Stable covariance matrices for PDF fits

Michael Wilson*

The Higgs Centre for Theoretical Physics, University of Edinburgh

September 17, 2019

*In collaboration with Zahari Kassabov and Emanuele R. Nocera

Intro

Use the χ^2 statistic to determine agreement between data and theory

$$\chi^2 = d^t \Sigma d \tag{1}$$

where d is vector of differences between data and theory and Σ is convariance matrix.
For a dataset which is not included in the fit we have the following

$$\langle \chi^2 \rangle = N, \tag{2}$$

and

$$\mathrm{std}\,(\chi^2) = \sqrt{2N}. \tag{3}$$

provided we have a compatibility between data and theory

Take ATLAS W/Z production $7\,TeV$, differential cross section in rapidity [arxiv. 1612.03016].

- with NNPDF3.1 [arxiv. 1706.00428] $\chi^2/N_{\rm data} = 2.2$ - one of the datasets described poorly by fit.
- with 34 data points this corresponds to 5σ discrepancy

Why might we get a bad χ^2?

- theory gives poor description of data
- uncertainties are underestimated
- correlations are difficult to estimate (this talk)

- We can perform a regularization* on the covariance matrix, $\Sigma \to \tilde{\Sigma}$

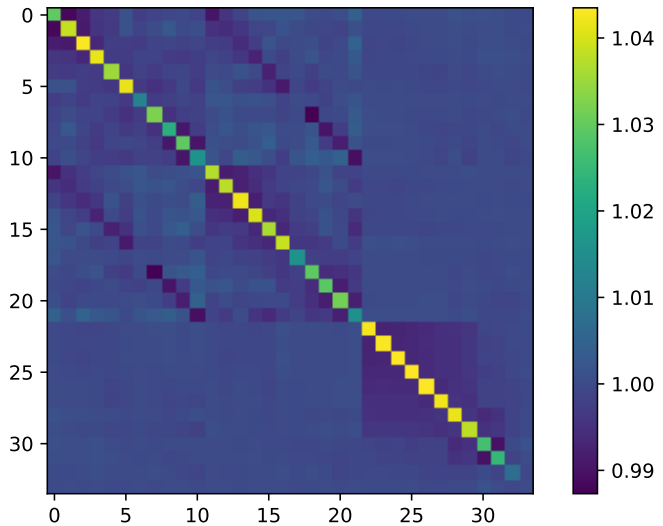


Figure: Ratio of regularized covariance to original covariance $\tilde{\Sigma}_{ij}/\Sigma_{ij}$. The maximum ratio of standard deviation is 1.02. The average ratio of standard deviations is 1.015.

*to be defined later

- We can perform a regularization* on the covariance matrix, $\Sigma \rightarrow \tilde{\Sigma}$

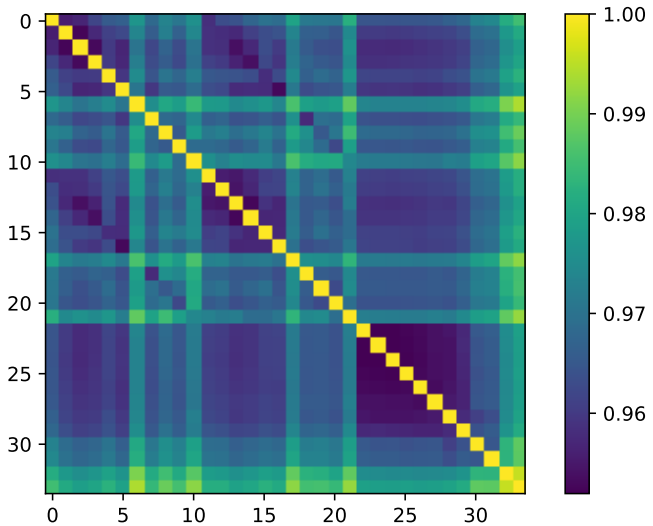


Figure: Ratio of elements of regularized correlation matrix, $\tilde{c}$, to original correlation matrix, c. $\max |1 - \tilde{c}_{ij}/c_{ij}| = 0.05$. The average relative change is 0.03.

