Tolerance in global PDF analysis

Graeme Watt

University College London

PDF4LHC workshop, CERN, Geneva
22nd February 2008

Uncertainties in global PDF analysis

Theoretical errors

- *Examples:* input parameterisation form, neglected higher-order and higher-twist QCD corrections, electroweak corrections, choice of cuts, nuclear corrections, heavy flavour treatment.
- Difficult to quantify ($\rightarrow$ talks by A. Guffanti, R. Thorne, S. Forte).

Experimental errors

- In principle there **should** be a well-defined procedure for propagating experimental uncertainties on the fitted data points through to the PDF uncertainties.
  1. *Hessian method:* based on linear error propagation, produce eigenvector PDF sets suitable for use by the end user.

Uncertainties in global PDF analysis

Theoretical errors

- *Examples:* input parameterisation form, neglected higher-order and higher-twist QCD corrections, electroweak corrections, choice of cuts, nuclear corrections, heavy flavour treatment.
- Difficult to quantify (→ talks by A. Guffanti, R. Thorne, S. Forte).

Experimental errors

- In principle there **should** be a well-defined procedure for propagating experimental uncertainties on the fitted data points through to the PDF uncertainties.
  1. *Hessian method:* based on linear error propagation, produce eigenvector PDF sets suitable for use by the end user.
  2. *Lagrange multiplier method:* does not rely on linear error propagation, but requires access to global fit code.
  3. *Neural networks:* work in progress (→ talk by A. Guffanti).

Uncertainties in global PDF analysis

Theoretical errors

- *Examples:* input parameterisation form, neglected higher-order and higher-twist QCD corrections, electroweak corrections, choice of cuts, nuclear corrections, heavy flavour treatment.
- Difficult to quantify ($\rightarrow$ talks by A. Guffanti, R. Thorne, S. Forte).

Experimental errors

- In principle there **should** be a well-defined procedure for propagating experimental uncertainties on the fitted data points through to the PDF uncertainties.
  1. *Hessian method:* based on linear error propagation, produce eigenvector PDF sets suitable for use by the end user.
  2. *Lagrange multiplier method:* does not rely on linear error propagation, but requires access to global fit code.
  3. *Neural networks:* work in progress ($\rightarrow$ talk by A. Guffanti).

Traditional propagation of experimental uncertainties

- Assume $\chi^2_{\rm global}$ is quadratic about the global minimum $\{a_i^0\}$:

$$\Delta\chi^2_{\rm global} \equiv \chi^2_{\rm global} - \chi^2_{\rm min} = \sum_{i,j} H_{ij}(a_i - a_i^0)(a_j - a_j^0),$$

where the Hessian matrix has components

$$H_{ij} = \frac{1}{2}\frac{\partial^2\,\chi^2_{\rm global}}{\partial a_i \partial a_j}\bigg|_{\rm minimum}$$

- Uncertainty on quantity $F(\{a_i\})$ from linear error propagation:

$$\Delta F = T\sqrt{\sum_{i,j}\frac{\partial F}{\partial a_i}C_{ij}\frac{\partial F}{\partial a_j}},$$

where $C \equiv H^{-1}$ is the covariance matrix, and $T = \sqrt{\Delta\chi^2_{\rm global}}$ is the tolerance for the required confidence interval.

Traditional propagation of experimental uncertainties

- Assume $\chi^2_{\rm global}$ is quadratic about the global minimum $\{a_i^0\}$:

$$\Delta\chi^2_{\rm global} \equiv \chi^2_{\rm global} - \chi^2_{\rm min} = \sum_{i,j} H_{ij}(a_i - a_i^0)(a_j - a_j^0),$$

where the Hessian matrix has components

$$H_{ij} = \frac{1}{2}\frac{\partial^2\,\chi^2_{\rm global}}{\partial a_i \partial a_j}\bigg|_{\rm minimum}$$

- Uncertainty on quantity $F(\{a_i\})$ from linear error propagation:

$$\Delta F = T\sqrt{\sum_{i,j}\frac{\partial F}{\partial a_i}C_{ij}\frac{\partial F}{\partial a_j}},$$

where $C \equiv H^{-1}$ is the covariance matrix, and $T = \sqrt{\Delta\chi^2_{\rm global}}$ is the tolerance for the required confidence interval.

