Challenges and progress associated to PDF determinations

QCD@LHC 2019, Buffalo

Zahari Kassabov

July 15, 2019

Cavendish Laboratory, University of Cambridge

A common trend in presentations

"The LHC has found no New Physics. Therefore further progress relies on precision studies."

Consequently we have to

- *Improve experimental data.*
- *Improve theory.*

- Any low hanging fruits?

Precision studies

Precision studies typically answer two kinds of question:

Parameter estimation What are the values and errors of the best fitting parameters given a model and a dataset?

Hypothesis testing Does a given dataset allow to reject a given model?

- Note that these two do not commute.
- Significative improvements can be made in the process of comparing data and theory, namely in the statistical treatment.
- Crucially related to PDF determinations.

Improvements in statistical treatment

- Maximize the sensitivity w.r.t. the things we want to understand.
- Minimize it w.r.t. the things we don't understand (*robustness*).
- Consider all relevant degrees of freedom.
- Understand the underlying optimization problem.
- Perform the optimization effectively.
- Consider all relevant sources of uncertainty.

This talk:

- Examples of how improvements in each of these can result in better precision studies.
- Related progress in PDF determination.

Improvements in statistical treatment

- Maximize the sensitivity w.r.t. the things we want to understand.
- Minimize it w.r.t. the things we don't understand (*robustness*).
- Consider all relevant degrees of freedom.
- Understand the underlying optimization problem.
- Perform the optimization effectively.
- Consider all relevant sources of uncertainty.

This talk:

- Examples of how improvements in each of these can result in better precision studies.
- Related progress in PDF determination.

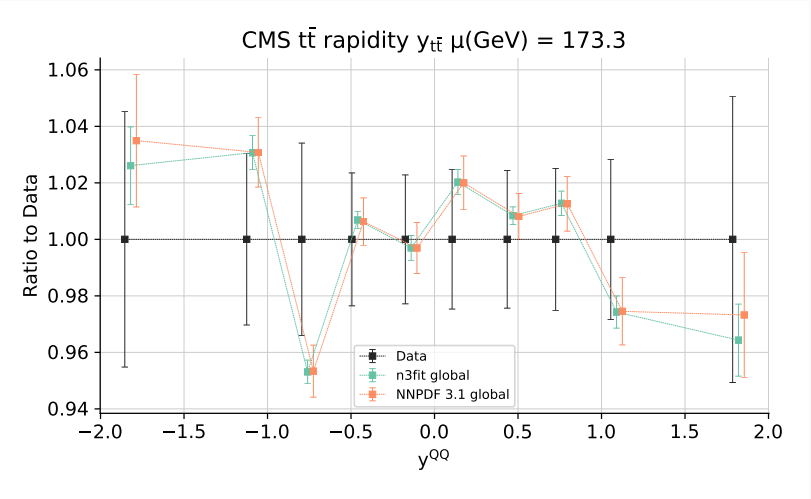
Overall effect is to improve consistency of parameter estimations and decrease sensitivity to new models. But can inform on the best way forward.

Unstable covariance matrices: Introduction

Can compare data and theory by looking at plots

But:

- Need a quantitative measure.
- Insensitive to correlations.



The χ^2 statistic is typically used

$$\chi^2 = \sum_i^N \sum_j^N (\text{data}_i - \text{prediction}_i)\Sigma_{ij}^{-1}(\text{data}_j - \text{prediction}_j) = \delta^T \Sigma^{-1} \delta$$

- Predictions supplied by the theoretical model.
- Central measurement of data and covariance matrix Σ supplied by experiments.
- Expected value (under suitable assumptions): $\langle \chi^2 \rangle = N$. Larger values indicate disagreement between data and theory.

Problems with covariance matrices

- Datasets with problematic correlation models have become common recently
  - ATLAS Jets at 7 TeV (arxiv: 1410.8857): Enormous sensitivity to correlations studied in detail in [Harland-Lang, Martin, Thorne arxiv:1711.05757].

	Full	21	62	21,62
$\chi^2/N_{\text{pts.}}$	2.85	1.58	2.36	1.27

Table 1: χ^2 per number of data points ($N_{\text{pts}} = 140$) for fit to ATLAS jets data [23], with the default systematic error treatment ('full') and with certain errors, defined in the text, decorrelated between jet rapidity bins.

  - CMS 8 TeV double-differential Drell-Yan data at 8 TeV (arXiv:1412.1115) had to be discarded from NNPDF 3.1 (arxiv:1706.00428).
  - Similar issues found in *"several"* newer datasets.
- We now study the issue within a toy model. Will showcase a request to experimentalists and a warning to theorists.

Systematics dominated covariance matrices

- Experiments have reached an impressive level of statistical precision.
  - Statistical component of the uncertainty (typically uncorrelated across bins) less important.
  - Systematic uncertainties (correlated across bins) tend to dominate.
- A somewhat realistic toy model for a covariance matrix from HepData:

$$\Sigma \propto \begin{bmatrix} \epsilon^2 + 1 & 1 & 1 & 1 \\ 1 & \epsilon^2 + 1 & 1 & 1 \\ 1 & 1 & \epsilon^2 + 1 & 1 \\ 1 & 1 & 1 & \epsilon^2 + 1 \end{bmatrix}$$

with $\epsilon^2 \ll 1$.

  - Assumes 4 data points, and uncorrelated error of size ϵ and one completely correlated systematic of size 1.

Allowed fluctuations

$$\Sigma \propto \begin{bmatrix} \epsilon^2+1 & 1 & 1 & 1 \\ 1 & \epsilon^2+1 & 1 & 1 \\ 1 & 1 & \epsilon^2+1 & 1 \\ 1 & 1 & 1 & \epsilon^2+1 \end{bmatrix}$$

The eigenvector decomposition is:

$$\left(\lambda_{1,2,3} = \epsilon^2,\ e_{1,2,3} = \begin{bmatrix} -1 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \\ -1 \end{bmatrix}, \begin{bmatrix} -1 \\ -1 \\ 1 \\ 1 \end{bmatrix} \right), \left(\lambda_4 = \epsilon^2 + 4,\ e_4 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \right)$$

$$\chi^2 = \sum_\ell \langle \delta | e_\ell \rangle \frac{\lambda_\ell^{-1}}{\|e_\ell\|^2} \langle e_\ell | \delta \rangle$$

Fluctuations in $\delta = (\text{data} - \text{theory})$ larger than ϵ in the subspace spanned by $e_{1,2,3}$ lead to $\chi^2 \gg N$.

