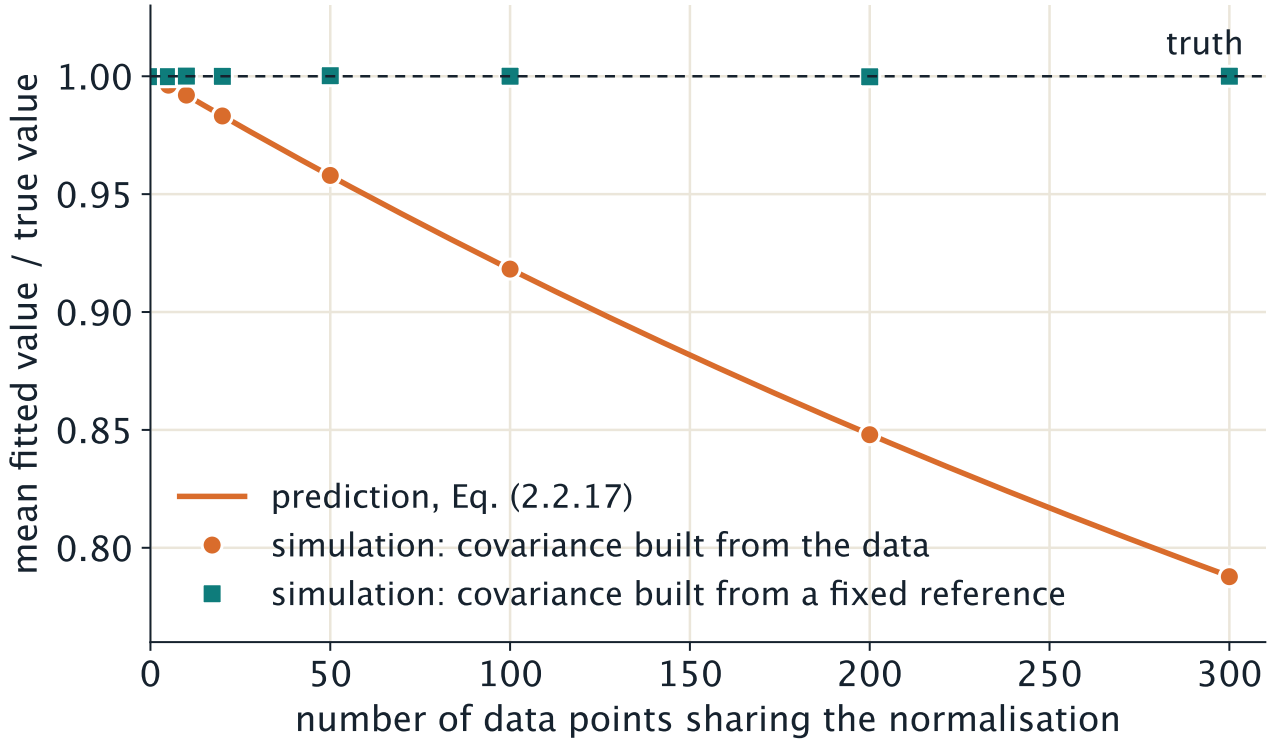

Figure 2. The d'Agostini bias in a simulation. Each pseudo-experiment draws n measurements of a quantity with true value 1, with independent errors of 5% and one common normalisation error of 3%, and fits a constant. Orange points: the covariance matrix is built from the measured values, Eq. (2.2.13); the mean of 40,000 fits follows Eq. (2.2.17), shown as the curve. Teal squares: the covariance is built from a fixed reference value T_0, Eq. (2.2.18); the fit is unbiased for every n. The toy is illustrative, not a fit to real data.

In likelihood terms, the fault is that the covariance (2.2.13) depends on the data. The function $\exp(-\chi^2/2)$ is then not a Gaussian in D and is not a normalised probability density for the data, so minimising the χ^2 is not maximising a likelihood [2]. The condition stated under Eq. (2.2.2) has been violated.

6. The t0 prescription

A covariance that does not move

A normalisation error is a fraction of the true value. The remedy is to build the multiplicative part of the covariance from a fixed estimate of the truth, T_0, chosen before the fit and held constant during it [4]:

$$C_{ij}^{(t_0)} = \sigma_i^2\,\delta_{ij} + s^2\,T_0^2. \tag{2.2.18}$$

The correlated term is now the same for every pair of points. In the shift form the nuisance parameter moves all predictions by the common amount $sT_0\lambda$ and no longer rescales the data. Minimising (derivation D4) gives

$$\hat{T} = w, \qquad \mathrm{Var}(\hat{T}) = \Sigma^2 + s^2\,T_0^2, \qquad \Sigma^2 = \Big(\sum_i \frac{1}{\sigma_i^2}\Big)^{-1}. \tag{2.2.19}$$

The central value is the weighted mean, whatever T_0 is. The variance is the statistical variance of the mean, Σ^2, plus the normalisation uncertainty on T_0, which is correct when T_0 is close to the fitted value. Applied to d'Agostini's example, the prescription returns 8.235.

