Pseudo-data fits*: some comments on error definitions

Ronan McNulty, Francesco deLorenzi
University College Dublin

PDF4LHC 19th May 2009

* plots preliminary and in some cases indicative

Pseudo-data fits

- From eigenvector phase space (assume multinomial distribution), choose one set: **'truth'**
- Generate many pseudo-data sets corresponding to given luminosity
- Fit each pseudo-data set: **'pseudo-measurement'**
- Compare **pseudo-measurement** to **truth**
  - centre of distribution gives bias
  - width of distribution gives precision
- Repeat

What is fit? (MSTW,CTEQ,Alekhin)

We considered $\frac{d\sigma}{dy}$ for W+,W-,Z.

$$f_0 = \frac{d\sigma}{dy}$$: distribution obtained with central eigenvectors

$$f_i = \frac{d\sigma}{dy}(\lambda_i = 1, \lambda_{\neq i} = 0)$$: distribution with i^{th} e.v. moved 1σ

Fit

$$\chi^2(\lambda_0, \lambda_i) = \sum_{j=1}^{\#bins} \left[\frac{x_j - \lambda_0\left(f_0 + \lambda_i\left(f_i - f_0\right)\right)}{\sigma_j} \right]^2 + \sum_{i=1}^{\#e.v.} \lambda_i^2$$

Normalisation (Luminosity)

data in j bins, each with uncertainty σ

What is fit? (NNPDF)

We considered $\frac{d\sigma}{dy}$ for W+,W-,Z.

$$f_i = \frac{d\sigma}{dy}$$ for i^{th} replica

Fit $$\chi^2(\lambda_0) = \sum_{j=1}^{\#bins} \left[\frac{x_j - \lambda_0 f_i}{\sigma_j} \right]^2$$

... and only consider consistent replicas
(Chisquared probability > 1 %)
[should explicitly include in NN procedure ... to do]

Thanks to Maria and Juan for suggestion on how to sample a large number of times from the NNPDF distribution

Results for precision on luminosity shown at DIS09....

DIS2009 – SF & low-x WG

	0.1 fb^{-1}			
	MSTW08	CTEQ66	Alekhin	NNPDF
W+	1.8	2.4	2.0	2.9
W-	1.9	2.6	2.2	2.7
Z	1.9	2.4	2.2	2.4
WWZ	1.7	2.3	1.8	2.0
	1 fb^{-1}			
	MSTW08	CTEQ66	Alekhin	NNPDF
W+	1.6	2.2	1.8	2.4
W-	1.6	2.3	2.1	2.4
Z	1.7	2.1	1.9	1.8
WWZ	1.3	2.1	1.4	2.2
	10 fb^{-1}			
	MSTW08	CTEQ66	Alekhin	NNPDF
W+	1.3	2.0	1.5	2.5
W-	1.2	1.9	1.6	3.0
Z	1.4	1.9	1.9	1.9
WWZ	0.8	1.7	1.0	-

Percentage statistical uncertainty on fitted luminosity

Precision doesn't scale with $\frac{1}{\sqrt{N_{events}}}$

Comments at September PDF4LHC meeting:

- If such a fit is a valid fit, it also has produced valid improved values for the eigenvectors.

- If you trust my luminosity derived in this way, you should trust my eigenvector values too.

- BUT – my eigenvector values are more precise than values that would be given by the global fit.

