PDFs and Run 2 inclusive jets

CTEQ meeting at ANL (December 5–7, 2008)

Jon Pumplin – MSU

Outline

1. Run 1 inclusive jet updates
2. K-factors
3. Fits with Run 2 inclusive jet data
4. Theory prejudices: the example of strangeness
5. New technique for measuring the compatibility of data sets in a global fit and characterizing the influence of selected data sets

Improved handling of Run 1 jet data

Run 1 inclusive jet experiments are now calculated by integrating LO predictions over each p_T bin instead of using effective p_T bin centers that were inferred from the data. The NLO predictions are then made by applying K-factors (NLO/LO) for each data point from FastNLO using the central CTEQ6.6 fit in place of EKS.

Find good agreement between the old and new calculations for run 1.

Run 2 jet K-factors from FastNLO

1. K-factors go far from 1 – even negative – for $\mu = \mathrm{ET}/4$. Hence that choice of scale is not usable, because if the NLO correction is so large, the NNLO corrections must be also large. This is the reason MRST choose $\mu = \mathrm{ET}$ for the central choice, so the factor-of-2 range between $\mu = 0.5\,\mathrm{ET}$ and $\mu = 2.0\,\mathrm{ET}$ can be used to estimate uncertainty.

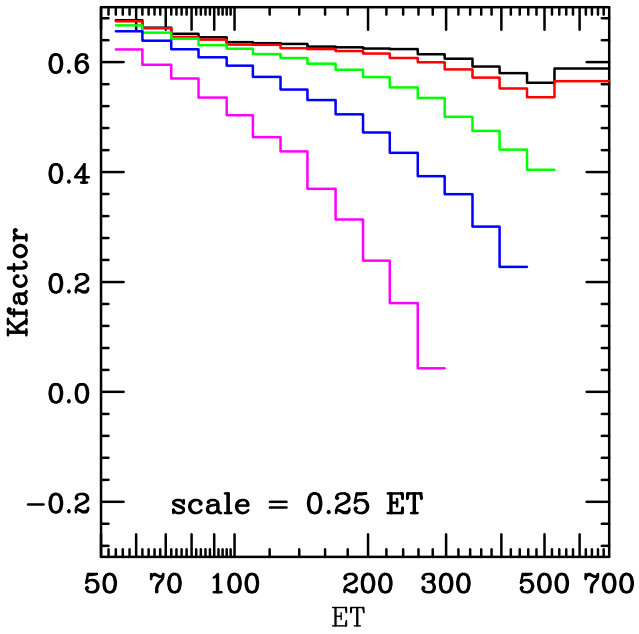
2. Have not compared these Run 2 results with EKS

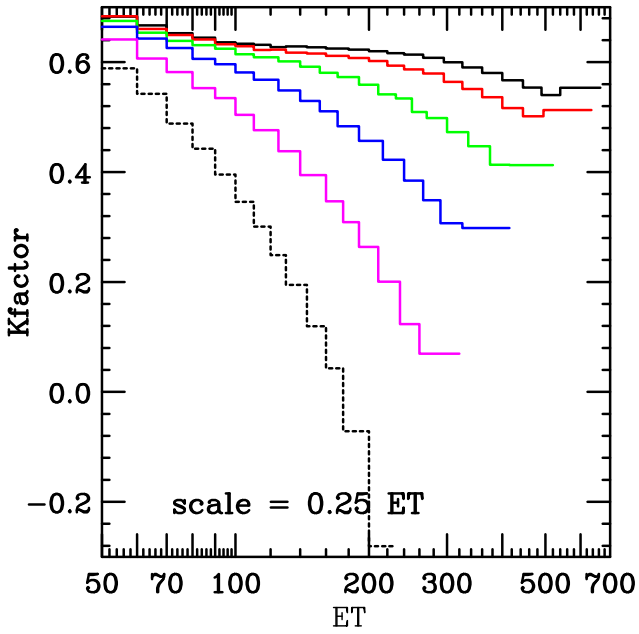
3. Using central CTEQ6.6 for K-factors, rather than computing PDF-dependent NLO/LO on the fly, as FastNLO intended.