Several experiments, and iteration

When n experiments each have their own normalisation uncertainty s_i, the prescription gives a diagonal covariance and [2]

$$\chi^2_{t_0}(T) = \sum_{i=1}^{n} \frac{(T - D_i)^2}{\sigma_i^2 + s_i^2\,T_0^2}, \qquad \hat{T} = \frac{\sum_i D_i/(\sigma_i^2 + s_i^2 T_0^2)}{\sum_i 1/(\sigma_i^2 + s_i^2 T_0^2)}. \tag{2.2.20}$$

Now $\hat{T}$ depends on T_0, because T_0 sets the balance between statistical and normalisation errors and so the relative weights of the experiments. The dependence is weak: T_0 enters only through the error sizes, so an error in T_0 affects $\hat{T}$ at second order, and it cancels entirely when the experiments agree or have equal errors. This makes iteration effective. Fit with a first guess for T_0, use the result as the new T_0, and repeat.

The check script runs a deliberately harsh case: three experiments with results 1.10, 0.95, 1.02, independent errors of 0.02, 0.05, 0.03 and normalisation uncertainties of 10%, 2% and 5%. Starting from $T_0 = 0$, which ignores the normalisations altogether, the iterates are 1.0629, 0.9956, 0.9978, 0.9977. The third is within 10^{-4} of the fixed point. Ubiali reports that two iterations were sufficient for the NNPDF1.2 fit [2].

In a PDF fit

In a real fit there are many observables and the reference is a set of predictions $T_i^{(0)}$, computed from a reference PDF set called the t_0 set. The covariance matrix used in the loss is

$$C_{ij}^{(t_0)} = \sigma_i^2\,\delta_{ij} + \sum_{a\,\in\,\text{add}} \beta_{ia}\beta_{ja} + \sum_{a\,\in\,\text{mult}} s_{ia}\,s_{ja}\,T_i^{(0)}\,T_j^{(0)}. \tag{2.2.21}$$

Uncorrelated and additive errors are taken as published. Each multiplicative source contributes its fractional sizes times the reference predictions, not the data. In the NNPDF2.0 fit, for which the prescription was devised, the terms treated in this way were the normalisation uncertainties, and the reference predictions were updated by iterating the fit [4]. The matrix is computed once, before the fit. The loss function $\chi^2_{t_0} = r^{\mathsf{T}}\big(C^{(t_0)}\big)^{-1} r$ is then a quadratic form with a constant matrix, as Eq. (2.2.2) requires.

Two alternatives are in use and both have a defect that t_0 avoids, shown in Ubiali's thesis [2]. The *penalty trick* fits a normalisation factor for each experiment with a Gaussian penalty. It is unbiased for one experiment and biased when several are combined. A *self-consistent* covariance, in which the fitted prediction replaces the data inside C during the minimisation, behaves in the same way.

In this project the t_0 set is fixed in the fit configuration: `240701-02-rs-nnpdf40-baseline` for all Phase B fits (Phase B ledger, fixed setup). One consequence deserves care. The covariance matrix defines the likelihood. Two fits can be compared by their χ^2, or by any quantity derived from the likelihood, only if they used the same t_0 set.

---

7. Goodness of fit

The expected value

Suppose the model is correct, with true parameters θ^*, and the covariance matrix is right. Then by Eq. (2.2.4) the χ^2 at θ^* is a sum of N_{dat} squared independent unit Gaussian variables. Its mean and variance are (derivation D5)

$$\big\langle \chi^2(\theta^*) \big\rangle = N_{\text{dat}}, \qquad \operatorname{Var}\chi^2(\theta^*) = 2N_{\text{dat}}. \tag{2.2.22}$$

Per data point,

$$\frac{\chi^2}{N_{\text{dat}}} = 1 \pm \sqrt{\frac{2}{N_{\text{dat}}}}. \tag{2.2.23}$$

The spread is 0.20 for 50 points, 0.082 for 300 and 0.0208 for 4,615. For large N_{dat} the distribution is close to Gaussian, so the deviation of χ^2/N_{dat} from one, in units of $\sqrt{2/N_{\text{dat}}}$, can be read as a number of standard deviations.

A fit does not evaluate the χ^2 at the truth. It minimises over p parameters, the number of free parameters of the fit, and the minimum is lower. If the predictions are linear in the parameters near the minimum (derivation D5),

$$\big\langle \chi^2_{\text{min}} \big\rangle = N_{\text{dat}} - p, \qquad \operatorname{Var}\chi^2_{\text{min}} = 2\,(N_{\text{dat}} - p). \tag{2.2.24}$$

Each fitted parameter removes one unit, on average, by fitting one direction of the noise. The combination $N_{\mathrm{dat}} - p$ is the number of degrees of freedom. For 4,615 points and 52 parameters the correction is small: the expected $\chi^2_{\mathrm{min}}/N_{\mathrm{dat}}$ is 0.9887.