*to be defined later

recalculate the χ^2 statistic with $\tilde{\Sigma}$ and compare to old value

	using Σ	using $\tilde{\Sigma}$
χ^2/N_{data}	2.2	1.2

new value is within 1σ

Defining stability

Consider a toy model, with small statistical uncertainties $\epsilon \ll 1$ and high correlations

$$\Sigma = \begin{pmatrix} \epsilon^2+1 & 1 & 1 & 1 \\ 1 & \epsilon^2+1 & 1 & 1 \\ 1 & 1 & \epsilon^2+1 & 1 \\ 1 & 1 & 1 & \epsilon^2+1 \end{pmatrix} \tag{4}$$

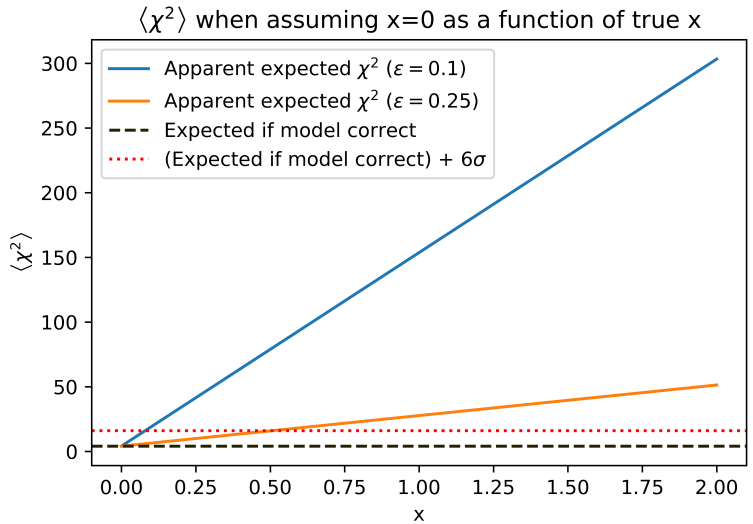
- matrix has eigenvalues $e_{1,2,3} = \epsilon^2$, $\epsilon_4 = \epsilon^2 + 4$
- L^2 condition number given by the ratio of smallest and largest eigenvalue $\kappa(\Sigma) = \frac{\epsilon^2+4}{\epsilon^2}$

Can introduce some input parameter $x \in [0, 2]$ which controls correlation of final datapoint.

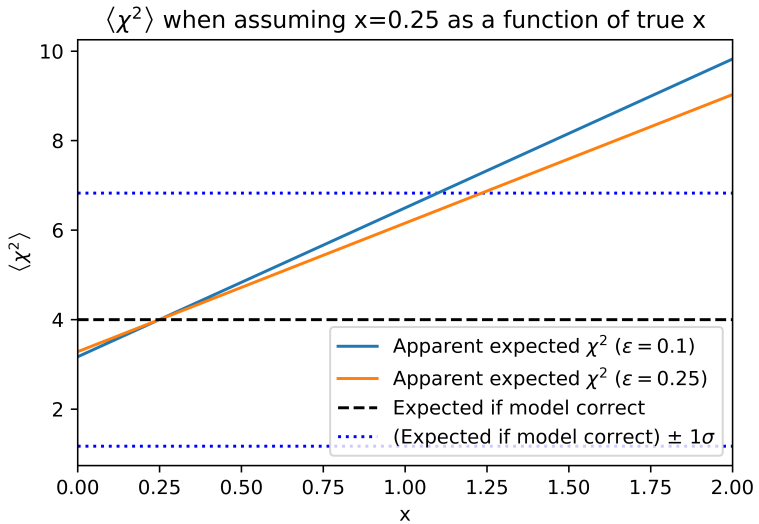
$$\Sigma = \begin{pmatrix} \epsilon^2 + 1 & 1 & 1 & 1 - x \\ 1 & \epsilon^2 + 1 & 1 & 1 - x \\ 1 & 1 & \epsilon^2 + 1 & 1 - x \\ 1 - x & 1 - x & 1 - x & \epsilon^2 + 1 \end{pmatrix} \tag{5}$$

but x has some uncertainty.

We can plot the expected χ^2 if we assume $x = 0$ but the data was actually generated with $x \in [0, 2]$



Can perform same exercise assuming $x = 0.25$. Note: the blue lines now are 1σ bands!



toy model prefers smaller values for $1 - x$ in terms of stability.