Comments at September PDF4LHC meeting:

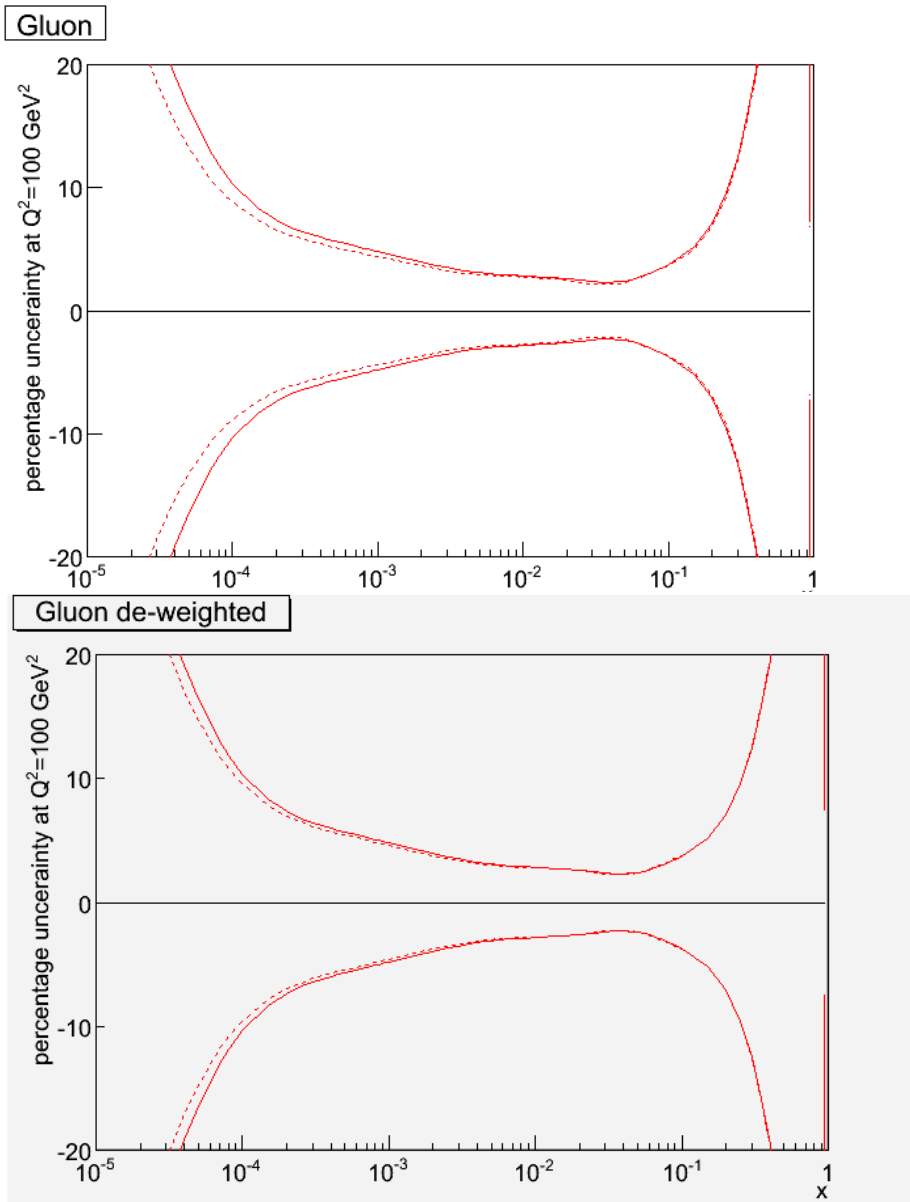
- If such a fit is a valid fit, it also has produced valid improved values for the eigenvectors.
- If you trust my luminosity derived in this way, you should trust my eigenvector values too.
- BUT – my eigenvector values are more precise than values that would be given by the global fit.

Approximate effect of global fit (should redo global fits):

$$\chi^2(\lambda_0,\lambda_i) = \frac{1.65^2}{(\Delta\chi^2_{90})}\sum_{j=1}^{\#bins}\left[\frac{y_j - \lambda_0\left(f_0 + \lambda_i\left(f_i - f_0\right)\right)}{\sigma_j}\right]^2 + \sum_{i=1}^{\#e.v.}\lambda_i^2$$