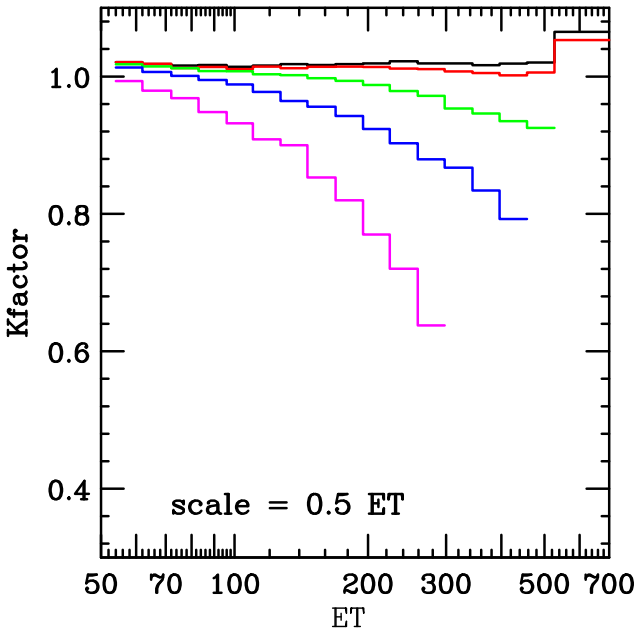
4. FastNLO offer "Midpoint (favored)" and "Rsep." I find "Midpoint" slightly favored by χ^2, but the difference is small.

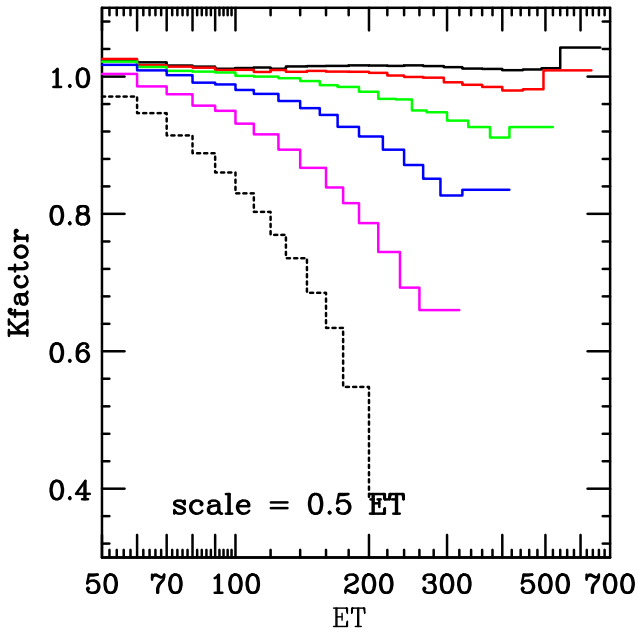
Figures on left are CDF, with Black=$(0.0 < y < 0.1)$, Red=$(0.1 < y < 0.7)$, Green=$(0.7 < y < 1.1)$, Blue=$(1.1 < y < 1.6)$, Magenta=$(1.6 < y < 2.1)$.

Figures on right are DØ, with Black=$(0.0 < y < 0.4)$, Red=$(0.4 < y < 0.8)$, Green=$(0.8 < y < 1.2)$, Blue=$(1.2 < y < 1.6)$, Magenta=$(1.6 < y < 2.0)$, Black dashed=$(2.0 < y < 2.4)$.