Eigenvector PDF sets (pioneered by CTEQ)

- Convenient to diagonalise covariance (or Hessian) matrix:

$$\sum_j C_{ij} v_{jk} = \lambda_k v_{ik},$$

where λ_k is the kth eigenvalue and v_{ik} is the ith component of the kth orthonormal eigenvector ($k = 1, \ldots, N_{\rm parameters}$).

- Expand parameter displacements from minimum in basis of rescaled eigenvectors $e_{ik} \equiv \sqrt{\lambda_k} v_{ik}$:

$$a_i - a_i^0 = \sum_k e_{ik} z_k.$$

- Then can show that

$$\chi^2_{\rm global} = \chi^2_{\rm min} + \sum_k z_k^2,$$

i.e. $\sum_k z_k^2 \leq T^2$ is the interior of a hypersphere of radius T.

Eigenvector PDF sets (pioneered by CTEQ)

- Convenient to diagonalise covariance (or Hessian) matrix:

$$\sum_j C_{ij} v_{jk} = \lambda_k v_{ik},$$

where λ_k is the kth eigenvalue and v_{ik} is the ith component of the kth orthonormal eigenvector ($k = 1, \ldots, N_{\rm parameters}$).

- Expand parameter displacements from minimum in basis of rescaled eigenvectors $e_{ik} \equiv \sqrt{\lambda_k} v_{ik}$:

$$a_i - a_i^0 = \sum_k e_{ik} z_k.$$

- Then can show that

$$\chi^2_{\rm global} = \chi^2_{\rm min} + \sum_k z_k^2,$$

i.e. $\sum_k z_k^2 \leq T^2$ is the interior of a hypersphere of radius T.

Use of eigenvector PDF sets

- Produce eigenvector PDF sets $S_k^{\pm}$ with parameters given by

$$a_i(S_k^{\pm}) = a_i^0 \pm t\, e_{ik},$$

with t adjusted to give the desired $T = \sqrt{\Delta\chi^2_{\rm global}}$.

- Then calculate uncertainties on a quantity F with

$$\Delta F = \frac{1}{2}\sqrt{\sum_k \left[F(S_k^+) - F(S_k^-)\right]^2},$$

or to account for asymmetric errors (S_0 = central PDF set):

$$(\Delta F)_+ = \sqrt{\sum_k \left[\max(F(S_k^+) - F(S_0), F(S_k^-) - F(S_0), 0)\right]^2}$$

$$(\Delta F)_- = \sqrt{\sum_k \left[\max(F(S_0) - F(S_k^+), F(S_0) - F(S_k^-), 0)\right]^2}$$

- Correlations between two quantities → talk by P. Nadolsky.

Use of eigenvector PDF sets

- Produce eigenvector PDF sets $S_k^\pm$ with parameters given by

$$a_i(S_k^\pm) = a_i^0 \pm t\, e_{ik},$$

with t adjusted to give the desired $T = \sqrt{\Delta\chi^2_{\rm global}}$.

- Then calculate uncertainties on a quantity F with

$$\Delta F = \frac{1}{2}\sqrt{\sum_k \left[F(S_k^+) - F(S_k^-)\right]^2},$$

or to account for asymmetric errors (S_0 = central PDF set):

$$(\Delta F)_+ = \sqrt{\sum_k \left[\max(F(S_k^+) - F(S_0), F(S_k^-) - F(S_0), 0)\right]^2}$$

$$(\Delta F)_- = \sqrt{\sum_k \left[\max(F(S_0) - F(S_k^+), F(S_0) - F(S_k^-), 0)\right]^2}$$

- Correlations between two quantities → talk by P. Nadolsky.

Criteria for choice of tolerance $T = \sqrt{\Delta\chi^2_{\mathrm{global}}}$

Parameter-fitting criterion

- $T^2 = 1$ for 68% (1-σ) C.L., $T^2 = 2.71$ for 90% C.L.
- Appropriate if fitting consistent data sets with ideal Gaussian errors to a well-defined theory.
- **In practice:** minor inconsistencies between fitted data sets, and unknown experimental and theoretical uncertainties, so not appropriate for global PDF analysis.