A calibration: closure tests

A closure test generates artificial data from a known set of PDFs, with noise drawn from the experimental covariance matrix, and fits them. The model and the covariance are then correct by construction, and Eq. (2.2.22) can be tested. In this project's Phase B closure test on the 19 deep inelastic data sets, the χ^2 of the pseudo-data about the truth is 3087.661 for $N_{\mathrm{dat}} = 3089$, a ratio of 0.9996 (Phase B ledger, PPDF-33). The expected spread is $\sqrt{2/3089} = 0.0254$, so this is 0.02 standard deviations from one. With no noise added, the same test gives $\chi^2 = 0.000$ at the truth. The machinery reproduces Eq. (2.2.22). Chapter 2.3 treats closure tests in full.

Real data

On real data the numbers are different. This project's fit to the full global data set, 82 data sets and 4,615 points, has $\chi^2 = 5713.228$ and

$$\frac{\chi^2}{N_{\mathrm{dat}}} = 1.2380$$

(Phase B ledger, PPDF-35). An earlier run of the same fit on a machine with half as many cores gave 5707.67, that is 1.2368 (PPDF-31), the value quoted in chapters 2.1 and 2.4. The ledger attaches a reproducibility caveat of $\pm 0.7\%$ across hardware to both. The same ledger records the published NNPDF4.0 value as 1.17, obtained with a neural network of about 763 parameters against our 52.

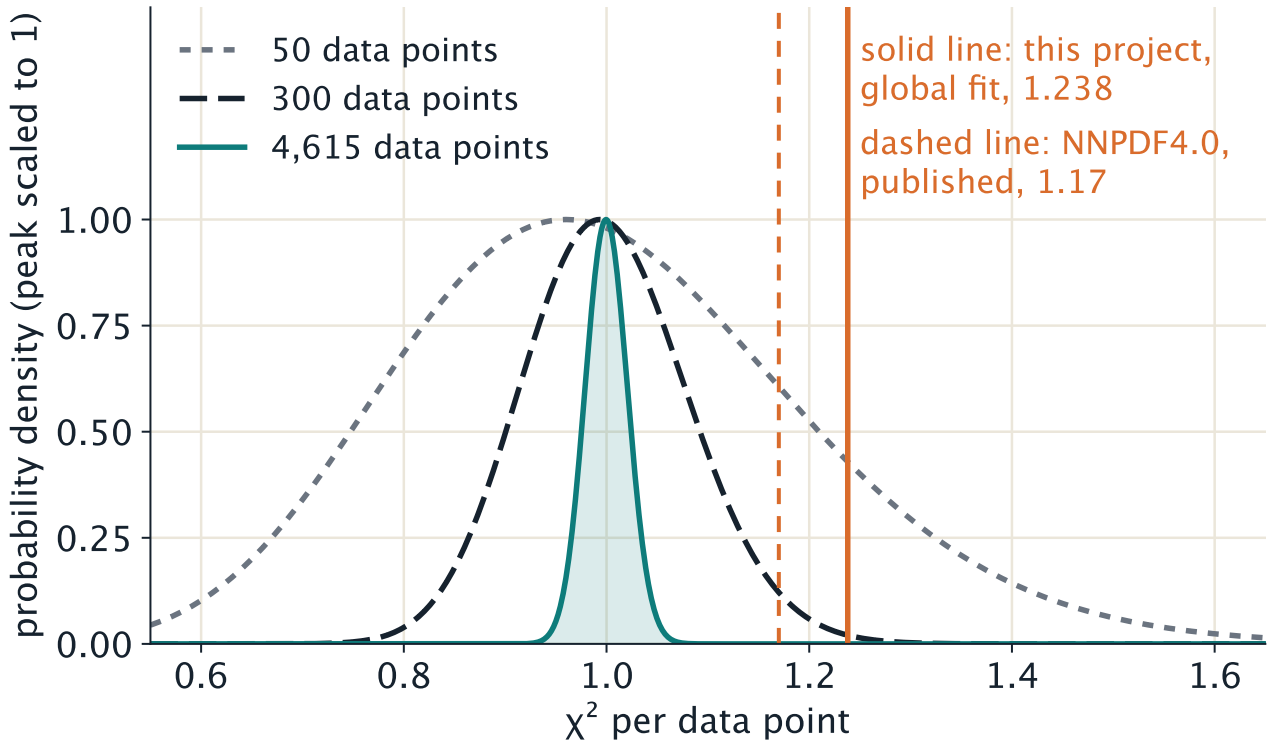

Figure 3. The distribution of χ^2/N_{dat} expected for a correct model and a correct covariance matrix, for three sizes of data set. For 4,615 points it is a narrow peak of width 0.021. The vertical lines are the values of two global fits to real data: this project's (1.238, Phase B ledger) and the published NNPDF4.0 value as recorded there (1.17). Both are far outside the peak. A value of 1.2 would be unremarkable for 50 points.