MSTW “deweight” data by ~4, CTEQ “deweight” data by ~6

Effect on gluon PDF of fit to 1fb^{-1} of LHCb Z data (MSTW08)



solid line: current uncertainty

dashed line: with LHCb data

<u>Straight fit</u>

x=1E-4, 11% -> 8%

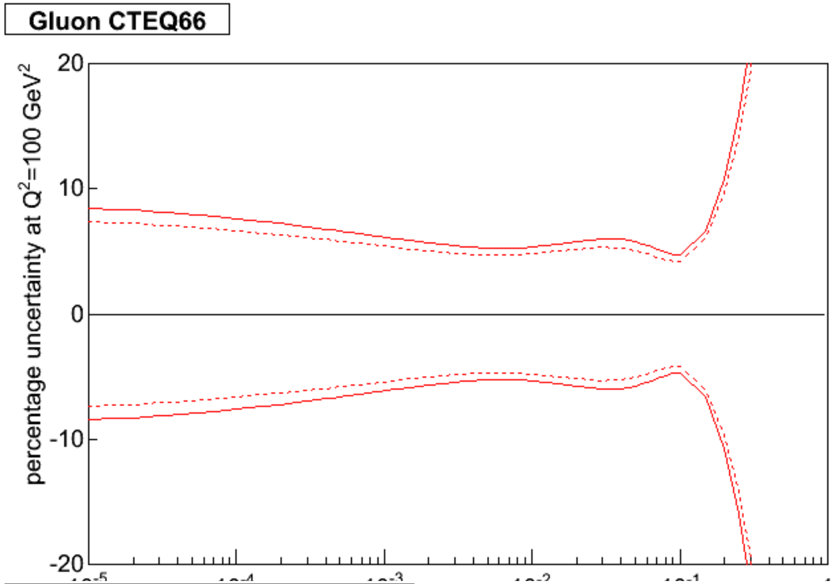
x=5E-5, 17%->13%

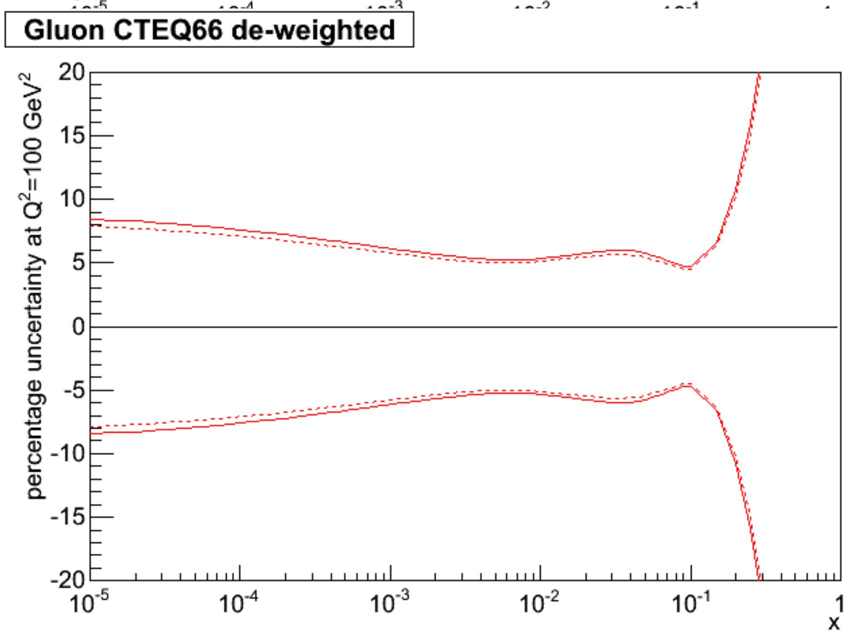
<u>Deweighted fit</u>

x=1E-4, 11%->10%

x=5E-5, 17%->15%

Effect on gluon PDF of fit to $1fb^{-1}$ of LHCb Z data (CTEQ66)