Hypothesis-testing criterion

- Much weaker than the parameter-fitting criterion: treat eigenvector PDF sets as alternative hypotheses.
- Determine T^2 from the criterion that each data set should be described within its 90% C.L. limit.

Criteria for choice of tolerance $T = \sqrt{\Delta\chi^2_{\rm global}}$

Parameter-fitting criterion

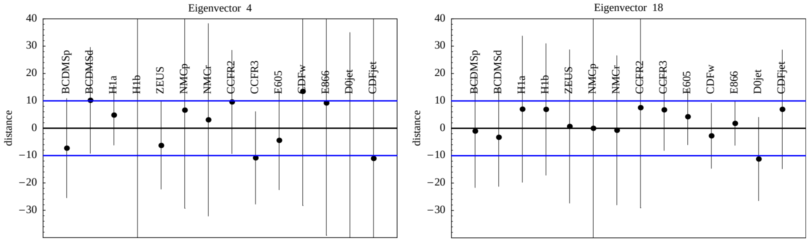
- $T^2 = 1$ for 68% (1-σ) C.L., $T^2 = 2.71$ for 90% C.L.
- Appropriate if fitting consistent data sets with ideal Gaussian errors to a well-defined theory.
- **In practice:** minor inconsistencies between fitted data sets, and unknown experimental and theoretical uncertainties, so not appropriate for global PDF analysis.

Hypothesis-testing criterion

- Much weaker than the parameter-fitting criterion: treat eigenvector PDF sets as alternative hypotheses.
- Determine T^2 from the criterion that each data set should be described within its 90% C.L. limit.

Choice of tolerance by CTEQ [hep-ph/0201195]

- For each eigenvector, plot location of the minimum for each data set and the 90% C.L. limits as the distance from the global minimum in units of $\sqrt{\Delta\chi^2_{\text{global}}}$:



- A rough "average" over all eigenvectors gives $T = 10$ . . .
- . . . But $T = 10$ exceeds the 90% C.L. limits of some data sets.

Choice of tolerance by MRST [hep-ph/0211080]

"We estimate $\Delta\chi^2 = 50$ to be a conservative uncertainty (perhaps of the order of a 90% confidence level or a little less than 2σ) due to the observation that **an increase of 50 in the global χ^2**, *which has a value $\chi^2 = 2328$ for 2097 data points,* **usually signifies that the fit to one or more data sets is becoming unacceptably poor**. *We find that* **an increase $\Delta\chi^2$ of 100 normally means that some data sets are very badly described** *by the theory."*

- Fairly qualitative statements.
- $\Rightarrow$ Study more quantitatively in new MSTW analysis.

Data sets fitted in MSTW 2008 NLO (prel.) analysis

Data set	$\chi^2/N_{\rm pts.}$
H1 MB 99 e^+p NC	9 / 8
H1 MB 97 e^+p NC	42 / 64
H1 low Q^2 96–97 e^+p NC	45 / 80
H1 high Q^2 98–99 e^-p NC	122 / 126
H1 high Q^2 99–00 e^+p NC	132 / 147
ZEUS SVX 95 e^+p NC	35 / 30
ZEUS 96–97 e^+p NC	86 / 144
ZEUS 98–99 e^-p NC	54 / 92
ZEUS 99–00 e^+p NC	62 / 90
H1 99–00 e^+p CC	29 / 28
ZEUS 99–00 e^+p CC	38 / 30
H1/ZEUS ep $F_2^{\rm charm}$	108 / 83
H1 99–00 e^+p incl. jets	19 / 24
ZEUS 96–97 e^+p incl. jets	29 / 30
ZEUS 98–00 $e^{\pm}p$ incl. jets	16 / 30
DØ I $p\bar{p}$ incl. jets	68 / 90
CDF II $p\bar{p}$ incl. jets	73 / 76
CDF II $W \to l\nu$ asym.	29 / 22
DØ II $W \to l\nu$ asym.	23 / 10
DØ II Z rap.	19 / 28
CDF II Z rap.	35 / 29