Measured against Eq. (2.2.23), 1.2380 is $(1.2380 - 1)/0.0208 = 11.4$ standard deviations high. The probability of a χ^2 this large from a correct model is about 10^{-26}. Taken at face value, the hypothesis "this model, with this covariance matrix, generated these data" is excluded, for our fit and for every published global fit. The χ^2/N_{dat} of 1.2 that looks close to one is, over thousands of points, a decisive failure of the formal test.

What a value of 1.2 means

The hypothesis under test is a conjunction, and the test does not say which part fails. There are four candidates.

- **Underestimated or mis-correlated experimental errors.** If every error were too small by a common factor, that factor would be $\sqrt{1.238} = 1.113$: errors 11% larger would restore $\chi^2/N_{\text{dat}} = 1$. Systematic errors are estimates, and their correlations are harder to determine than their sizes.
- **Tension between data sets**, where two measurements pull a distribution in incompatible directions (section 8).
- **Missing theory uncertainty.** The predictions are computed at a fixed perturbative order, and their own error is absent from C (section 9).
- **An inflexible parametrisation.** Our 52-parameter form gives 1.238, and the ledger records 1.17 for a far more flexible neural network. The two numbers were not obtained with identical settings, so

the comparison is indicative. Even if the whole difference of 0.07 per point were due to flexibility, it would be the smaller part of the excess over one.

The consequence for the rest of Part 2 is this. An uncertainty method that assumes the likelihood is exactly right, and allows parameters to move only until χ^2 rises by one, inherits the failure and produces errors that are too small. The tolerance of the Hessian method (chapters 2.3 and 2.4) is the practical response.

---

8. Per-dataset values and tension

Reading a table of fit qualities

Global fits report χ^2/N_{dat} for each data set separately. Equation (2.2.23) applies to each, with its own number of points. A set of 50 points at 1.3 is 1.5 standard deviations high and unremarkable. A set of 500 points at 1.3 is 4.7 standard deviations high.

The per-dataset values add up to the total only when no systematic source is shared between sets. Where sources are shared, as between measurements made with the same detector, the total contains cross terms.

What a tension costs

Two data sets are in tension when they prefer different values of the same combination of parameters. Reduce each to what it says about that one combination θ: set 1 measures $a_1 \pm e_1$ and set 2 measures $a_2 \pm e_2$. The part of the χ^2 that depends on θ is

$$\chi^2(\theta) = \frac{(\theta - a_1)^2}{e_1^2} + \frac{(\theta - a_2)^2}{e_2^2}.$$

Setting the derivative to zero gives the weighted mean, $\hat\theta = (a_1/e_1^2 + a_2/e_2^2)/(1/e_1^2 + 1/e_2^2)$, so $\hat\theta - a_1 = -(a_1 - a_2)\, e_1^2/(e_1^2 + e_2^2)$ and likewise for set 2 with the sign reversed. Substituting,

$$\chi^2_{\mathrm{min}} = \frac{(a_1 - a_2)^2}{e_1^2 + e_2^2}, \qquad \text{of which set 1 pays the fraction } \frac{e_1^2}{e_1^2 + e_2^2}. \tag{2.2.25}$$

The cost of a compromise is the squared difference of the two preferences in units of the combined error. The more precise set pays less, because the compromise lies closer to it.

This result explains why tension is hard to see in fit qualities. A disagreement of three standard deviations adds 9 to the total χ^2, whatever the sizes of the data sets. Spread over two sets of 150 points, that is 0.03 per point, well inside the expected spread of 0.082 for 300 points. A serious disagreement about one direction in parameter space is invisible in χ^2/N_{dat}, globally and per set. It is found by other means: by fitting with and without a data set and comparing, or by scanning the χ^2 of each set along each eigenvector direction, which is how the dynamic tolerance of MSHT is set (chapter 2.4).

Tension from the correlation model

The excess in χ^2 is often a statement about systematics. The Snowmass whitepaper [5] describes the case of the ATLAS 7 TeV inclusive jet data. Fits to each rapidity bin separately were good. The fit to all bins together was very poor, because the published correlations of the systematic sources across rapidity did not allow the shifts that the separate bins preferred. The responses were to fit a single rapidity bin, or to loosen the correlation of selected sources between bins. The whitepaper notes that such decorrelation models are somewhat ad hoc, and recommends that experiments publish each source separately so that its correlation can be revisited.

---

9. Theory uncertainties in the covariance matrix

The predictions $T_i(\theta)$ are not exact. They are truncated at a fixed order in α_S, and for data taken on nuclear targets they rely on a model of the nucleus. The loss function of section 2 treats them as exact, which overweights data whose predictions are least reliable.