- often given datasets with highly correlated uncertainties
- correlations are hard to estimate, some datasets provide multiple correlation models, see e.g:
  - ATLAS jets at 7 TeV [arxiv. 1410.8857]
  - sensitivity studied in detail by Harland-Lang, Martin, Thorne [arxiv. 1711.05757]
- default correlations should be chosen to maximise stability

How to define stability? Assume the underlying model is correct, but that the covariance is wrong, another possible covariance could be given by $\tilde{\Sigma} = \Sigma + \delta S$

- S is some symmetric NxN matrix
- δ is a dimensionless number measuring size of fluctuation

take data distributed according to $\tilde{\Sigma}$

$$\tilde{d} \in \mathcal{N}(0, \tilde{\Sigma}), \tag{6}$$

and define

$$\bar{\chi}^2 \equiv \tilde{d}^t \, \Sigma \, \tilde{d}. \tag{7}$$

take the difference between χ^2 calculated on d and $\tilde{d}$ keeping Σ fixed

$$\begin{aligned} \Delta\chi^2 &= |\langle\chi^2\rangle - \langle\bar{\chi}^2\rangle| \\ &= |N - \langle\bar{\chi}^2\rangle| \end{aligned} \tag{8}$$

stability condition is that this difference is much less than statistical fluctuations of χ^2

$$|N - \langle\bar{\chi}^2\rangle| \ll \sqrt{2N} \tag{9}$$

Generic analysis

Without knowledge of S we can get the *approximate* relation

$$\Delta\chi^2 \ll \sqrt{2N} \Rightarrow \kappa(\Sigma) \ll 1/\delta \tag{10}$$

where $\kappa(\Sigma)$ is the L^2 condition number of the covariance matrix.

Disdvantages:

- slightly heuristic bound, working on more rigorous proof
- in practise covariance matrices span uncertainties with many orders of magnitude

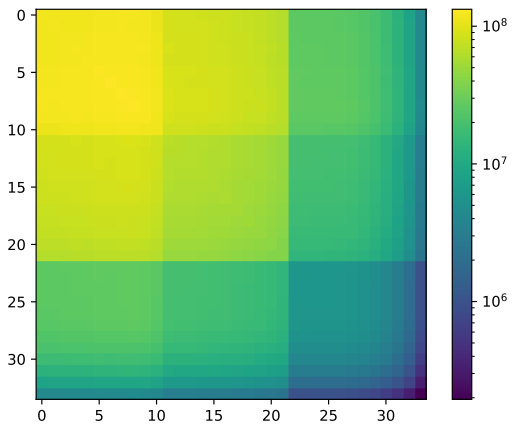


Figure: Covariance matrix for ATLAS WZ production dataset

- don't want to mix uncertainties with big magnitudes and uncertainties with small magnitudes

What about regularizing the correlation matrix? Correlation matrix is covariance of reduced variables: d/σ

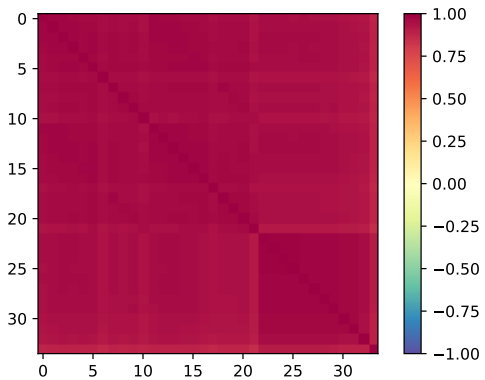


Figure: Correlation matrix for ATLAS WZ production dataset

looks a lot like toy model!