Problem: Uncertainties in the correlations

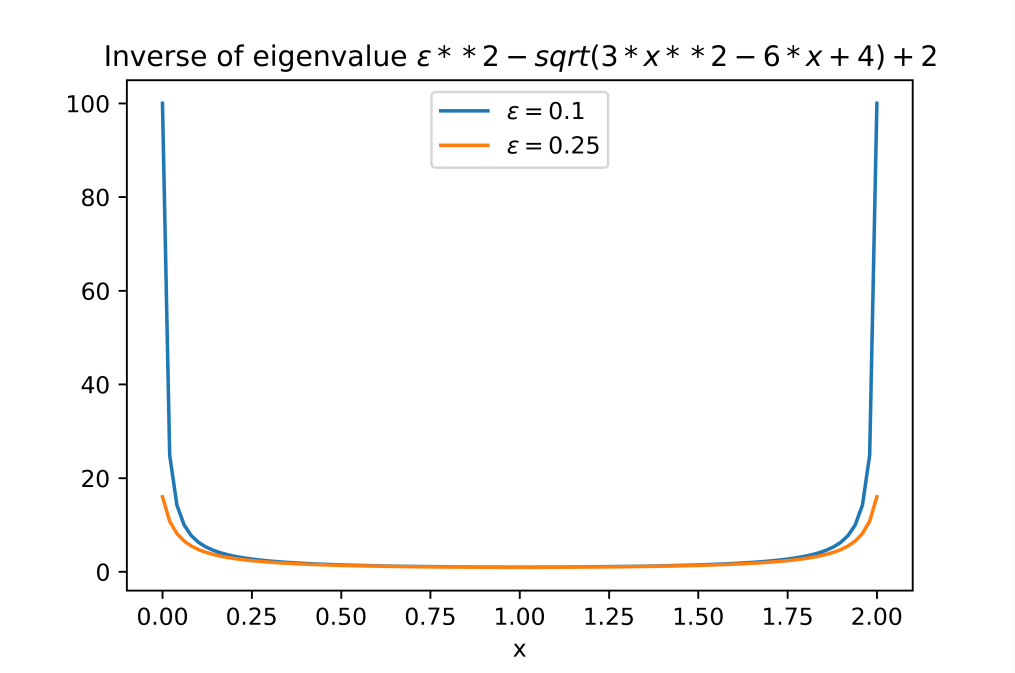
- Well known that exact experimental correlations are hard to determine precisely.
- Model the uncertainty in correlations with unknown parameter $x \in [0, 2]$ controlling the correlations of the last bin.

$$\begin{bmatrix} \epsilon^2 + 1 & 1 & 1 & 1 - x \\ 1 & \epsilon^2 + 1 & 1 & 1 - x \\ 1 & 1 & \epsilon^2 + 1 & 1 - x \\ 1 - x & 1 - x & 1 - x & \epsilon^2 + 1 \end{bmatrix}$$

  - We are keeping the total variance fixed. It is realistic to think that x could be anywhere in the range.
- Experimental results often presented by default with the highest correlation (i.e. $x = 0$).

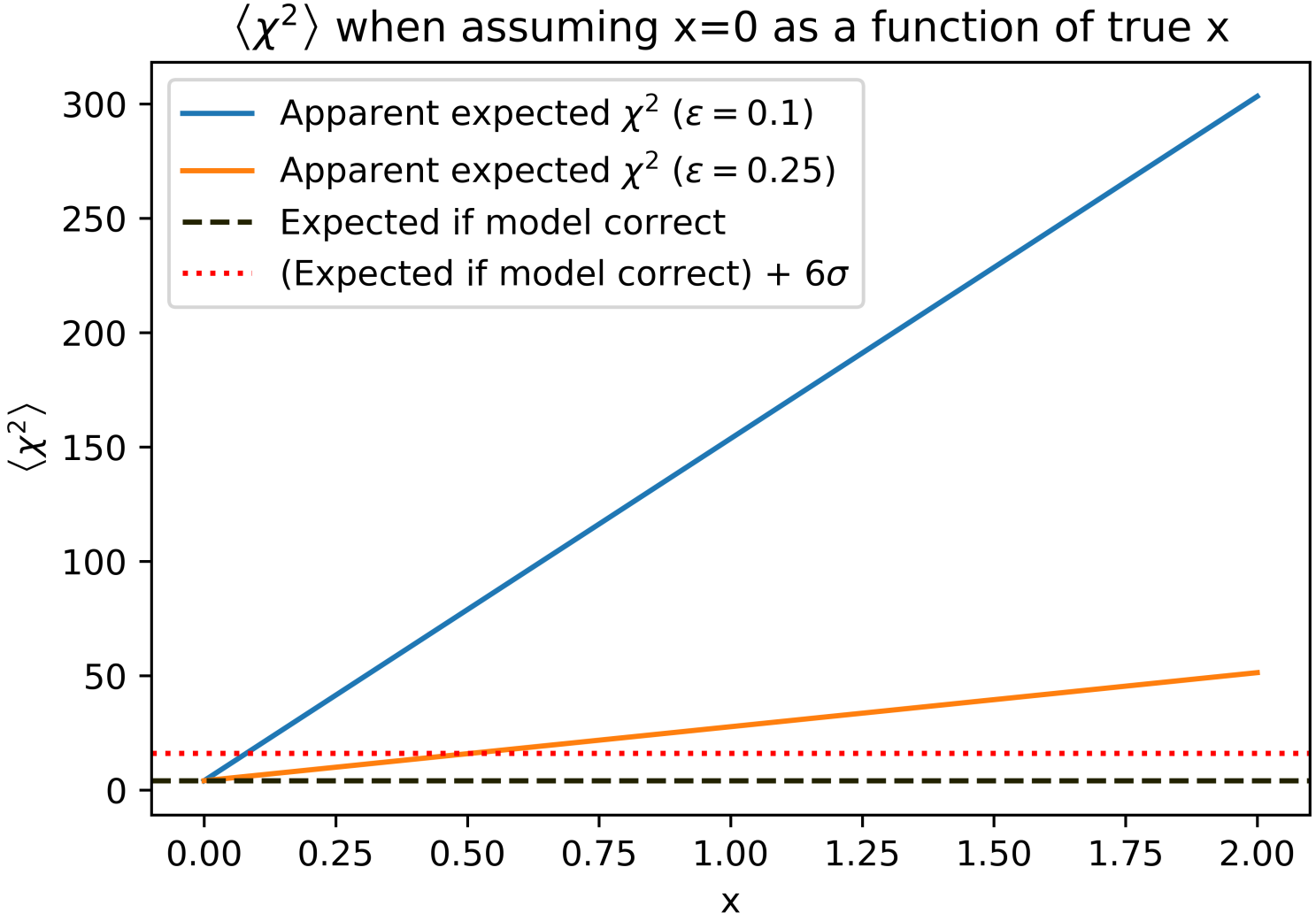
Modified eigenvalue

- Now one eigenvalue depends critically on x.



- The problematic situation is where:
  - Experimental analysis claims high correlations ($x = 0$).
  - Actual correlations are lower (e.g. $x = 0.3$).
- Instability translates directly to the χ^2.

χ^2 with incorrectly predicted correlations



Note: this is all assuming that there are no free parameters to fit.

Choices of correlation models

- A plethora of correlations models (up to 18) was proposed to address issues with ATLAS Jets at 7 TeV, (B. Malescu QCD@LHC'18).
- Also done in other datasets (e.g. ATLAS dijets offer three correlation models)
- This is a positive development but:
  - Pushes the work to test each correlation model downstream.
  - Causes fragmentation among different analyses.
  - Data should inform PDF (and other theory) models; not the other way around.