<u>Straight fit</u>

x=1E-4, 7.5% -> 6.5%

x=5E-5, 7.5%-> 6.5%

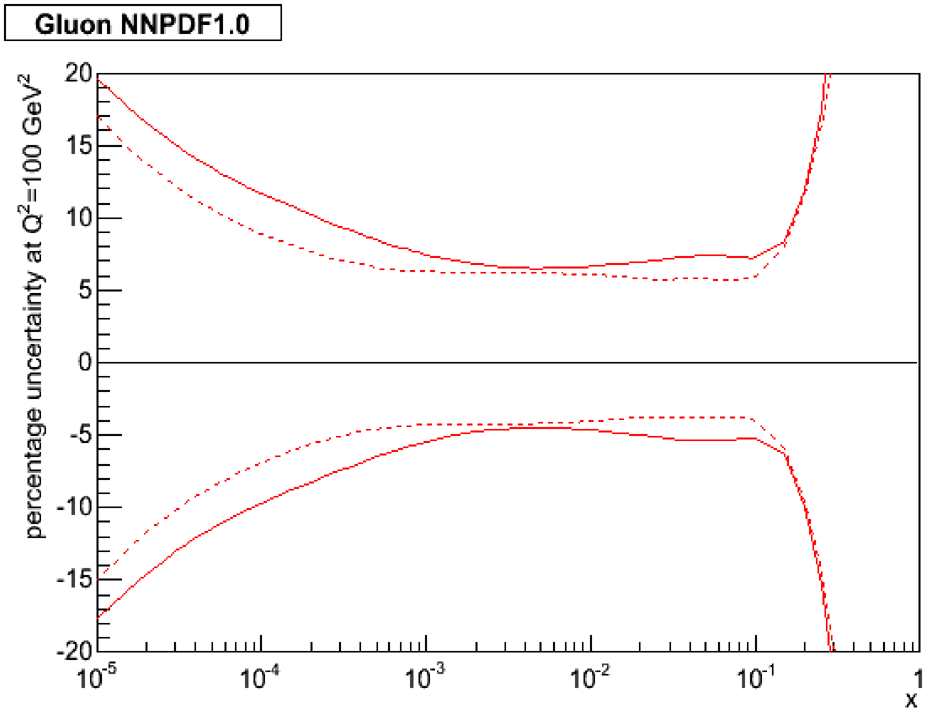
<u>Deweighted fit</u>

x=1E-4, 7.5%->7%

x=5E-5, 7.5%->7%

(Smaller difference because impact of data is less)

Effect on gluon PDF of fit to $1fb^{-1}$ of LHCb Z data (NNPDF1.0)