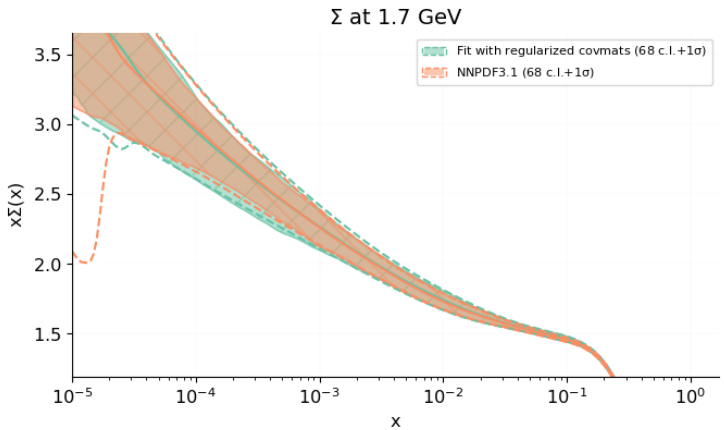
- obtain correlation matrix from covariance matrix $c_{ij} = \frac{\Sigma_{ij}}{\sqrt{\Sigma_{ii}\Sigma_{jj}}}$
- perform eigenvalue decomposition on c giving Λ and U such that $c = U^t \Lambda U$.
- obtain new eigenvalues $\tilde{\Lambda}_{ij} = \delta_{ij} \min(\Lambda_{ij}, \hat{\lambda})$ where $\hat{\lambda} = \max(\Lambda_{ij})/k$ where k is an input parameter specifying a threshold condition number
- construct $\tilde{c} = U^t \tilde{\Lambda} U$ and use to obtain new, regularized covariance matrix $\tilde{\Sigma}_{ij} = \tilde{c}\sqrt{\Sigma_{ii}\Sigma_{jj}}$

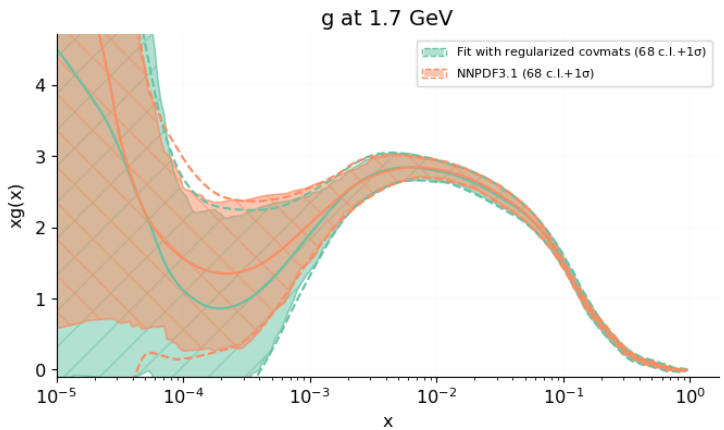
This is our regulariation procedure!

fitting with $\tilde{\Sigma}$

Perform regularization with condition number threshold 500 on each dataset correlation matrix then perform fit using all other settings of NNPDF3.1 we find (preliminary results - paper on in depth study in preparation)

Stat Estm.	Fit using $\tilde{\Sigma}\|_{k=500}$	fit using Σ
$\chi^2/N_{\rm data}$	1.00035	1.16328
$\langle\chi^2/N_{\rm data}\rangle$	1.095 ± 0.038	1.253 ± 0.033





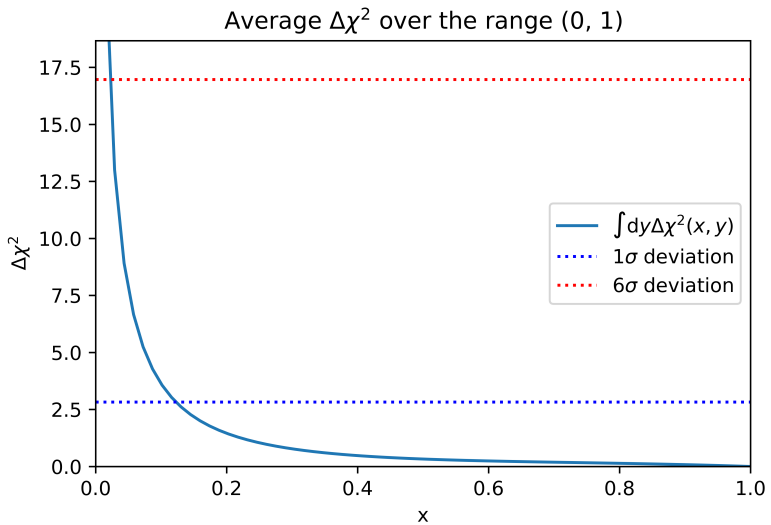
PDFs are unchanged from 3.1, despite dramatic change in global χ^2