The remedy uses section 4 in reverse. Suppose the true prediction differs from the computed one by a shift Δ_i, and that the shifts are Gaussian with zero mean and covariance $C^{\text{th}}_{ij} = \langle \Delta_i \Delta_j \rangle$, independent of the experimental errors. The shifts are nuisance parameters with a Gaussian prior, entering linearly. Integrating them out, as in Eq. (2.2.12), adds their covariance to the experimental one:

$$\chi^2 = \big(D - T\big)^{\mathsf{T}} \big(C + C^{\text{th}}\big)^{-1} \big(D - T\big). \tag{2.2.26}$$

This is the *theory covariance matrix* formalism [5, 7]. Its content lies in the choice of Cth, which is a prior on the size and correlation of the theory error. The whitepaper describes two applications.

- **Nuclear uncertainties.** For data on heavy nuclei and on deuterium, C^{th} is determined empirically from fits to nuclear data. This was the first use of the method, and it is part of the NNPDF4.0 fit.
- **Missing higher orders.** The prior is constructed from variations of the renormalisation and factorisation scales, in schemes with 5, 7 or 9 scale combinations, with correlations between processes built in. Applied to the NLO NNPDF3.1 fit, the overall uncertainties grew only a little, while central values moved towards the NNLO result, because data with large theory uncertainty were given less weight.

Two points remain open. The theory error in the PDFs may be correlated with the theory error in the cross-section they are later used to predict [8]. And estimates of missing higher orders that do not rely on scale variation generally give non-Gaussian priors, which do not fit into Eq. (2.2.26) [5].

---

10. Penalty terms and the complete loss

Positivity and integrability

Two physical requirements on the distributions are imposed by adding terms to the loss. *Positivity*: cross-sections are probabilities and cannot be negative, and the NNPDF methodology takes the $\overline{\text{MS}}$

distributions themselves to be positive, on the basis of Ref. [9]. *Integrability*: the combinations that enter the sum rules of chapter 2.1 must have finite integrals at small x.

A fixed functional form can satisfy these by construction in part. A flexible form cannot, and they are enforced with penalties. Moore's Eq. (1.104) [1], following NNPDF4.0 [6], gives the positivity term:

$$\begin{aligned} \chi^2_{\text{pos}}(\theta) &= \chi^2_{t_0}(\theta) + \sum_q \Lambda_q \sum_{k=1}^{N_{\text{grid}}} \text{ELU}_\alpha\big(-f_q(x_k,\, 5\ \text{GeV}^2)\big), \\ \text{ELU}_\alpha(t) &= \begin{cases} t, & t \geq 0, \\ \alpha\,(e^t - 1), & t < 0. \end{cases} \end{aligned} \tag{2.2.27}$$

The sum runs over the fitted distributions f_q and over a grid of points x_k. Where f_q is negative the argument is positive and the term grows linearly with the violation. Where f_q is positive the term is a small negative constant, at most α in magnitude, and exerts no force. The scale of $5\ \text{GeV}^2$ lies above the charm mass, and evolution preserves positivity at higher scales [1]. The multipliers Λ_q and the parameter α are hyperparameters; in NNPDF the Λ_q are increased as the minimisation proceeds. Further terms of the same form act on a set of pseudo-observables, the positivity data sets, because positive distributions do not guarantee positive cross-sections. Integrability penalties have the same structure.

The complete loss is then

$$\mathcal{E}(\theta) = \underbrace{\big(D - T(\theta)\big)^{\mathsf{T}} \big(C^{(t_0)}\big)^{-1} \big(D - T(\theta)\big)}_{\text{data}} + \underbrace{\mathcal{P}_{\text{pos}}(\theta) + \mathcal{P}_{\text{int}}(\theta)}_{\text{penalties}}. \tag{2.2.28}$$

The two parts have different standing. The first is $-2\ln\mathcal{L}$ up to a constant and has the statistical properties of section 7. The penalties are not a statement about the data. In Bayesian terms they are a prior, $-2\ln$ of a probability assigned to θ before the data are seen. Fit qualities are quoted for the first part alone.

In this project

The fits use NNPDF's full set of positivity and integrability constraints, recorded as 17 + 2 in the fixed setup of the Phase B ledger. Three results from the ledger show how the two parts of Eq. (2.2.28) interact.

FIT	DATA SETS	POINTS	FIT QUALITY PER POINT
Global, penalties on (PPDF-35)	82	4,615	1.2380
Global, penalties on (PPDF-31)	82	4,615	1.2368
Global, penalties off (PPDF-31)	82	4,615	16.35
Deep inelastic only, penalties on (PPDF-31)	19	3,089	1.1505

- **The penalties are needed for the fit to converge.** With positivity and integrability removed, the same model on the same data reached $\chi^2/N_{\text{dat}} = 16.35$, and the fit to deep inelastic data alone did not return a finite value. The ledger's reading is that the penalty block acts as a regulariser: without it the 52-parameter form reaches regions where the likelihood is not finite. The two kinds of penalty were switched off together, so the effect cannot be assigned to positivity alone.
- **At the best fit the penalty is negligible.** For the penalised fit to the deep inelastic data, the data term is 3530.6 and the penalty 8.9, a ratio of 0.0025. The constraints shape the path to the minimum and the region explored, and contribute almost nothing at the minimum itself.
- **Away from the best fit it can dominate.** At the published MSHT20 parameter values the same penalty is 70,729.1. In the closure test planted at that point, this was 23 times the data term of 3,078.65. A parameter point that satisfies one collaboration's constraints can be heavily penalised by another's.