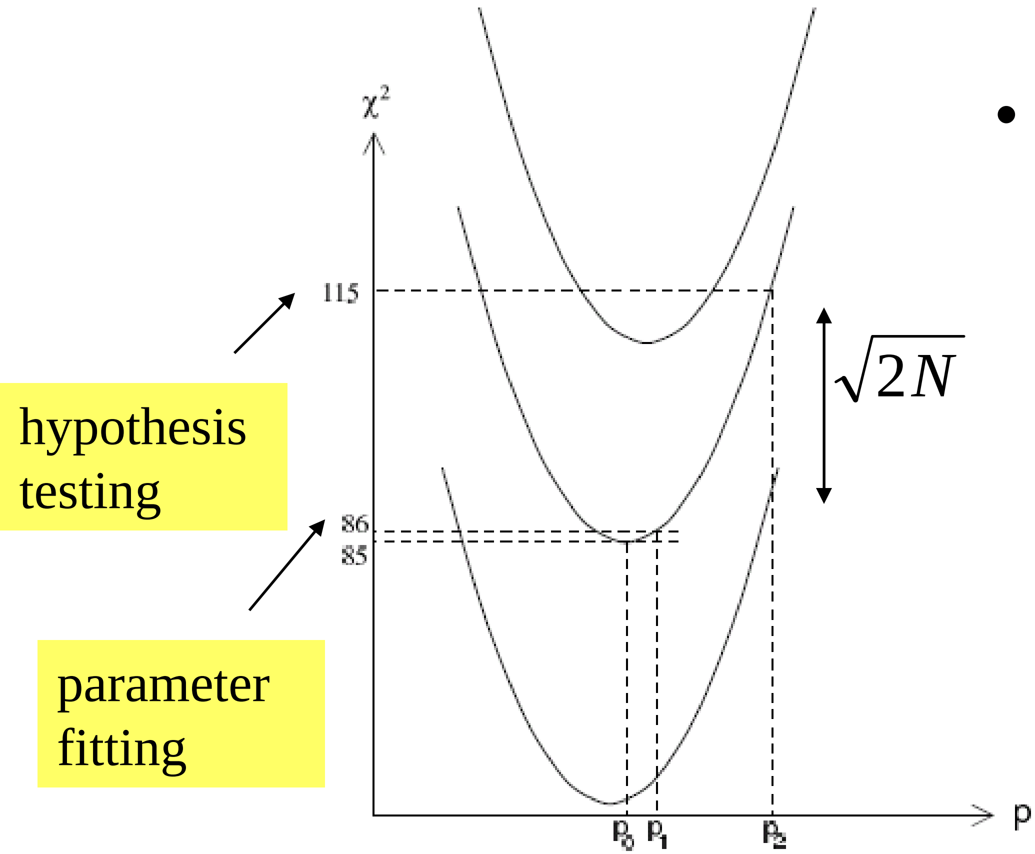


Fit

x=1E-4, 12% -> 9%

x=5E-5, 13%-> 11%

Dynamic Tolerance

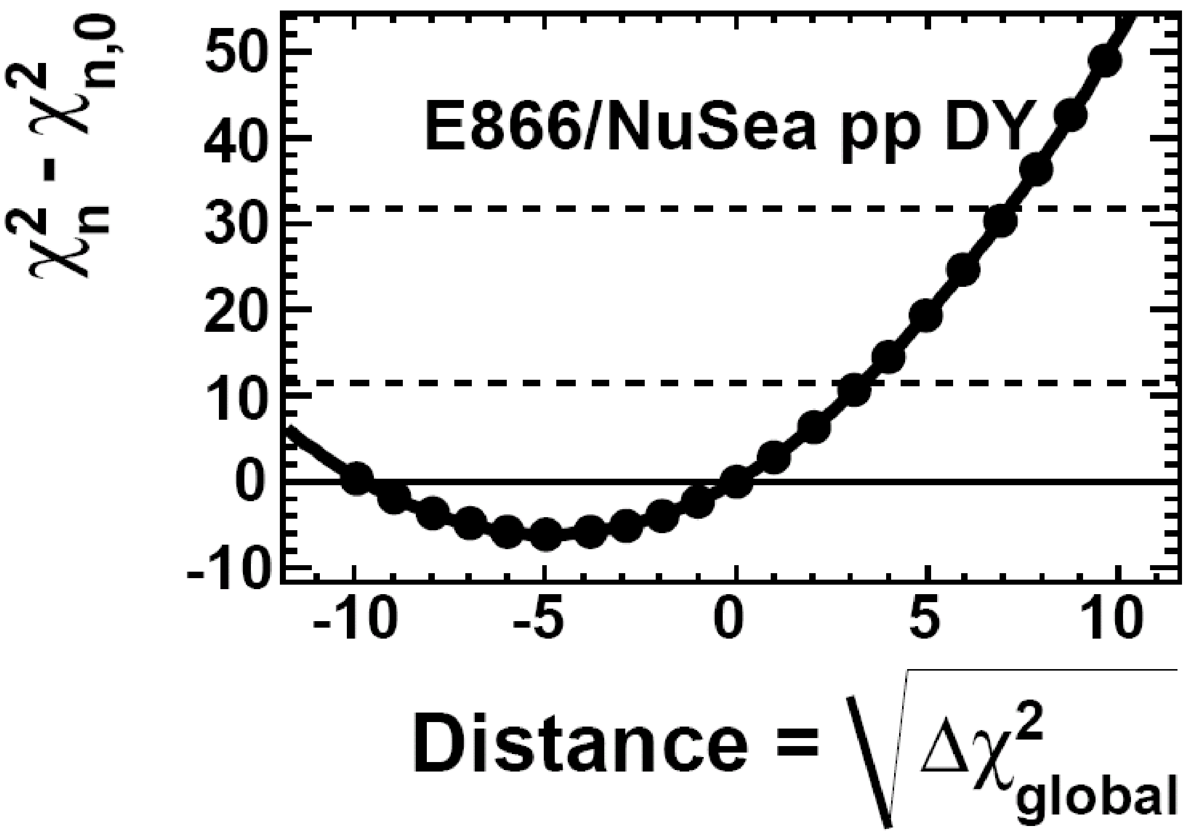


- In principle the dynamic tolerance method (see Watt PDF4LHC Feb 08) seems a better way to proceed.

•Collins&Pumplin hep-ph/0105207

Dynamic Tolerance

- In principle the dynamic tolerance method (see Watt PDF4LHC Feb 08) seems a better way to proceed.