Other solutions and take away

- Experimental uncertainties should be **robust** w.r.t. unknown parameters, particularly correlations.
- Possible improvements:

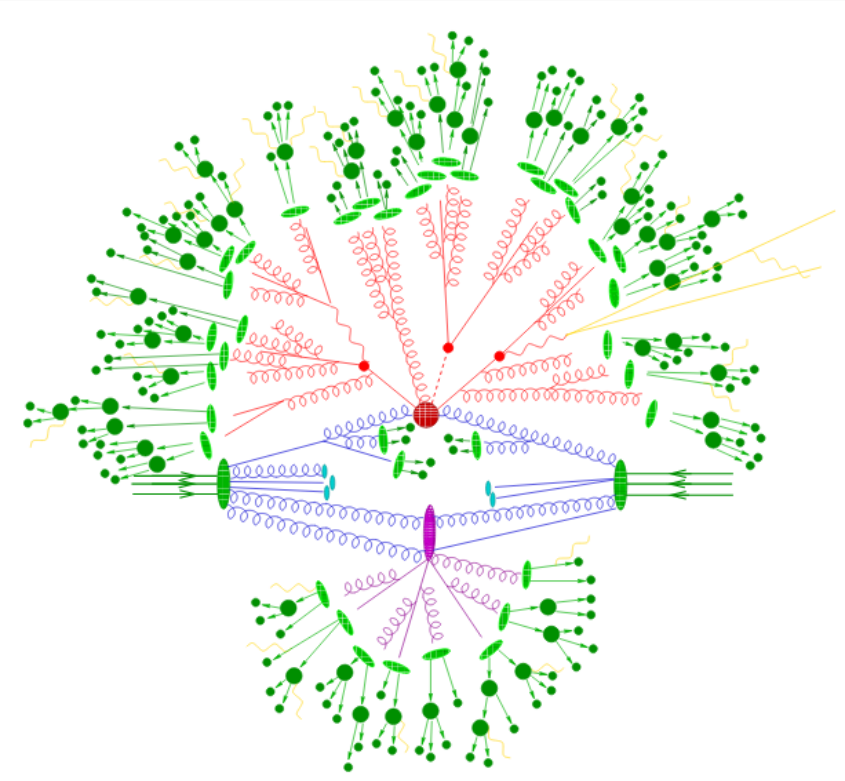
  Decorrelation Assume lower correlations unless proven otherwise. Note that even a small decorrelation (e.g. make $x = 0.3$) makes a big difference.

  Regularization Add a term to Σ that makes Σ^{-1} more stable. E.g. Tikhonov regularization: Add a diagonal term making $\Sigma \to \Sigma + \rho I$

    - Advatage: Best fit theory could be the same.

- Both experimentalist and theorists should be aware of these issues when comparing results.
- Improvements here crucial. Can't do precision physics with unstable precision tests.

On partial fits to partial data



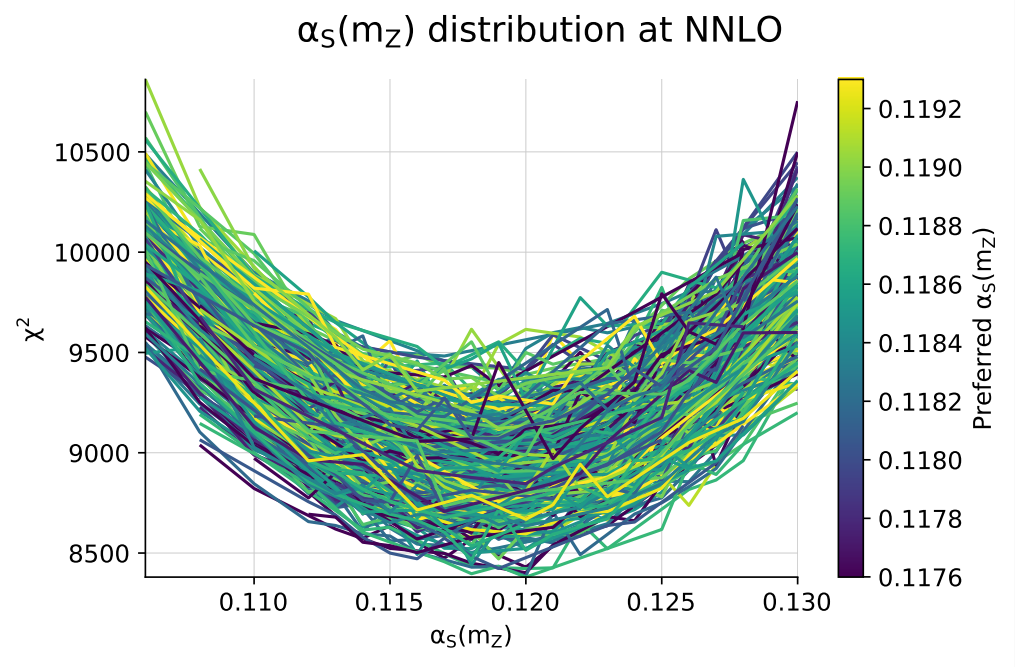
From SHERPA, arxiv:0811.4622

If a given parameter is to be fitted from hadronic data, its best fit value depends on its effect on the full *picture.*

Example: Interaction of α_s and PDF fits.

α_S from PDFs

- α_S Can be determined from global PDF fits. Most recently [NNPDF, arxiv:1802.03398].
- Methodology: Determine simultaneously (α_s, PDF), by minimizing the global χ^2 of the full (NN)PDF dataset (see Z.K., QCD@LHC'18).



- Best attainable total χ^2 changes **strongly** with α_s (leading to a very precise prediction, if theory uncertainties are ignored).

α_S from hadronic data

- Many determinations of α_S based on hadronic data exist. For example CMS $t\bar{t}$ determination at 7 TeV (arXiv:1307.1907), included as "independent" category in the PDF average.
- Methodology:
  - Compute the χ^2 of the particular dataset in a range of α_S and determine the minimum.
  - Use external PDFs scanned fitted at each scanned value of α_S. Value of PDF χ^2 ignored.
- Some also provide simultaneous PDF determinations although with limitations in dataset and parametrization (e.g. H1 jets, arXiv:1709.07251; see talk by K. Rabbertz).

Interplay between partial and global χ^2

(See Z.K., arxiv:1802.05236)

- Hadronic predictions (and thus the partial χ^2) depend on PDFs.
- The PDFs are themselves the result of a global χ^2 optimization.
- Suddenly there are *two* χ^2s in the problem as well as *two* datasets: Partial and global.
    - Certainly PDF and hadronic based determinations cannot be considered independent.
- The methodology, i.e.
    1. Restricting to the best fit PDF and
    2. then minimizing the partial χ^2 along α_S

  minimizes neither the partial or the total χ^2 nor a combination of them.

Erroneous minimization in a real example

- Can perform an "hadronic" determination with any subset of data used in the global NNPDF α_s fit.
- Choose Z p_T data and find that *"best fit"* is $\alpha_S = 0.124$.
- Could find a *"better best fit"* at $\alpha_S = 0.120$. Both Z p_T and the rest of the data agree better.