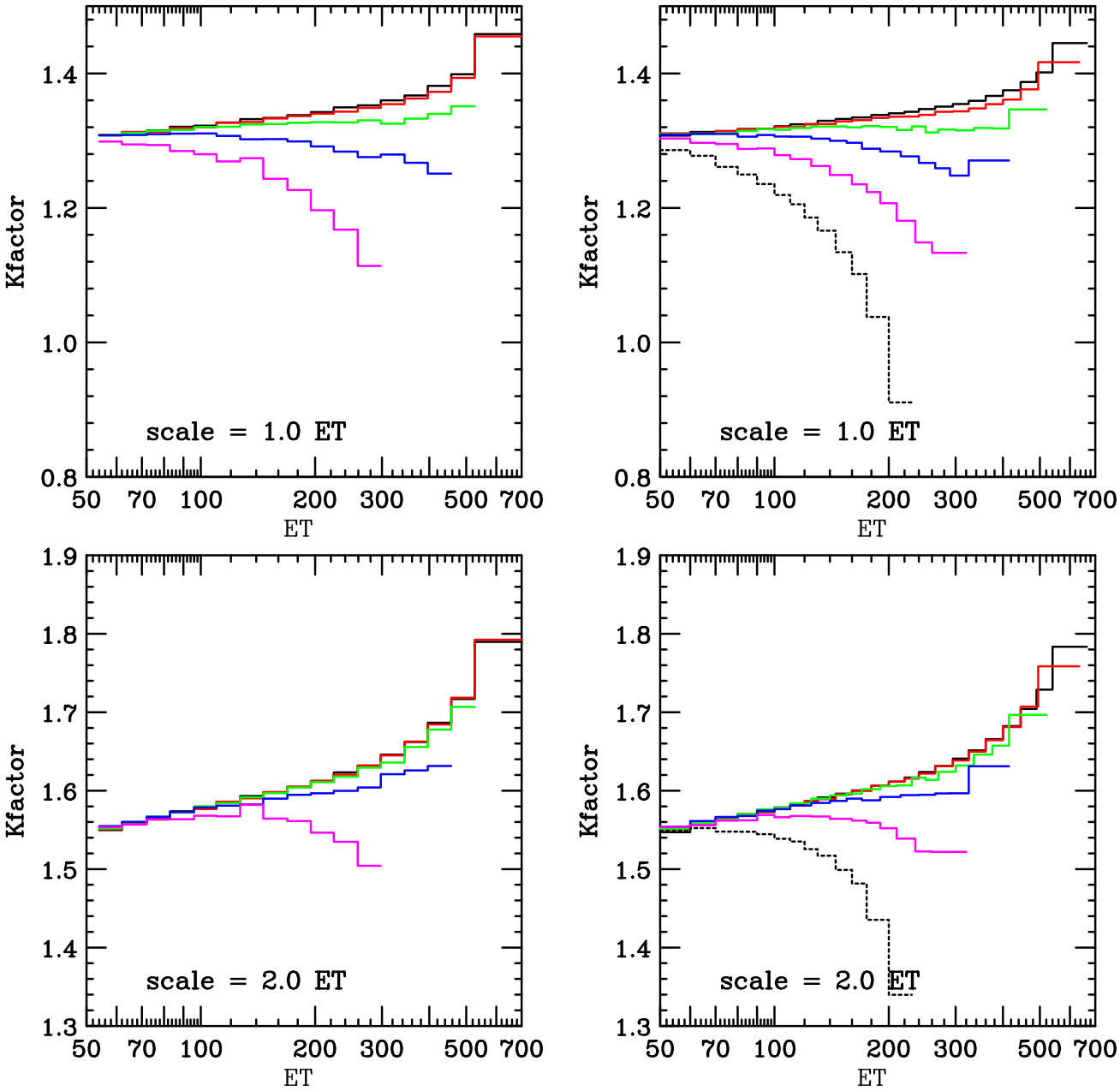






Figures on left are CDF, with Black=$(0.0 < y < 0.1)$, Red=$(0.1 < y < 0.7)$, Green=$(0.7 < y < 1.1)$, Blue=$(1.1 < y < 1.6)$, Magenta=$(1.6 < y < 2.1)$.

Figures on right are DØ, with Black=$(0.0 < y < 0.4)$, Red=$(0.4 < y < 0.8)$, Green=$(0.8 < y < 1.2)$, Blue=$(1.2 < y < 1.6)$, Magenta=$(1.6 < y < 2.0)$, Black dashed=$(2.0 < y < 2.4)$.

