Progress on characterising PDF uncertainties

PDF4LHC Meeting, CERN

Zahari Kassabov

November 23, 2022

DAMTP, University of Cambridge

Details

This talk

Progress in LHC phenomenology requires much improved understanding of uncertainties

- Improved understanding of experimental uncertainties
  - *Regularising experimental correlations in LHC data: theory and application to a global analysis of parton distributions* (ZK, Nocera, Wilson, arxiv:2207.00690)
- Improved understanding of theory uncertainties
  - *Parton distributions with scale uncertainties: a MonteCarlo sampling approach* (ZK, Ubiali, Voisey, arxiv:2207.07616)

Covariance matrix regularization

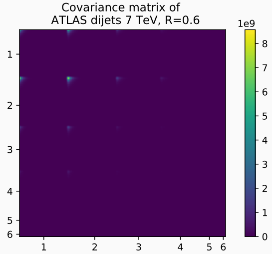
arxiv:2207.00690 studies the problem of covariance matrix stability

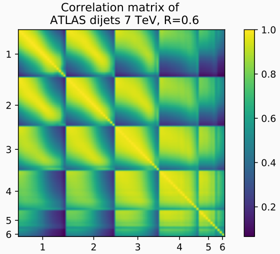
$$\chi^2 = (\text{data} - \text{theory})(\text{inverse covariance matrix})(\text{data} - \text{theory})$$

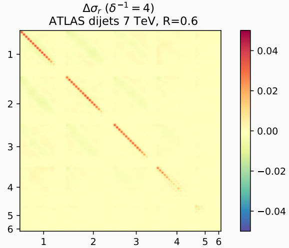
- If the covariance matrix is close to singular, then χ^2 is unstable
  - Small inaccuracies in the matrix make χ^2 appear much larger.
- Large correlations between systematics cause the matrix to be unstable
  - But correlation models usually difficult to determine, hence can be inaccurate

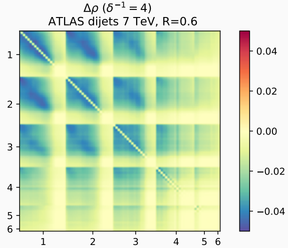
Derive a regularization procedure to make the covariance matrix stable.

Example: ATLAS dijets at 7 TeV









- Original $\chi^2/N_{\text{dat}} = 2.14$
- Regularized $\chi^2/N_{\text{dat}} = 1.10$

Analysis framework

- Consider matrix of uncertainties $N_{\mathrm{dat}} \times N_{\mathrm{err}}$ such that the covariance is AA^t.
- Assuming the theory is known, fixed, and correct

$$\mathbf{d} - \mathbf{t} = A\mathbf{n}\,, \qquad \mathbf{n} \sim \mathcal{N}(\mathbf{0}, I)\,.$$

  Then the expected value of the χ^2 is

$$\langle \chi^2 \rangle = \left\| A^+ A \right\|_F^2 = N_{\mathrm{dat}}\,,$$

- If the χ^2 is measured with a different matrix $\bar{A}$ then the expected value is instead

$$\langle \bar{\chi}^2 \rangle = \left\| \bar{A}^+ A \right\|_F^2$$

Stability

Standard deviation of χ^2 distribution is $\sqrt{2N_{\mathrm{dat}}}$ hence we have stability if

$$\Delta\chi^2 = \left\|\bar{A}^{+}A\right\|_F^2 - N_{\mathrm{dat}} < \sqrt{2N_{\mathrm{dat}}}$$

- No non trivial assumptions so far - Assumptions needed since we don't know A.