χ^2/d.o.f..	$\alpha_S = 0.120$ weighted Zp_T	$\alpha_S = 0.124$ default
Total	1.226	1.281
Zp_T	0.94	1.11

- Found it by minimizing $(\text{global } \chi^2) + 31(Zp_T\ \chi^2)$, but that's a detail.

Erroneous minimization in a toy model

- Imagine PDF characterized by a single parameter b. (α, PDF) parameter space simply a plane.
- Assume both partial and total χ^2 are paraboloids

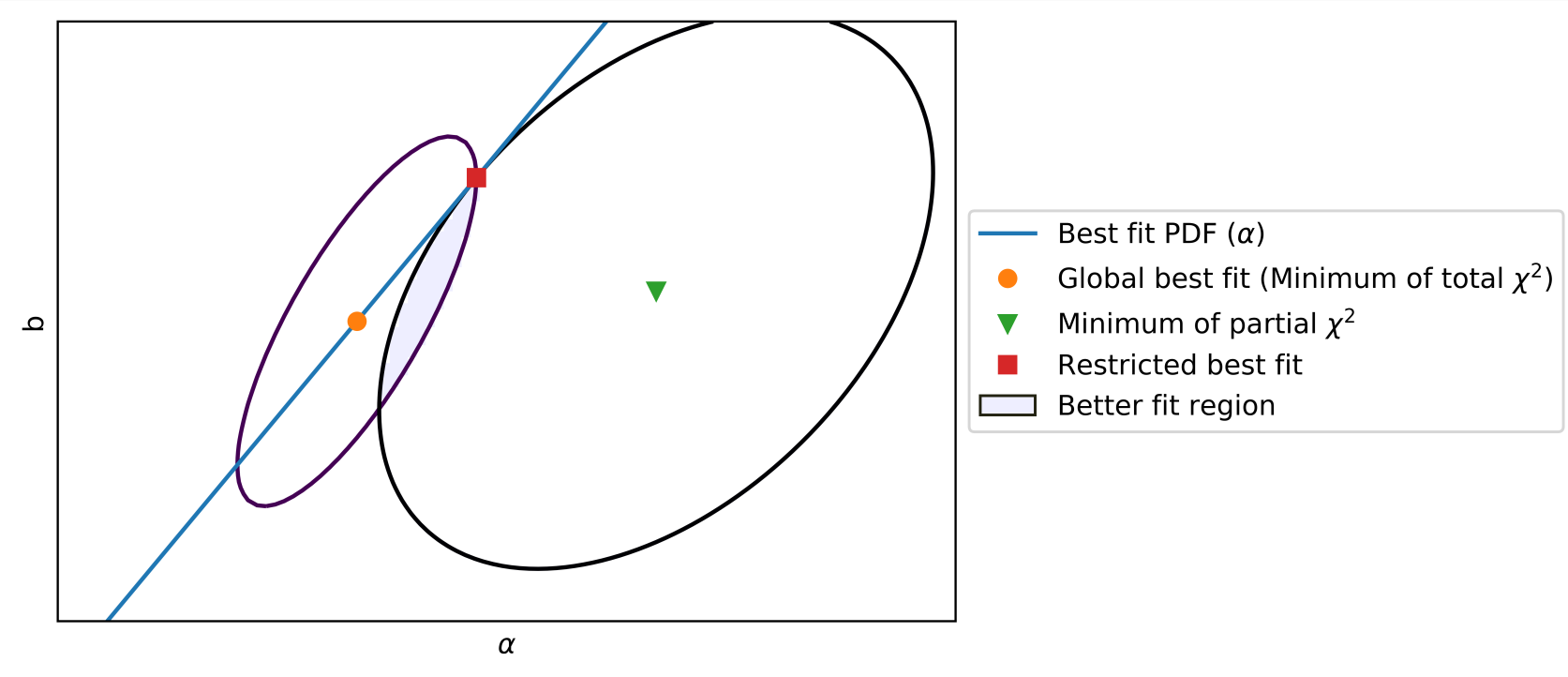
$$\chi^2_{\text{total}}(\alpha, b) = t_1 \left(\alpha \cos\theta + b \sin\theta\right)^2 + t_2 \left(-\alpha \sin\theta + b \cos\theta\right)^2$$

$$\chi^2_{\text{partial}}(\alpha, b) = p_1 \left((\alpha - \delta_\alpha)\cos\phi + (b - \delta_b)\sin\phi\right)^2$$
$$+ p_2 \left(-(\alpha - \delta_\alpha)\sin\phi + (b - \delta_b)\cos\phi\right)^2$$