Conclusions

- datasets with high correlations and low statistical uncertainties can have unstable χ^2 (toy model)
- experimentalists have knowledge of the input parameters, much better positioned to choose stable correlation models
- we can perform a generic regularization of covariance matrices based on SVD of correlation matrix
- Investigation into effects of regularization on fits still in progress but results look promising, in particular want to study sensitivity on condition number threshold [paper in progress Z. Kassabov, E. R. Nocera, MW]

backup slides

toy model prefers no correlations with uniform prior on x



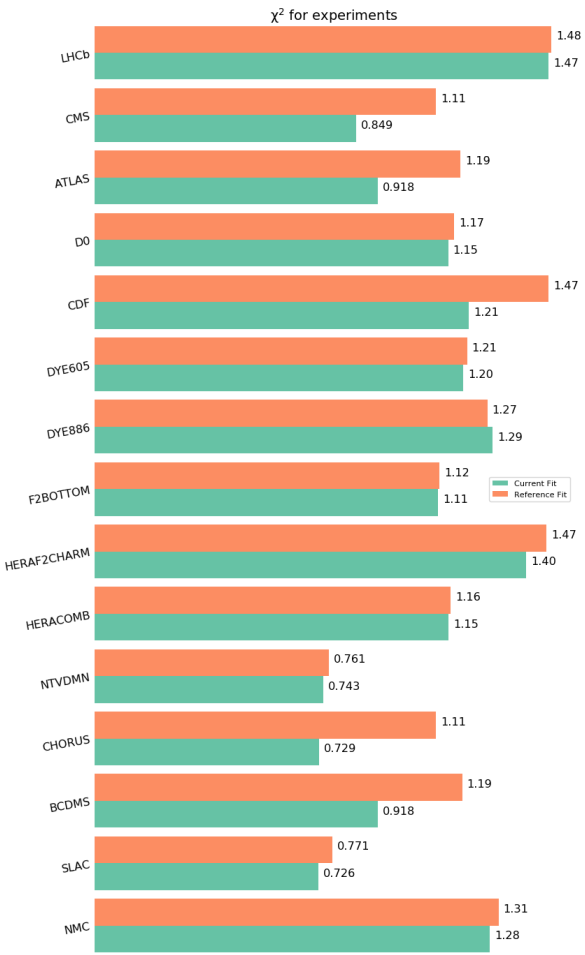
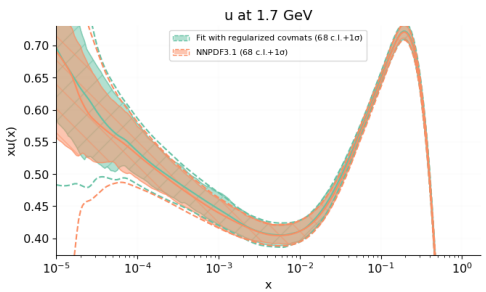
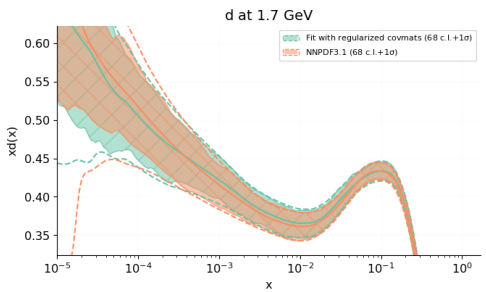
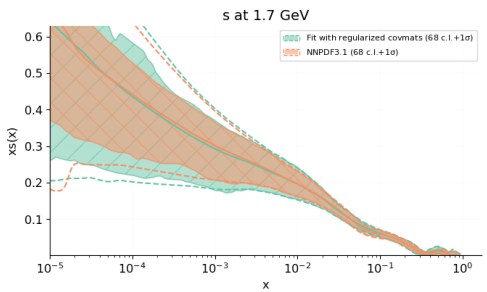


Figure: χ^2 by experiment for fit with $\tilde{\Sigma}|_{k=500}$ (current) and fit with Σ (reference)


u at 1.7 GeV
Fit with regularized covmats (68 c.l.+1σ)
NNPDF3.1 (68 c.l.+1σ)
xu(x)
x


d at 1.7 GeV
Fit with regularized covmats (68 c.l.+1σ)
NNPDF3.1 (68 c.l.+1σ)
xd(x)
x


s at 1.7 GeV
Fit with regularized covmats (68 c.l.+1σ)
NNPDF3.1 (68 c.l.+1σ)
xs(x)
x