A toy model with all the information

$$A(x) = \begin{pmatrix} \epsilon & 0 & 0 & 0 & 1 & 0 \\ 0 & \epsilon & 0 & 0 & 1 & 0 \\ 0 & 0 & \epsilon & 0 & 1 & 0 \\ 0 & 0 & 0 & \epsilon & 1-x & \sqrt{1-(1-x)^2} \end{pmatrix}$$

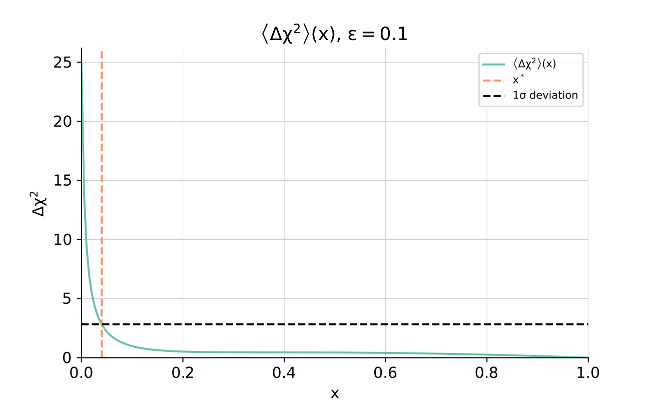
Assume $\epsilon \ll 1$ and $x \in [0,1]$ unknown, sampled from

$$f_x(\xi) = 5(1-\xi)^4$$

- $x = 0$ is the most likely value. But it kills stability! - Measure with

$$\left\langle \Delta\chi^2 \right\rangle (x) = \int_0^1 \left| \left\| \bar{A}^+(\xi) A(x) \right\|_F^2 - N \right| f_x(\xi) \mathrm{d}\xi$$

Why not pick the highest correlation?



- $x = 0$ leads to an expected error of over 8 standard deviations
- $x = 0.04$ reduces the error to 1 standard deviation

Regularization

Assuming that

- All inaccuracies are in correlations $\bar{A} = D\bar{A}_{\mathrm{corr}}$
  - D: matrix of standard deviations
- Inaccuracies come from a small $\mathcal{O}(1)$ number of systematics

The stability of A can be measured by the condition number

$$Z = \left\|\bar{A}^{+}_{\text{corr}}\right\|_2 = \left\|\bar{A}_{\text{corr}}\right\|_2^{-1}$$

Closest matrix to $A = DUSV^t$ with $Z = \delta^{-1}$ for some acceptable δ

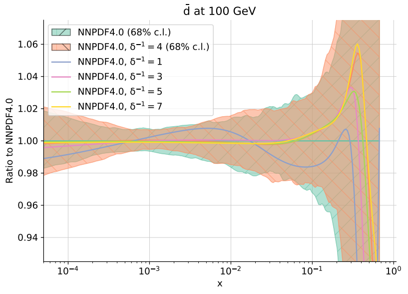
$$\bar{A}_{\mathrm{reg}} = DUS_{\text{reg}}V^t$$

$$S_{\text{reg}(ii)} = \begin{cases} \delta & s_i < \delta \\ s_i & \text{otherwise} \end{cases}$$

Effect on global fit

Applying our preferred regularization to the full NNPDF dataset:

- Relative covariance differences smaller than 5%
- Correlation differences smaller than 0.05
- χ^2/N_{dat}: $1.16 \to 1.11$
- Almost no effect on best fit PDFs



What to do

Experimentalist (target audience)

- Measure stability of covmats
- Provide stable covmats
  - If sources of inaccuracy not known use the regularization procedure
  - If more information and resources available carry out detailed analysis (see Sect 3.2 of arxiv:2207.00690)

Fitters (fallback)

- Measure stability of covmats
- Seek stable versions of the covmat
  - If not available, regularize them
- Discuss regularized χ^2 only

What not to do

Correlation models (so far)

Advantages:

- Made by experimentalists using complete information

Disadvantages:

- Appear much later than the original data, causing versioning confusion
- Enormously laborious to analyse (feedback loop with experimentalists)
- Stability of the covariance matrices not guaranteed

Regularization procedure

Advantages

- Simple quick to apply formula
- Minimal modification of the covariance matrix
- Guaranteed stability
- Seems to yield similar results as correlation models