The best PDF for a given α is

$$b_{\text{best}}(\alpha) = \arg\min_b \chi^2_{\text{total}}(\alpha, b)$$

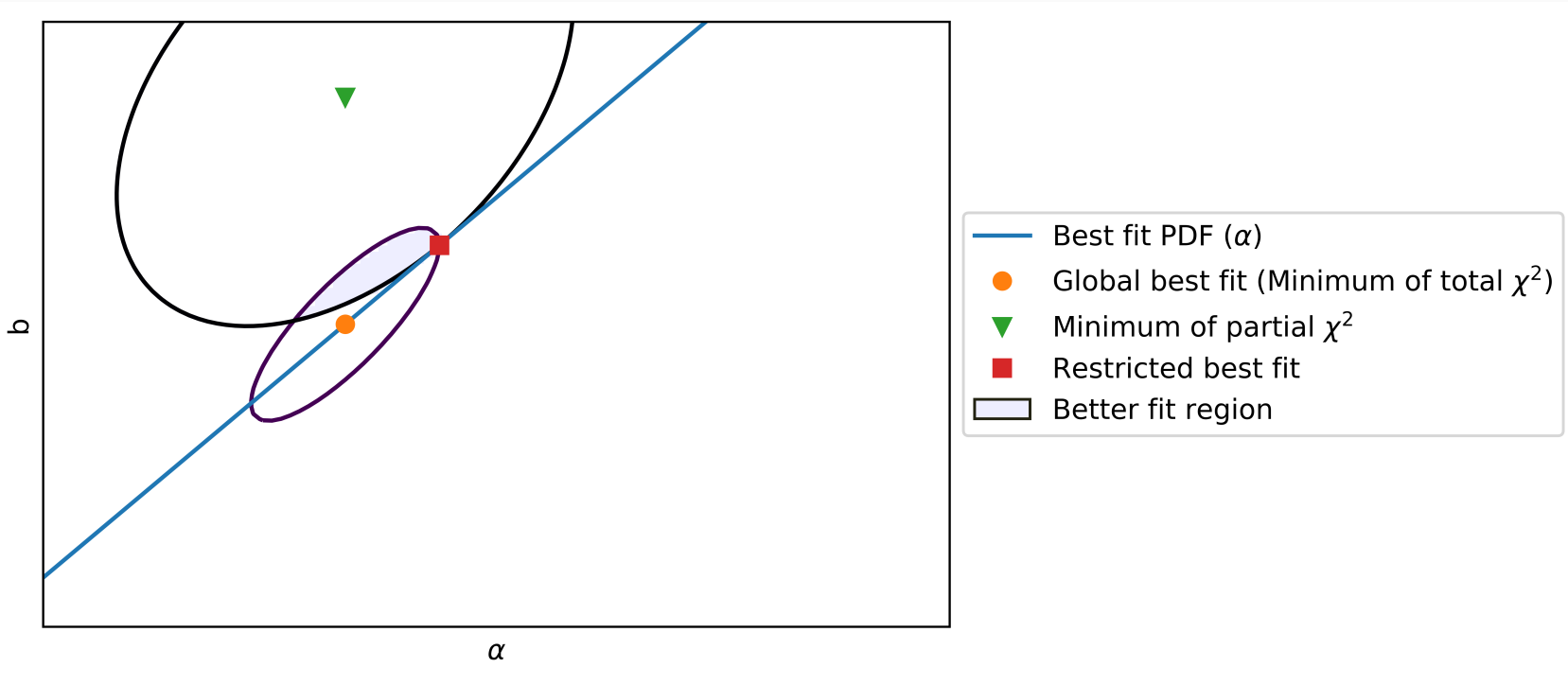
Better fit regions



- The procedure using best fit PDFs would yield the minimum of the partial χ^2 along the best fit PDF (red square)
- In fact the overlapping region describes better both datasets.

Making it worse

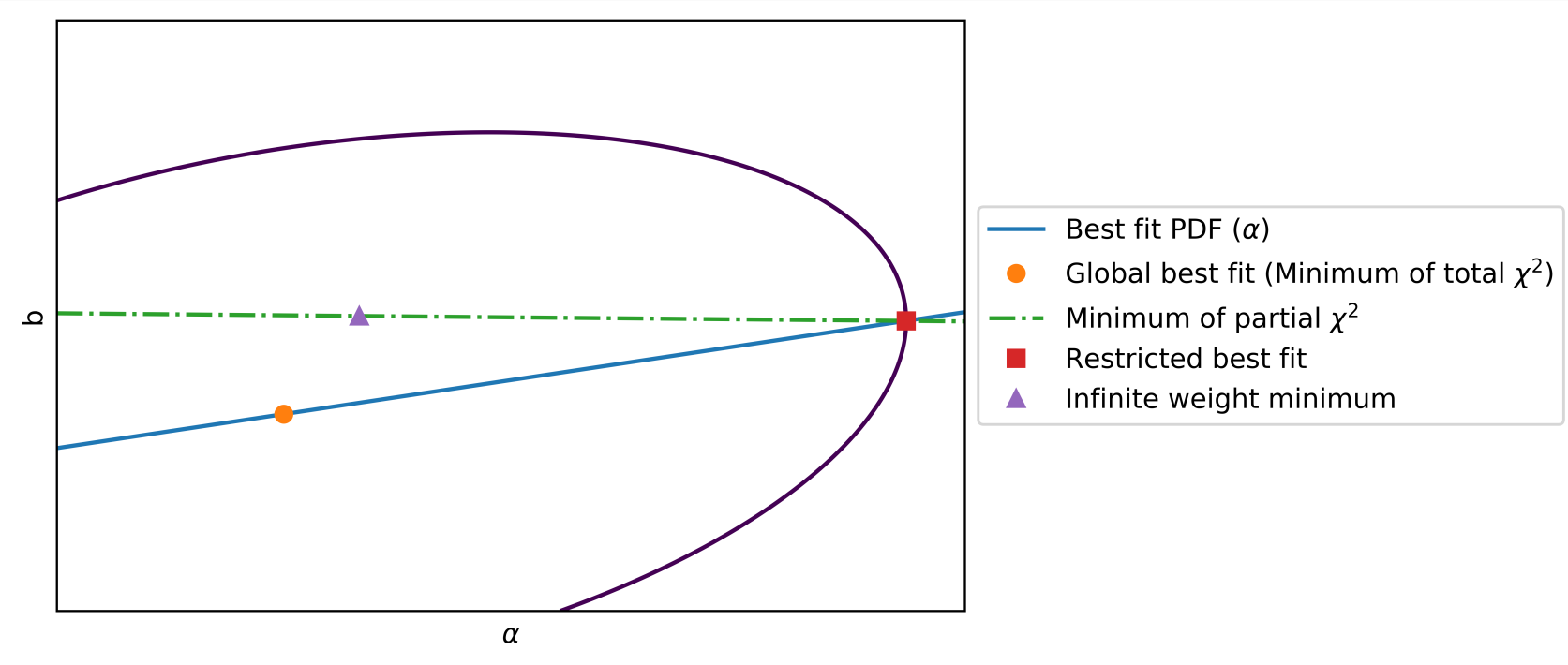
- Parameters can be tweaked so that the restricted minimum is far away from a desirable value.



A slightly more realistic model

- A more realistic situation is where the partial set doesn't determine both α_s and the PDF on its own, but only a combination of them.

$$\chi^2_{\text{partial}}(\alpha, b) = p_1 \left((\alpha - \delta_\alpha) \cos\phi + (b - \delta_b) \sin\phi \right)^2$$



- Similar situation: Now there is a better fit segment.
- Datasets with little handle on α can look artificially inconsistent.
- Can consider the $\lim_{w\to\infty} \chi^2_{\text{total}} + w\chi^2_{\text{partial}}$