Data set	$\chi^2/N_{\rm pts.}$
BCDMS μp F_2	182 / 163
BCDMS μd F_2	187 / 151
NMC μp F_2	121 / 123
NMC μd F_2	103 / 123
NMC $\mu n/\mu p$	130 / 148
E665 μp F_2	57 / 53
E665 μd F_2	53 / 53
SLAC ep F_2	30 / 37
SLAC ed F_2	40 / 38
NMC/BCDMS/SLAC F_L	38 / 31
E866/NuSea pp DY	227 / 184
E866/NuSea pd/pp DY	15 / 15
NuTeV νN F_2	50 / 53
CHORUS νN F_2	26 / 42
NuTeV νN xF_3	40 / 45
CHORUS νN xF_3	31 / 33
CCFR $\nu N \to \mu\mu X$	65 / 86
NuTeV $\nu N \to \mu\mu X$	39 / 40
All data sets	**2497 / 2723**

- Red = Update to last MRST fit.

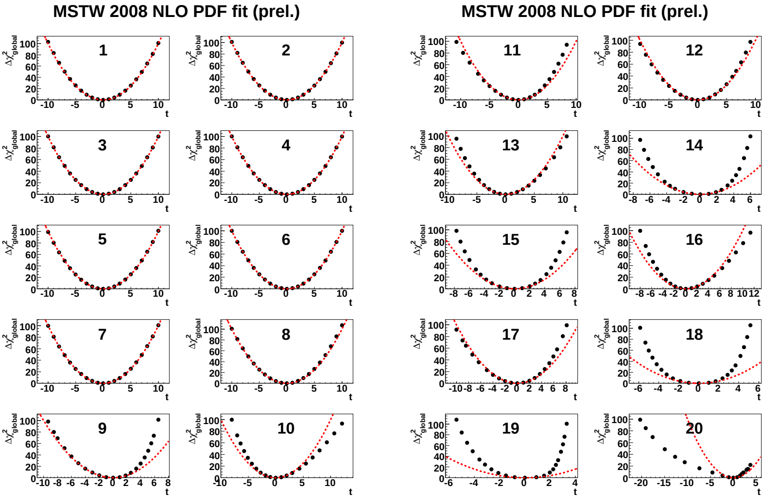
Input parameterisation in MSTW 2008 NLO (prel.) fit

At input scale $Q_0^2 = 1$ GeV^2:

$$
\begin{aligned}
xu_v &= A_u\, x^{\eta_1}(1-x)^{\eta_2}(1+\epsilon_u\,\sqrt{x}+\gamma_u\,x)\\
xd_v &= A_d\, x^{\eta_3}(1-x)^{\eta_4}(1+\epsilon_d\,\sqrt{x}+\gamma_d\,x)\\
xS &= A_S\, x^{\delta_S}(1-x)^{\eta_S}(1+\epsilon_S\,\sqrt{x}+\gamma_S\,x)\\
x\bar{d}-x\bar{u} &= A_\Delta\, x^{\eta_\Delta}(1-x)^{\eta_S+2}(1+\gamma_\Delta\,x+\delta_\Delta\,x^2)\\
xg &= A_g\, x^{\delta_g}(1-x)^{\eta_g}(1+\epsilon_g\,\sqrt{x}+\gamma_g\,x)+A_{g'}\,x^{\delta_{g'}}(1-x)^{\eta_{g'}}\\
xs+x\bar{s} &= A_+\, x^{\delta_S}\,(1-x)^{\eta_+}(1+\epsilon_S\,\sqrt{x}+\gamma_S\,x)\\
xs-x\bar{s} &= A_-\, x^{\delta_-}(1-x)^{\eta_-}(1-x/x_0)
\end{aligned}
$$