The fit quality of this project, $\chi^2/N_{\text{dat}} = 1.2380$ on 4,615 points, is therefore the value of the first term of Eq. (2.2.28), with the t_0 covariance of Eq. (2.2.21), at the minimum of the whole expression. Chapter 2.3 takes this function and asks how far the parameters can move away from that minimum.

Derivations

D1. Profiling the nuisance parameters

Derives (2.2.8), (2.2.9), (2.2.11)

The stationarity condition. Write Eq. (2.2.7) in matrix form, $\chi^2(\theta, \lambda) = (r - B\lambda)^{\mathsf{T}} U^{-1} (r - B\lambda) + \lambda^{\mathsf{T}}\lambda$. Differentiate with respect to λ and set the result to zero:

$$-2B^{\mathsf{T}}U^{-1}(r - B\lambda) + 2\lambda = 0 \quad \Longrightarrow \quad \left(1_K + B^{\mathsf{T}}U^{-1}B\right)\hat{\lambda} = B^{\mathsf{T}}U^{-1}r.$$

With $A = 1_K + B^{\mathsf{T}}U^{-1}B$ this is Eq. (2.2.8). A is the unit matrix plus a positive semi-definite matrix, so it is invertible and the stationary point is a minimum.

The value at the minimum. Expand the χ^2:

$$\chi^2(\theta, \lambda) = r^{\mathsf{T}}U^{-1}r - 2\lambda^{\mathsf{T}}B^{\mathsf{T}}U^{-1}r + \lambda^{\mathsf{T}}A\,\lambda.$$

At $\lambda = \hat{\lambda}$ the stationarity condition gives $A\hat{\lambda} = B^{\mathsf{T}}U^{-1}r$, so the last term equals $\hat{\lambda}^{\mathsf{T}}B^{\mathsf{T}}U^{-1}r$ and cancels half of the middle term:

$$\chi^2(\theta, \hat{\lambda}) = r^{\mathsf{T}}U^{-1}r - \hat{\lambda}^{\mathsf{T}}B^{\mathsf{T}}U^{-1}r = r^{\mathsf{T}}U^{-1}r - r^{\mathsf{T}}U^{-1}BA^{-1}B^{\mathsf{T}}U^{-1}r,$$

which is Eq. (2.2.9).

One source. For $K = 1$, B is a single column β, A is the number $1 + \sum_i \beta_i^2/\sigma_i^2$, and $B^\mathsf{T} U^{-1} r = \sum_i \beta_i r_i/\sigma_i^2$. This gives Eq. (2.2.11).

The spread of a fitted shift. For one source and a correct model, $r = \sigma\varepsilon + \beta\lambda$ has covariance C, so $\mathrm{Var}(\hat\lambda) = \beta^\mathsf{T} U^{-1} C\, U^{-1}\beta/A^2 = (A-1)A/A^2 = 1 - 1/A$, which is below one.

D2. The Woodbury identity, and integrating out the shifts

Derives (2.2.10), (2.2.12), (2.2.26)

The identity. Multiply the covariance matrix by the claimed inverse, $M = U^{-1} - U^{-1}BA^{-1}B^\mathsf{T}U^{-1}$:

$$\big(U + BB^\mathsf{T}\big)M = 1 - BA^{-1}B^\mathsf{T}U^{-1} + BB^\mathsf{T}U^{-1} - B\big(B^\mathsf{T}U^{-1}B\big)A^{-1}B^\mathsf{T}U^{-1}.$$

In the last term $B^\mathsf{T}U^{-1}B = A - 1_K$. Collecting what stands between B and $B^\mathsf{T}U^{-1}$:

$$-A^{-1} + 1_K - (A - 1_K)A^{-1} = -A^{-1} + 1_K - 1_K + A^{-1} = 0.$$

So $(U + BB^\mathsf{T})M = 1$, which is Eq. (2.2.10). The identity trades the inverse of an $N_\mathrm{dat} \times N_\mathrm{dat}$ matrix for the inverse of a $K \times K$ one.