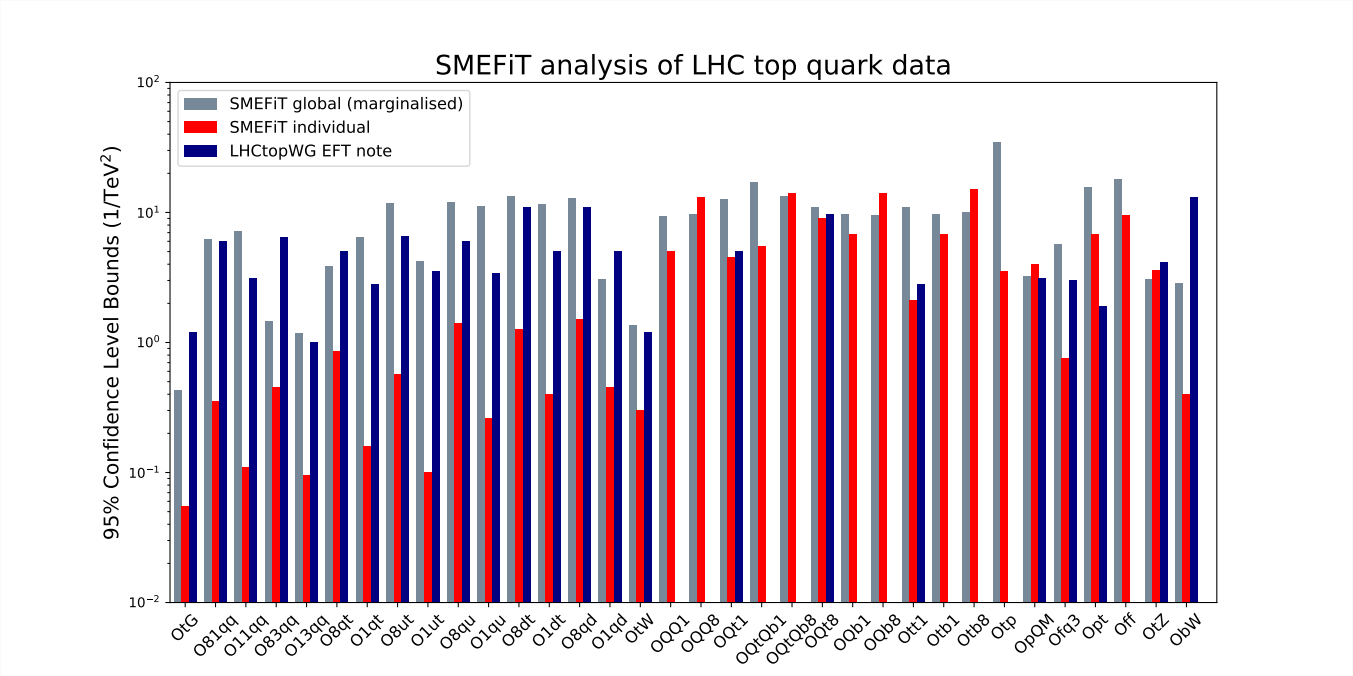
Dangers of partial fits

- Inconsistent minimization puts into question several determinations of α_S, as well as the QCD average.
- PDFs generally cannot be considered external to fitting problems.
- Problems not limited to α_S, but affect any parameter that changes PDFs.

Fits of EFT coefficients

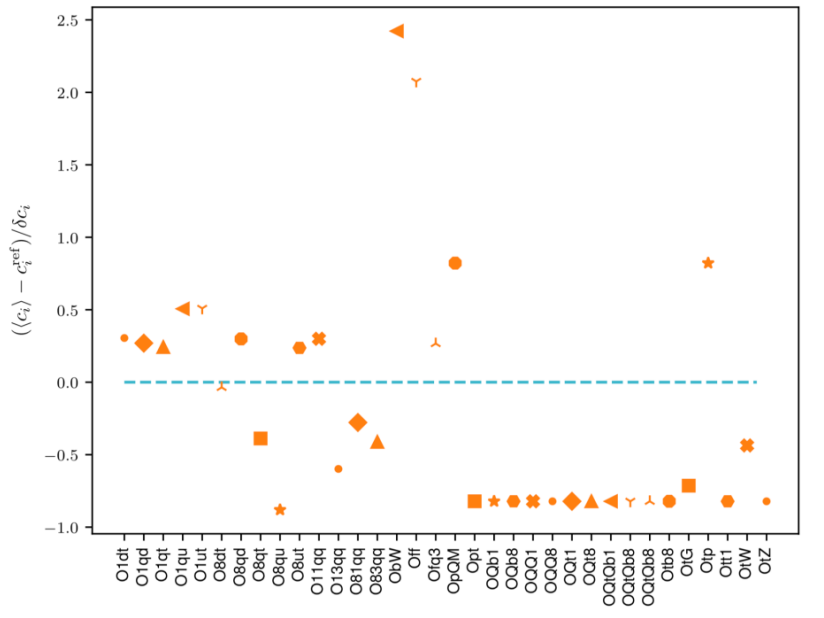
(See talk by Emma Slade and Hartland et al, arxiv:1901.05965)

- Used fitting methodology and validation inspired by NNPDF to constrain EFT coefficients in the top sector.
- Some results:
  - Bounds derived from fitting only one operator at a time (and fixing the rest to zero) much tighter than in a full fit.
  - Including more operators results in looser bounds.

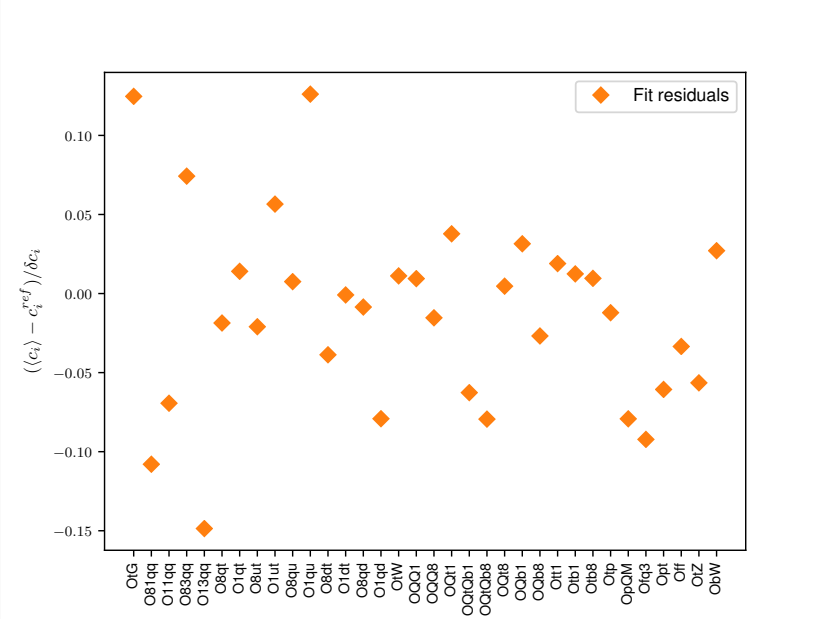


Residuals on a closure test

- Methodology tested on simulated data assuming SM only.



(a) Individual fit. Note: Marker shapes not meaningful



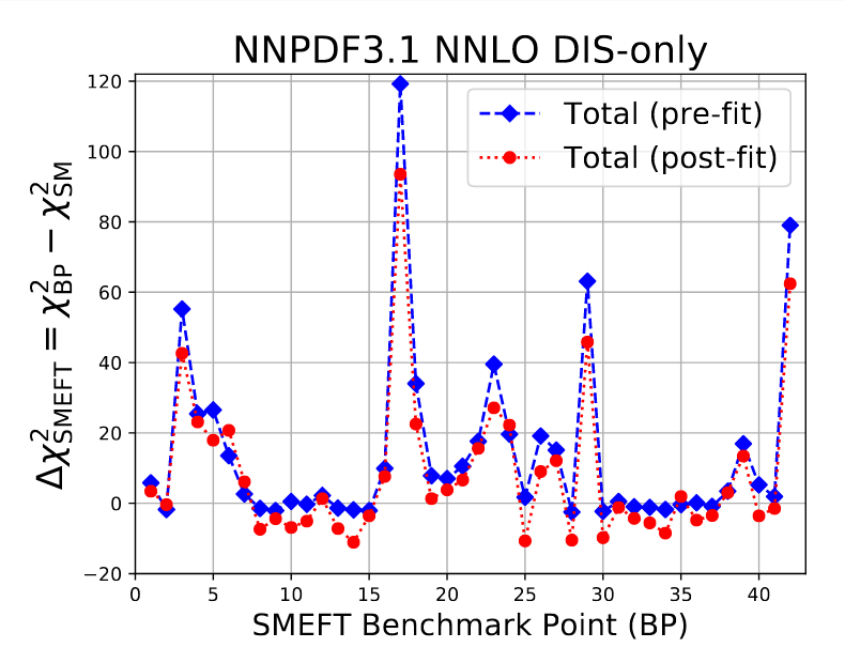
(b) Simultaneos fit

- Suggests EFT bounds only reliable when constrained globally.