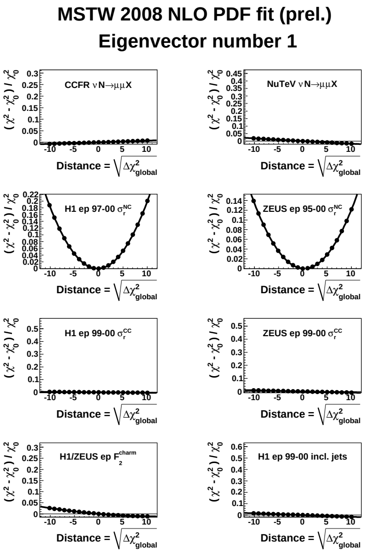
- A_u, A_d, A_g and x_0 are determined from sum rules.
- 20 parameters allowed to go free for eigenvector PDF sets, cf. 15 for MRST eigenvector PDF sets.

$\Delta\chi^2_{\rm global}$ vs. distance along each eigenvector, t



- Deviations from ideal quadratic behaviour (red dashed lines) for higher eigenvector numbers.

Fractional change in χ^2 for each data set



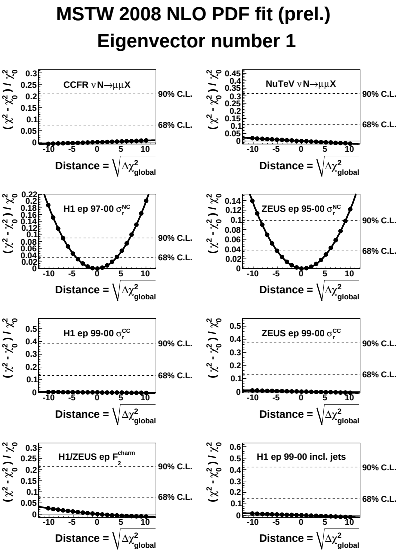
- Plot $(\chi^2 - \chi_0^2)/\chi_0^2$ versus the distance along a particular eigenvector.
- Define 90% C.L. region for each data set as

$$(\chi^2 - \chi_0^2)/\chi_0^2 < (\xi_{90} - \xi_{50})/\xi_{50}.$$

ξ_{90} is the 90th percentile of the χ^2-distribution with $N_{\rm pts.}$ d.o.f. $\xi_{50} \simeq N_{\rm pts.}$ is the most probable value.

- Similarly for the 68% C.L.

Fractional change in χ^2 for each data set



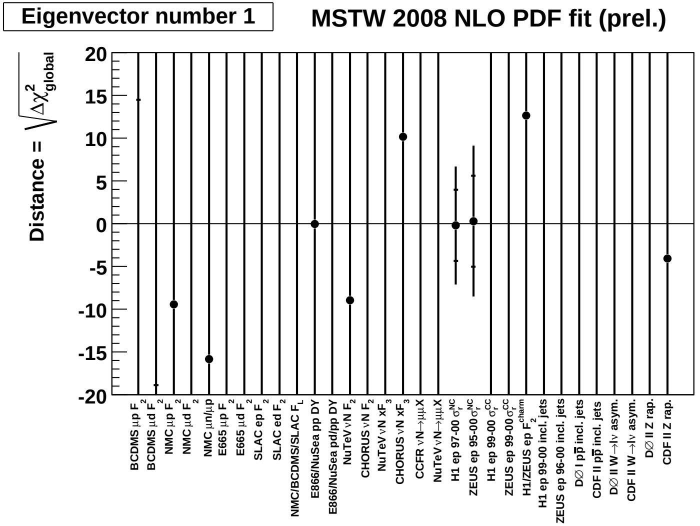
- Plot $(\chi^2 - \chi_0^2)/\chi_0^2$ versus the distance along a particular eigenvector.
- Define 90% C.L. region for each data set as

$$(\chi^2 - \chi_0^2)/\chi_0^2 < (\xi_{90} - \xi_{50})/\xi_{50}.$$

ξ_{90} is the 90th percentile of the χ^2-distribution with $N_{\rm pts.}$ d.o.f. $\xi_{50} \simeq N_{\rm pts.}$ is the most probable value.