Completing the square. From the expansion in D1 and $A\hat\lambda = B^\mathsf{T}U^{-1}r$,

$$\chi^2(\theta,\lambda) = \chi^2(\theta,\hat\lambda) + (\lambda - \hat\lambda)^\mathsf{T} A\,(\lambda - \hat\lambda).$$

The integral. The joint probability density of the data and the shifts is the product of N_dat Gaussians of width σ_i and K unit Gaussians, which is the integrand on the left of Eq. (2.2.12). The λ integral is a Gaussian with inverse covariance A:

$$\int d^K\lambda\; e^{-\frac{1}{2}(\lambda-\hat\lambda)^\mathsf{T} A(\lambda-\hat\lambda)} = \frac{(2\pi)^{K/2}}{\sqrt{\det A}}.$$

The result is $e^{-\chi^2(\theta,\hat\lambda)/2}\big/\sqrt{(2\pi)^{N_\mathrm{dat}}\det U\;\det A}$.

The determinant. $\det C = \det U\;\det\big(1_N + U^{-1}BB^\mathsf{T}\big)$. For any $N \times K$ matrix X and $K \times N$ matrix Y, $\det(1_N + XY) = \det(1_K + YX)$. With $X = U^{-1}B$ and $Y = B^\mathsf{T}$ this gives $\det C = \det U\;\det A$. Together with D1 the integral equals the Gaussian (2.2.1) with covariance C, which is Eq. (2.2.12).

D3. The d'Agostini bias for any number of points

Derives (2.2.14), (2.2.15), (2.2.16), (2.2.17)

The shift form. By section 4 the fit with covariance (2.2.13) is the minimum over T and λ of Eq. (2.2.7) with $r_i = D_i - T$ and $\beta_i = sD_i$. Then $r_i - \beta_i\lambda = (1 - s\lambda)D_i - T$, which is Eq. (2.2.16) with $f = 1 - s\lambda$.

Minimising over T. At fixed f the data are fD_i with errors σ_i. The minimum is at their weighted mean, $T = fw$, and the value there is $\sum_i (fD_i - fw)^2/\sigma_i^2 = f^2 X$. So

$$\min_T \chi^2 = f^2 X + \lambda^2 = (1 - s\lambda)^2 X + \lambda^2.$$

Minimising over λ. Setting the derivative to zero, $-2s(1 - s\lambda)X + 2\lambda = 0$, so

$$\lambda = \frac{sX}{1 + s^2 X}, \qquad f = 1 - s\lambda = \frac{1}{1 + s^2 X}, \qquad \hat{T} = fw = \frac{w}{1 + s^2 X}.$$

This is Eq. (2.2.14). The χ^2 at the minimum is $f^2 X + \lambda^2 = X/(1 + s^2 X)$, below the scatter X that the data really have.

Two points. With $w = (D_1\sigma_2^2 + D_2\sigma_1^2)/(\sigma_1^2 + \sigma_2^2)$, $D_1 - w = (D_1 - D_2)\sigma_1^2/(\sigma_1^2 + \sigma_2^2)$ and $D_2 - w = -(D_1 - D_2)\sigma_2^2/(\sigma_1^2 + \sigma_2^2)$. Then $X = (D_1 - D_2)^2(\sigma_1^2 + \sigma_2^2)/(\sigma_1^2 + \sigma_2^2)^2$, which gives Eq. (2.2.15).

The expectation. Generate data as $D_i = n_0(T + \sigma_i \varepsilon_i)$ with a common factor n_0 of mean one and variance s^2. The scatter is $X = n_0^2 \sum_i (\varepsilon_i - \bar{\varepsilon})^2$ for equal errors, with mean $(1 + s^2)(n - 1)$. To leading order in s^2 this is $n - 1$, and $\langle w \rangle = T$. Replacing X and w in Eq. (2.2.14) by their means gives Eq. (2.2.17). The replacement is an approximation, since the mean of a ratio is not the ratio of means. In the simulation of Figure 2 it holds to better than 10^{-3} for n from 2 to 300.

D4. The to fit is unbiased

Derives (2.2.19), (2.2.20)

The shift form. With covariance (2.2.18) the correlated shift is $\beta_i = sT_0$ for every point, and Eq. (2.2.7) reads

$$\chi^2(T, \lambda) = \sum_{i=1}^{n} \frac{(D_i - u)^2}{\sigma_i^2} + \lambda^2, \qquad u = T + sT_0\lambda.$$

The data are not rescaled. The nuisance parameter moves the prediction, and T and λ enter the first term only through u.

The central value. The first term is minimised at $u = w$, where it equals X. The second is minimised at $\lambda = 0$. Both can be satisfied at once, with $T = w$. So $\hat{T} = w$ and $\chi^2_{\min} = X$, the true scatter of the data.

The variance. Use the identity $\sum_i (D_i - u)^2/\sigma_i^2 = X + (u - w)^2/\Sigma^2$ with $\Sigma^2 = (\sum_i 1/\sigma_i^2)^{-1}$. At fixed T, write $\delta = T - w$ and minimise $(\delta + sT_0\lambda)^2/\Sigma^2 + \lambda^2$ over λ. This is Eq. (2.2.11) for a single point with residual δ, error Σ and shift sT_0:

$$\chi^2_{t_0}(T) = X + \frac{(T - w)^2}{\Sigma^2 + s^2 T_0^2}.$$

The χ^2 is an exact parabola in T. It rises by one when T moves by $\sqrt{\Sigma^2 + s^2 T_0^2}$ from its minimum, which is the variance in Eq. (2.2.19).