EFT effects inside the PDF

(See Carrazza et al, arxiv:1905.05215)

- PDF fits sensible to inclusion of EFT effects in the determination itself.
- Added EFT corrections to DIS structure functions and used the result to determine PDFs .



- Results might vary more with an increased dataset.
- Putting it all together should be interesting: Global dataset + “complete” set of operators + PDFs + other parameters.

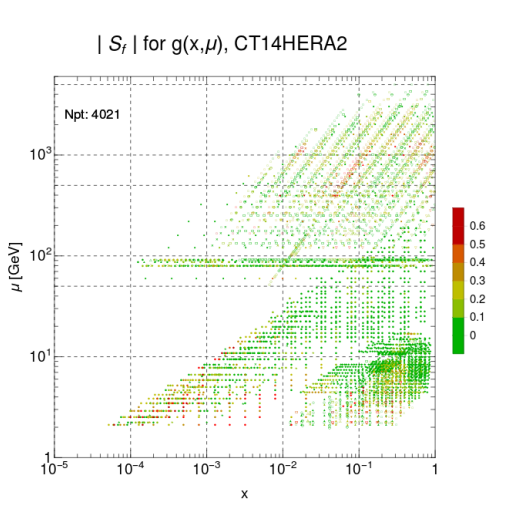
Visualization and analysis tools

(See talks by Carl Schmidt and Timothy Hobbs)

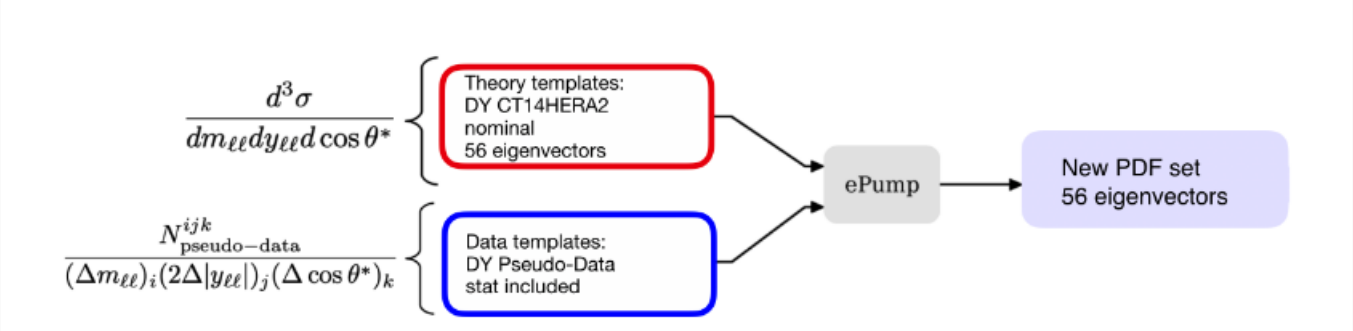
- Existing ways to understand how PDFs change upon inclusion of new data and which datasets cause discrepancies are ad hoc and manual.
- It would be good to be able to *visualize* data dependency.
- Lots of work on open source tools done by CTEQ lately.

Examples of tools

PDFSense (Wang et al, arxiv:1808.07470) code to study various data dependency, such as cluster data points by their effect on PDFs.



ePump (Willis et al, arxiv:1809.09481) Can estimate the effect of new data in PDFs.



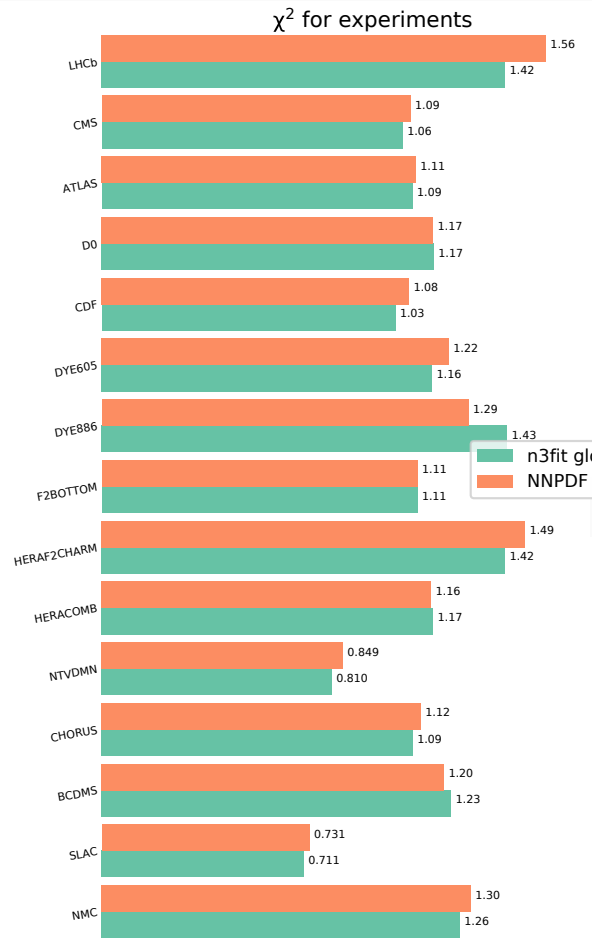
Improvements in the NNPDF fitting methodology

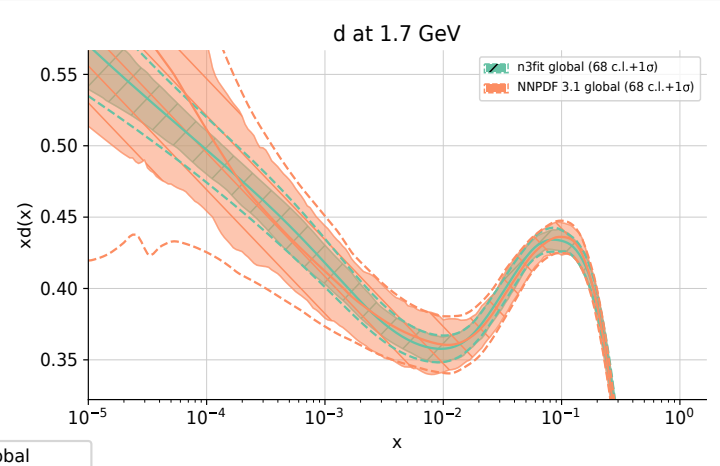
(See talk by Juan Cruz, and Carrazza, Cruz-Martinez arxiv:1907.05075)