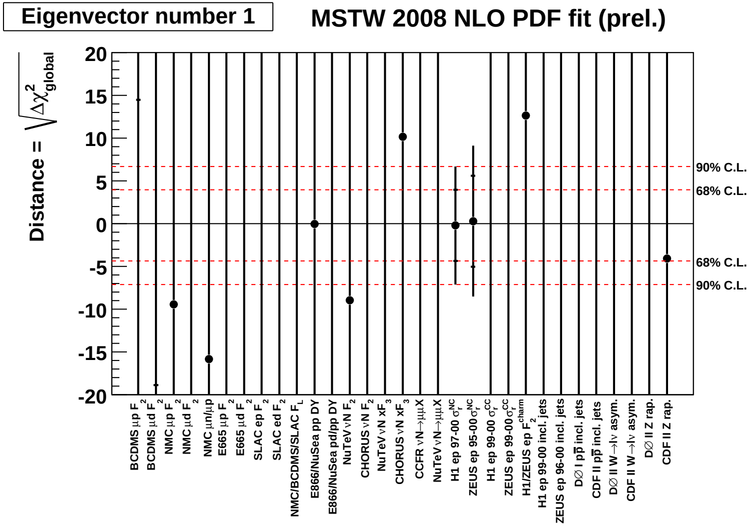
- Similarly for the 68% C.L.

Determination of tolerance for eigenvector number 1



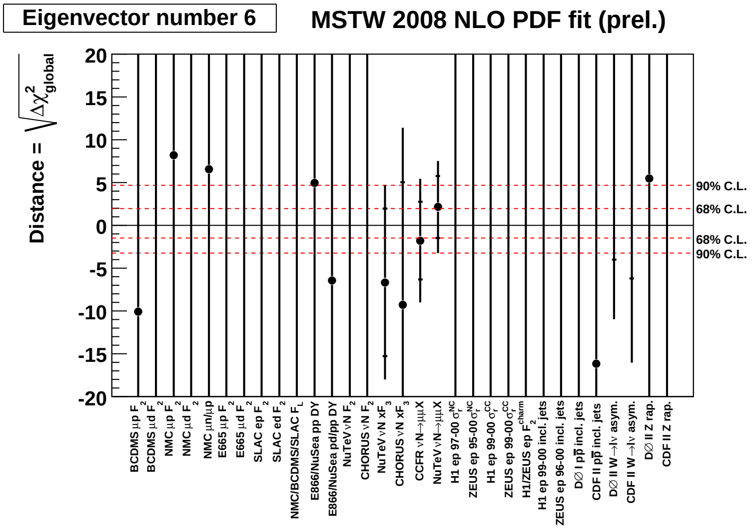
- Eigenvector direction sensitive to low-x gluon distribution.

Determination of tolerance for eigenvector number 1



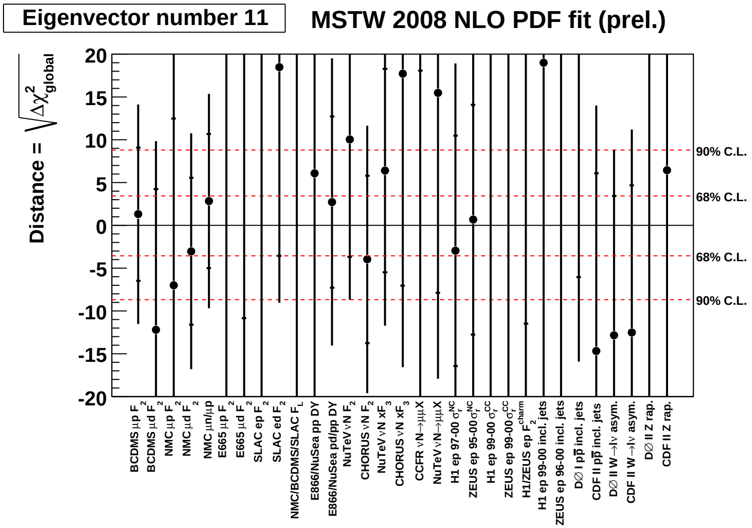
- Eigenvector direction sensitive to low-x gluon distribution.