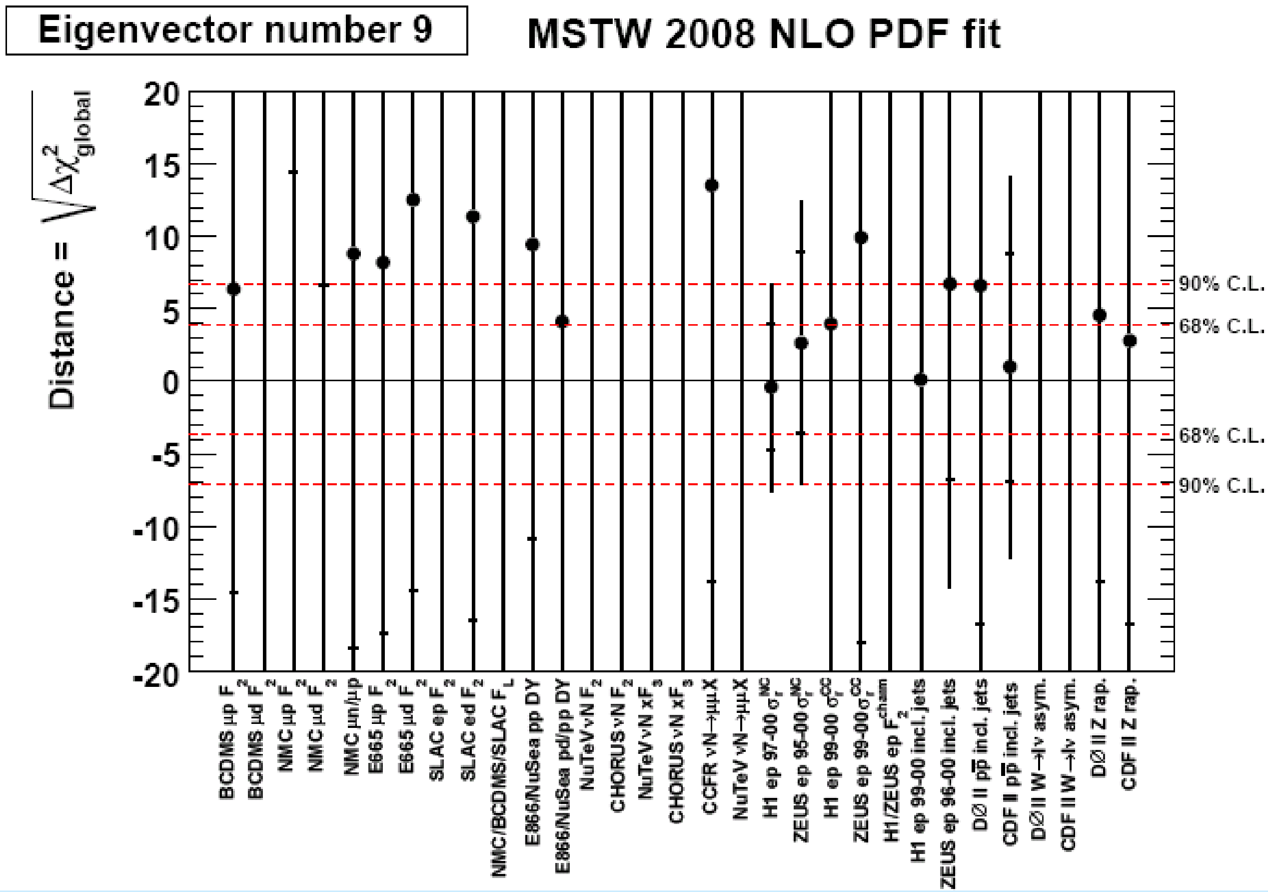


90% C.L. $< \left(\frac{\chi^2_{n,0}}{\xi_{50}}\right)\xi_{90}$

68% C.L.

$$\int_0^{\xi_{90}} \mathrm{d}\chi^2 \; P_N(\chi^2) = 0.90.$$

Dynamic Tolerance



If LHCb were the dominant experiment, how small could the tolerance be?

Would we reproduce the simple Chisq fit? i.e. tolerance 1.

Dynamic Tolerance

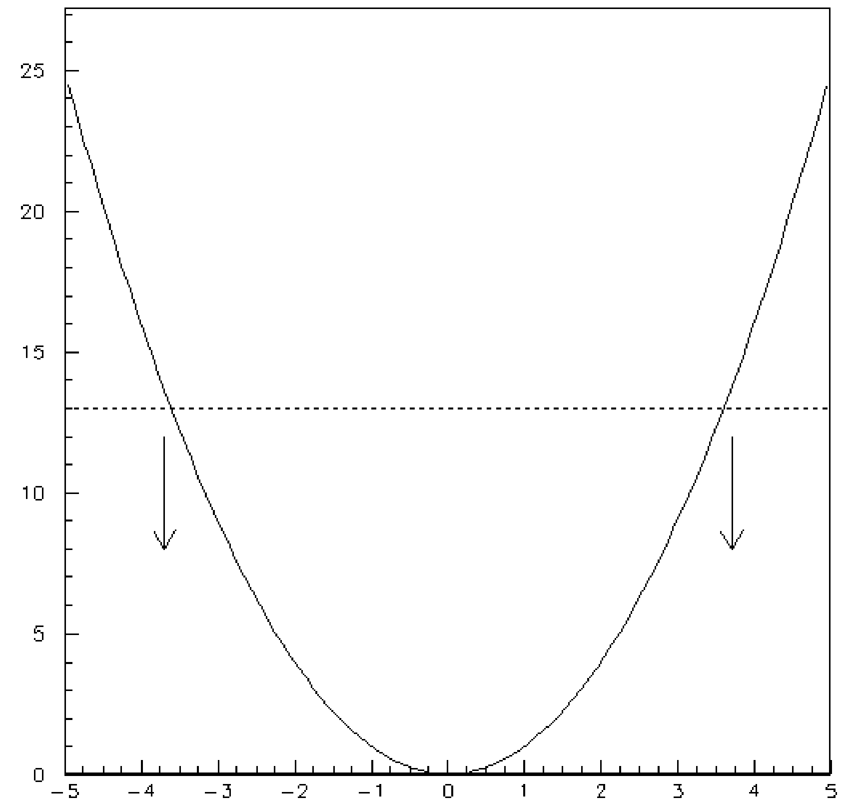
- Assume we bin data in 50 rapidity bins
- If dominant experiment then