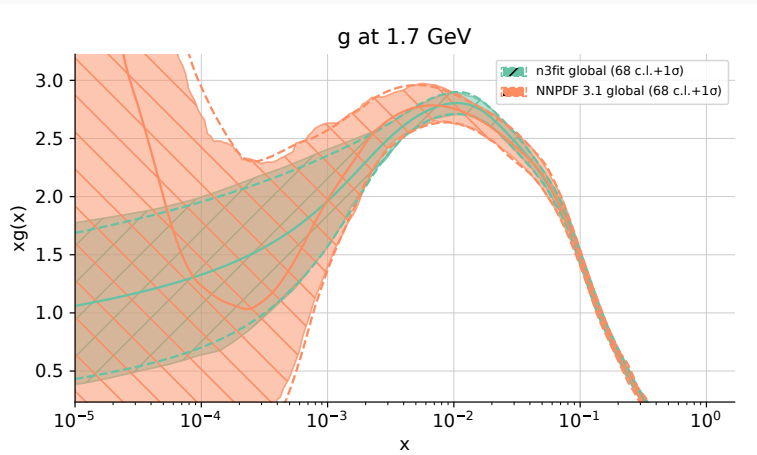
- Current NNPDF methodology was state of art Machine Leaning some 10 years ago. But the field has moved:
  - Gradient based optimization of large networks and complicated function.
  - Quality industry backed library present.
- The NNPDF methodology has been implemented with Keras + Tensorflow using gradient techniques.
  - Performance increased by a factor 20.
  - Allows to remove a lot of legacy code.
  - Still have to patch Tensorflow a bit to get good memory usage with convolutions.
- Increased speed will open the door to new classes of studies!

Results (so far)

- Central values and fit quality remarkably stable.
- PDF uncertainties significatively reduced.
  - Not so much at the level of predictions: High frequency component removed.





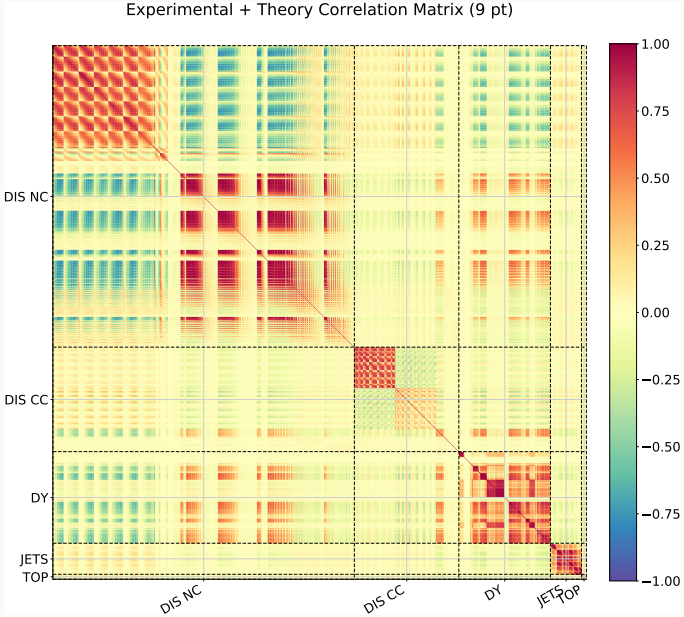


Theory uncertainties in PDFs

- Most important issue in current PDF sets is that they do not account for any uncertainty in the underlying theory, and specifically Missing Higher Order Uncertainties.
- Scale variations typically used as a proxy. Only method known to generalize to multiscale problems like PDF fits.
- Much work towards improving on it lately. E.g. (Harland-Lang, Thorne, arxiv:1811.08434)
- A First Determination of Parton Distributions with Theoretical Uncertainties (NNPDF, arxiv:1905.04311)
    - NLO PDFs with a model for MHOUs available for first time.
    - General formalism: NNPDF, arxiv:1906.10698.

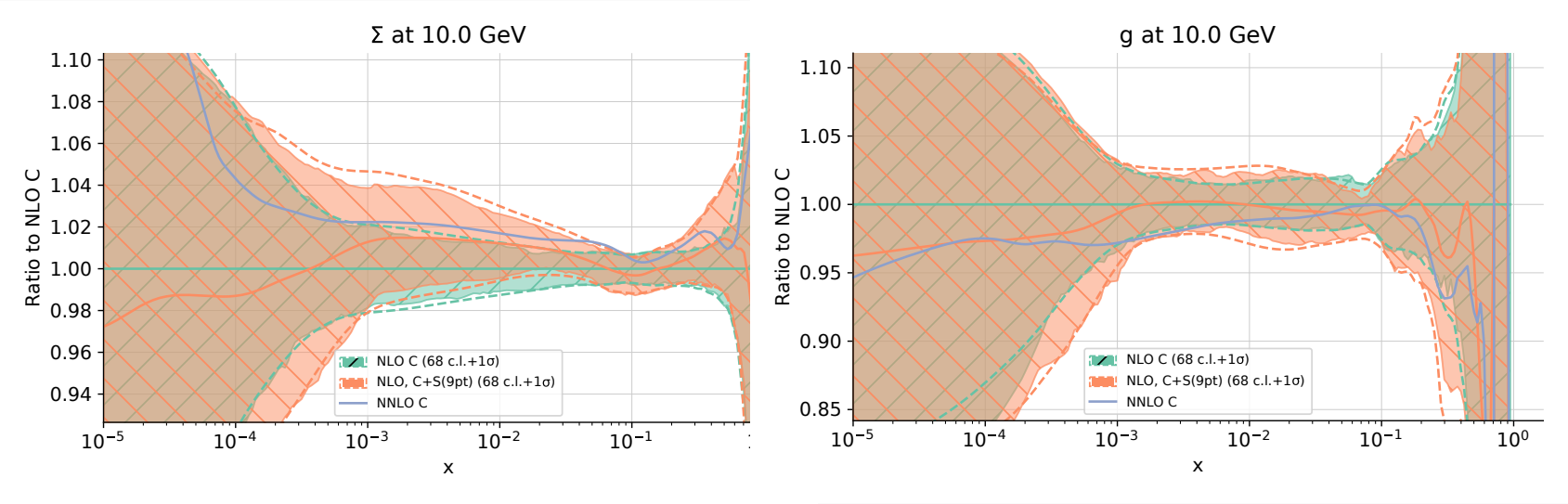
Theory covariance fit in NNPDF

- Idea: Include the integrated effect of scale variations in each PDF replica.
- In that case “scale variation systematics” can simply be added in quadrature to the experimental uncertainties. Define by:
  1. Labelling each dataset with a *process*: DIS CC, DIS NC, DY, Top, Jets
  2. Defining a covariance model for points belonging to the same or a different process.
     - Note that both the size of the uncertainties and the correlation model are heuristic. Choose to generalize envelopes with plausible correlations (“9 point prescription”).



Effects of the theory covariance matrix

- Data with large scale scale uncertainties weighted down in favour of more perturbatively stable data.
- Central value shift towards NNLO.



Concluding remarks

- PDF fits might be obsolete for precision physics. Need global fits including PDF parameters.
- Frameworks (both the theoretical and software ones) need to become more open and integrated.
- Independent cross checks very valuable
  - E.g. photon PDF in (Harland–Lang, Martin, Nathvani, Thorne, arxiv: 1907.02750, talk by Thomas Cridge) re-derives a LUXQED-like framework in an elegant way and finds consistent results.

Thank you!