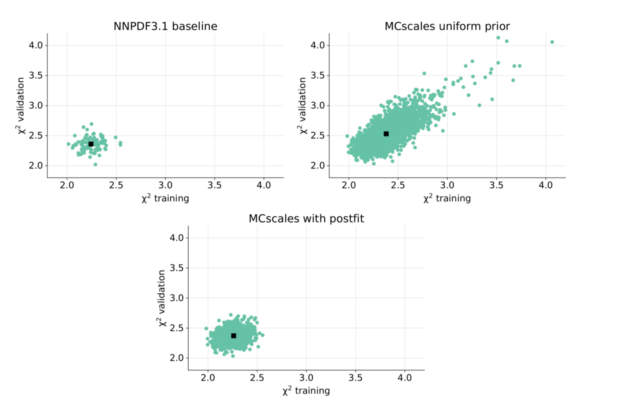
Disadvantages

- Grounded on assumptions and incomplete information

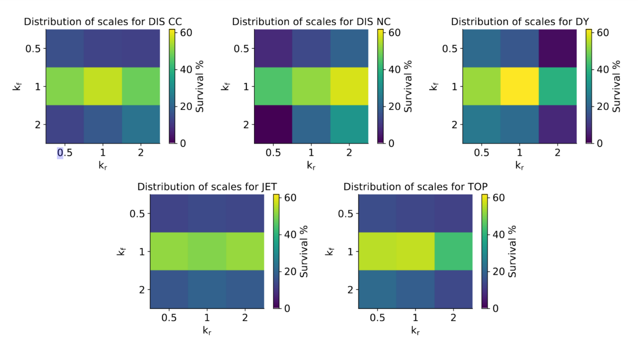
Theoretical uncertainties: MCscales

- arxiv:2207.07616 Studies the problem of matching PDF fits with scale variations
- Theory predictions require specifying factorisation and renormalisation scales:
    - Result depends on scale choice $\longrightarrow$ scale uncertainty.
- Idea: assign different scale multipliers to each NNPDF replica.
- Record the information so scales can be matched between the PDF and the partonic cross section.

Fit quality allows assessing scale choices



Survival fraction



- Statistical interpretation of scale variations
- Assessment of ranges of variation

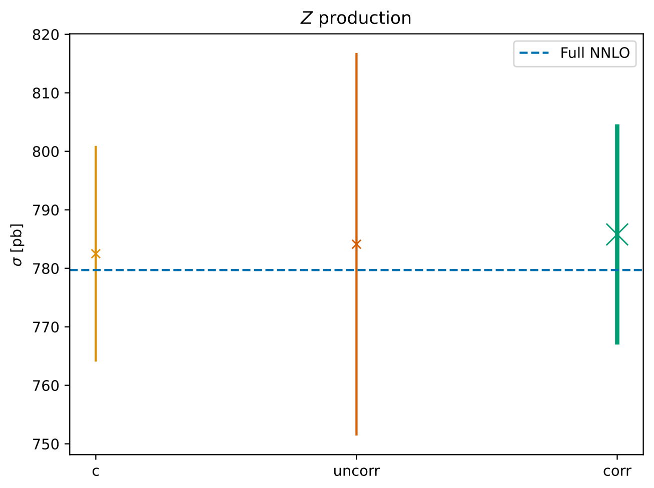
Matched scales convolution

We record the scale multiplier choices for each fitted replica. This allows matching the partonic cross section with the scale choices within each replica

- Monte Carlo sample of N_{rep} `MCscales` prediction including correlated PDF and scale uncertainty

$$\left\{\sigma_k = \hat{\sigma}_p(k_f^{(k)}, k_{r_p}^{(k)}) \otimes f_k(k_f^{(k)}, k_{r_p}^{(k)}) \;\; \forall\, k \in 1 \ldots N_{\text{rep}}\right\}$$

Scales must be matched: Example Z cross section



Treating scales as uncorrelated between PDF and partonic cross section largely overestimates the uncertainties

Why MCscales

- Correlation between scale variations in PDFs and partonic cross sections is large.
    - MCscales allows for exact matching
- Transparent specification of scale uncertainties, with tools allowing users to manipulate it.
- Largest benchmark of effect of scale variations of fit quality.
- NNLO implementation on NNPDF4.0 expected.