$$\Delta\chi^2 = \chi_n^2 - \chi_{n,0}^2$$

$$\xi_{50} \approx 50, \xi_{90} \approx 63$$

-
- Error defn $\approx \Delta\chi_{90}^2 = 13$
- Scale LHCb

data by $\approx \frac{\sqrt{13}}{1.65} \approx 2.2$



Dynamic Tolerance

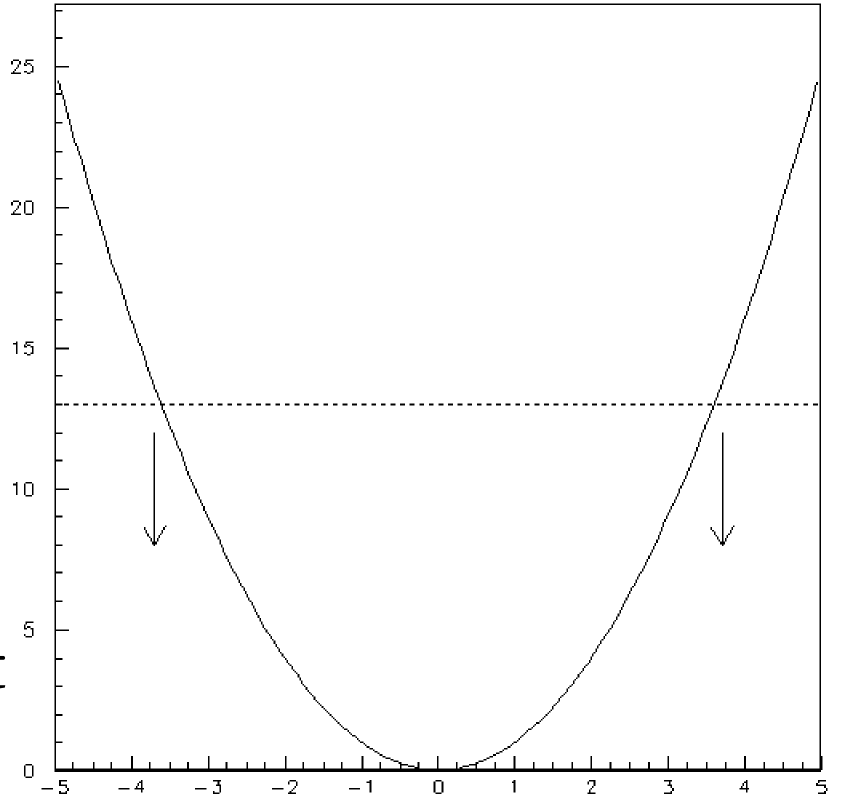
- Assume we bin data in N rapidity bins
- If data dominant then

$$\Delta\chi^2 = \chi_n^2 - \chi_{n,0}^2$$

$$\xi_{50} \approx N, \xi_{90} \approx N + \sqrt{1.65}\sqrt{2N}$$

- Error defn $\approx \Delta\chi_{90}^2 = \sqrt{1.65}\sqrt{2N}$

- Scale LHCb data by $\approx 0.7*(2N)^{1/4}$



Dynamic Tolerance

- Deweighting of data scales with $N^{1/4}$
  - Experimental Brinksmanship. Report your data in as few bins as possible (but not too few!)
  - Paradoxical situation (also for CTEQ) that if you have just one data set you will immediately deweight it
  - Tolerance may be reasonable when competing datasets
  - If you have a **dominant** dataset, this may not be the best approach.

Summary (Personal feelings)

- As an experimentalist, I feel uneasy that the full statistical impact of the data is not seen in the global fits.
- We would have to collect 10fb-1 of data in order to have the same statistical effect as 1fb-1 of data.
- It is unlikely that experiments are underestimating their systematics by such a large effect (see e.g. LEPEWWG or HERAPDF).
- More likely to be due to model dependence. (Have previously shown that 1fb-1 of data is sufficient to distinguish between models.)
- If so, try to determine a systematic error due to the model but don't scale experimental errors