Independent normalisations. If each point has its own normalisation, the covariance is diagonal with entries $\sigma_i^2 + s_i^2 T_0^2$, and Eq. (2.2.20) is the weighted mean of Eq. (2.2.3). With all σ_i and s_i equal, the weights are equal and $\hat{T}$ is the plain average, for any T_0.

D5. The mean and variance of the chi-squared

Derives (2.2.22), (2.2.24)

At the truth. By Eq. (2.2.4), $\chi^2(\theta^*) = \sum_i z_i^2$ with independent unit Gaussian z_i. A unit Gaussian has $\langle z^2 \rangle = 1$ and $\langle z^4 \rangle = 3$, so $\mathrm{Var}(z^2) = 3 - 1 = 2$. Means and variances of independent terms add, which gives Eq. (2.2.22).

The linear model. Near the minimum write $T(\theta) = T(\theta^*) + J\,(\theta - \theta^*)$, where J is the $N_{\mathrm{dat}} \times p$ matrix of derivatives $\partial T_i / \partial \theta_\alpha$, assumed to have full rank. In whitened variables, $z = L^{-1}\big(D - T(\theta^*)\big)$ is a vector of independent unit Gaussians and $\tilde{J} = L^{-1} J$, so

$$\chi^2(\theta) = \left| z - \tilde{J}\,(\theta - \theta^*) \right|^2 .$$

The minimum is a projection. Minimising the squared length over θ removes from z its component in the p-dimensional space spanned by the columns of $\tilde{J}$. With $P = \tilde{J}(\tilde{J}^{\mathsf{T}}\tilde{J})^{-1}\tilde{J}^{\mathsf{T}}$ the projector onto that space,

$$\chi^2_{\mathrm{min}} = z^{\mathsf{T}} (1 - P)\, z .$$

Counting. $1 - P$ is a projector onto a space of dimension $N_{\mathrm{dat}} - p$. In an orthonormal basis adapted to that space, z still has independent unit Gaussian components, because a rotation leaves the unit covariance unchanged. So χ^2_{min} is a sum of $N_{\mathrm{dat}} - p$ squared unit Gaussians, and the first step gives Eq. (2.2.24).

The result fails when the predictions are not linear over the range the noise explores, when a parameter has no effect on any prediction, or when a prior or a bound stops a parameter from reaching its unconstrained minimum. Each case removes less than one unit per parameter.

References

1. J. M. Moore, *Parton Distributions in Beyond the Standard Model Theories*, PhD thesis, University of Cambridge, section 1.3.2. The loss function, the positivity penalty (Eq. 1.104) and the motivation for the t_0 prescription.
2. M. Ubiali, *A new approach in the determination of Parton Distribution Functions. Application to processes with heavy quarks in the initial states*, PhD thesis, Université catholique de Louvain

and University of Edinburgh (2010), sections 2.1.3 and 2.2. The covariance matrix, the two-point bias (Eqs. 2.15–2.21), the penalty trick and the t_0 method (Eqs. 2.31–2.44).

3. G. D'Agostini, "On the use of the covariance matrix to fit correlated data", Nucl. Instrum. Meth. A346 (1994) 306.
4. R. D. Ball, L. Del Debbio, S. Forte, A. Guffanti, J. I. Latorre, J. Rojo and M. Ubiali, "Fitting parton distribution data with multiplicative normalization uncertainties", JHEP 05 (2010) 075, arXiv:0912.2276; the form used in the NNPDF2.0 fit is Eq. (1) of R. D. Ball et al. (NNPDF), Nucl. Phys. B838 (2010) 136, arXiv:1002.4407.
5. S. Amoroso et al., "Snowmass 2021 whitepaper: Proton structure at the precision frontier", arXiv:2203.13923. The sections on correlated systematic uncertainties and on theoretical uncertainties in PDF fits.
6. NNPDF Collaboration, R. D. Ball et al., "The path to proton structure at 1% accuracy", Eur. Phys. J. C82 (2022) 428.
7. NNPDF Collaboration, R. Abdul Khalek et al., Eur. Phys. J. C79 (2019) 838 and C79 (2019) 931 (missing higher orders); R. D. Ball, E. R. Nocera and R. L. Pearson, Eur. Phys. J. C79 (2019) 282 (nuclear uncertainties). As cited in [5].
8. L. A. Harland-Lang and R. S. Thorne, Eur. Phys. J. C79 (2019) 225. As cited in [5].
9. A. Candido, S. Forte and F. Hekhorn, "Can $\overline{\text{MS}}$ parton distributions be negative?", JHEP 11 (2020) 129.