Determination of tolerance for eigenvector number 6



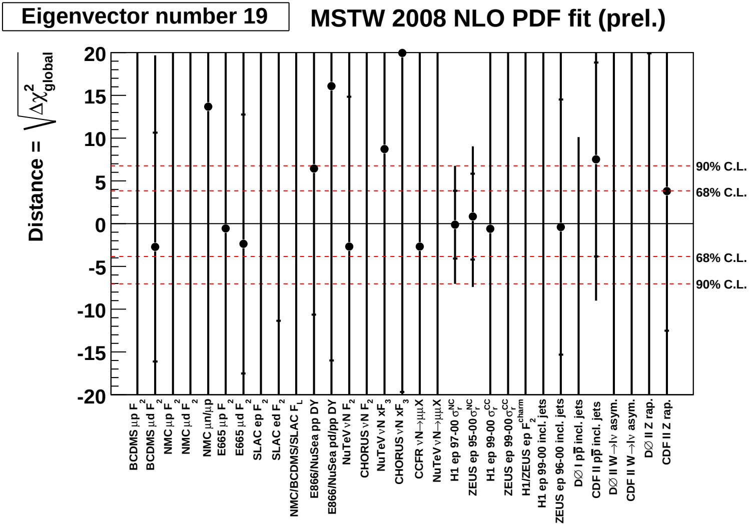
- Eigenvector direction sensitive to strange quark asymmetry.

Determination of tolerance for eigenvector number 11



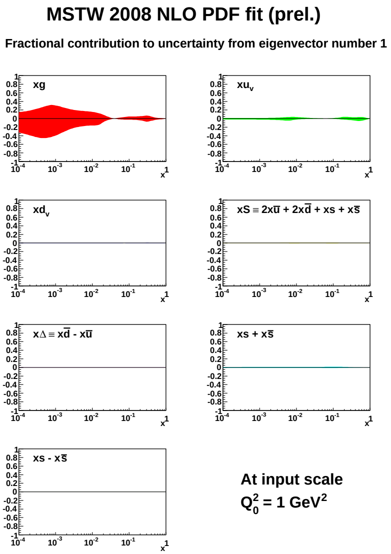
- Eigenvector direction sensitive to many parton flavours.

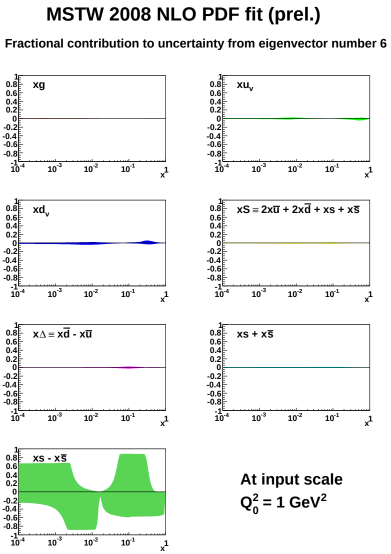
Determination of tolerance for eigenvector number 19



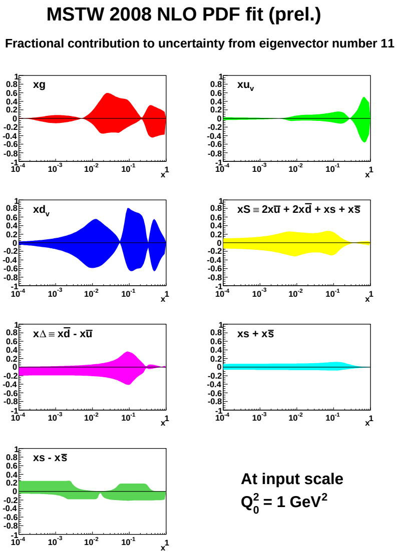
- Eigenvector direction sensitive to high-x gluon distribution.

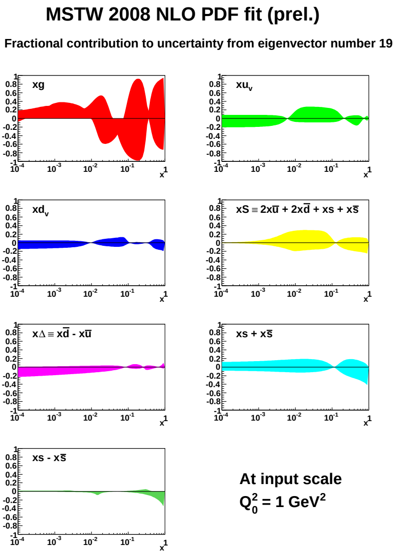
Contribution to PDF uncertainty from single eigenvector



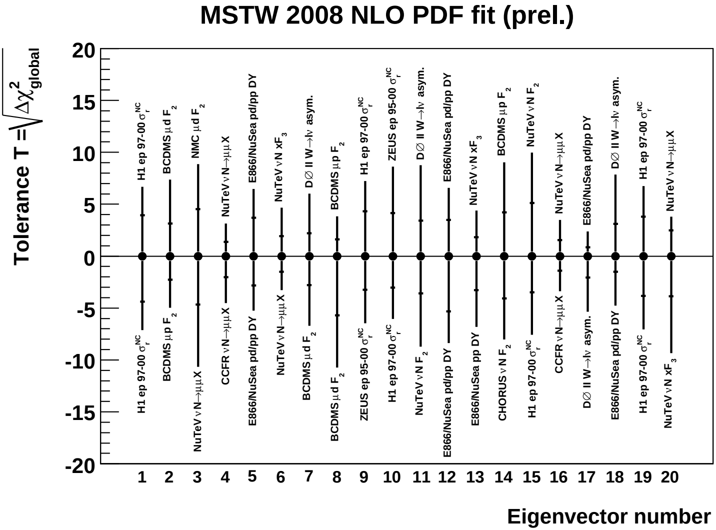


Contribution to PDF uncertainty from single eigenvector

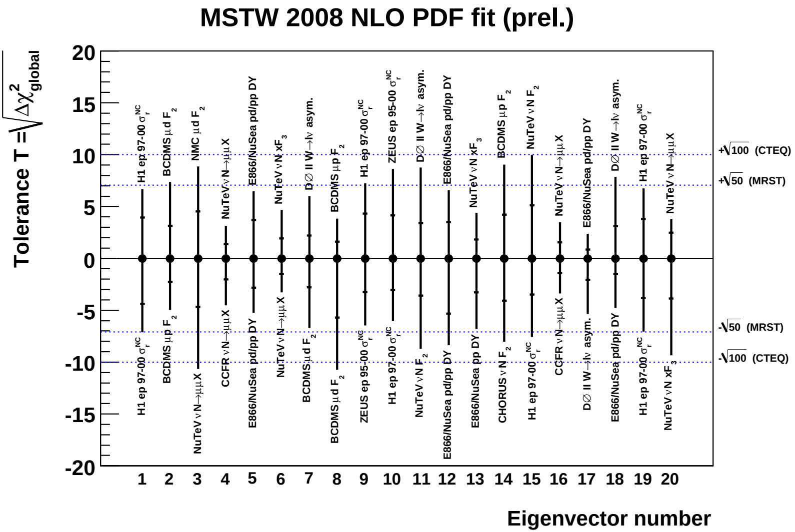




Tolerance vs. eigenvector number



Tolerance vs. eigenvector number



Summary

- CTEQ and MRST have so far used a fixed value of the tolerance $T = \sqrt{\Delta\chi^2_{\rm global}}$ in producing eigenvector PDF sets.
- Propose **dynamic** determination of tolerance: different for each eigenvector of the Hessian/covariance matrix.
- In general 90% C.L. given by $T \sim \sqrt{50}$. Close to MRST value. CTEQ tolerance ($T = \sqrt{100}$) too large?
- Smaller tolerance for some eigenvectors, e.g. strange quarks.

Outlook

- Will provide LO, NLO, NNLO (+ modified LO for MCs) PDFs, each with 40 additional eigenvector PDF sets.
- Will provide stand-alone FORTRAN, C++, MATHEMATICA interpolation code (in addition to inclusion in LHAPDF).
- **Timescale:** $\sim$ few weeks for publication and public release.

MSTW 2008 NLO (prel.) compared to MRST 2001 NLO