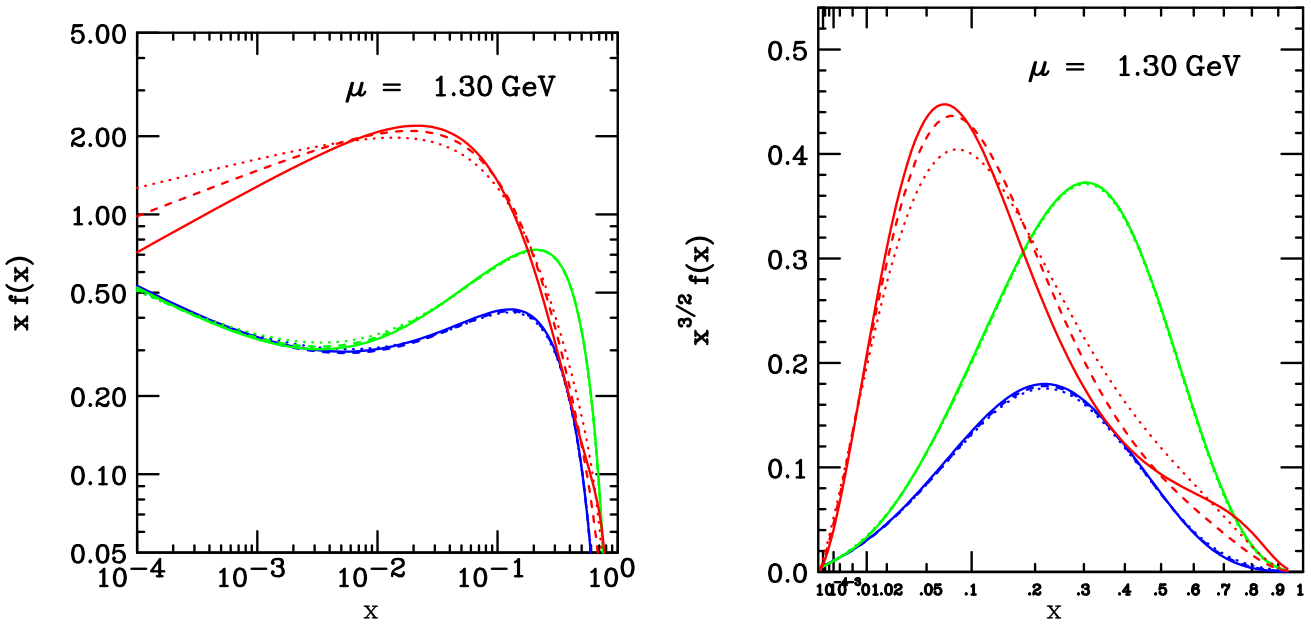


1. Have not (yet) tried the threshold resummation option of Owens et al., which is also an available option in FastNLO.
2. Progress from Soper and Olness on K-factor checking?

Initial fits with Run 2 jet data

1. Best fit has good weighted chisqr overall: 2758 for 2775 data points.
2. Chisqr/Npt OK for the run 2 data (we assume an additional 1% error for each data point)
3. CTEQ6.6 PDFs give weighted chisqr = 2787, only 29 higher, so CTEQ6.6 PDFs fit the run 2 data quite well.
4. Refitting (with weight 2) lowers chisqr for CDF from 101 to 89, and 126 to 121 for D0, a fairly small improvement.
5. Stronger weight for D0 doesn't do much to the fit—the fit to D0 is already quite good and can't be improved much.
6. Stronger weight for CDF improves its fit slightly, from 89 to 80 or 77.
7. Normalizations for the inclusive jet experiments are OK: N=0.983 for CDF, N=0.952 for D0, for the weight=2 fit.
8. Systematic error parameters are OK: a few are larger than 1.0, but none larger than 1.8

Solid: CTEQ6.6.
Dashed: fit with weight 2 for both CDF and D0
Dotted: fit with weight 20 for both CDF and D0
GREEN=u, BLUE=d, RED=g at $Q_0 = 1.3\,\mathrm{GeV}$



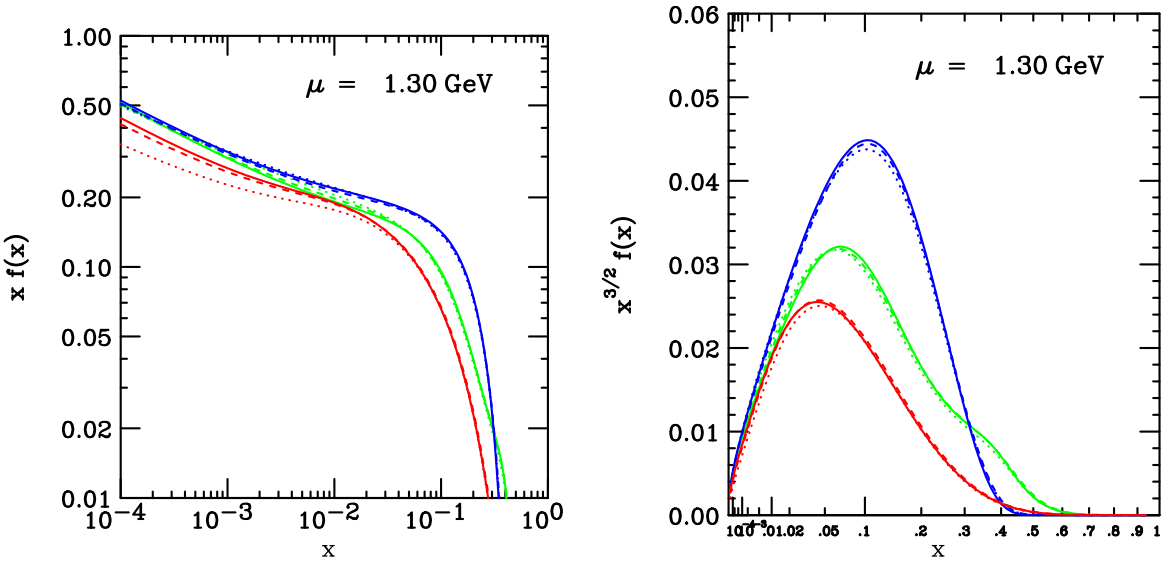
- "HJ" structure went away: No bump in the gluon at large x. When weight=2 is used for the jets, $g(x) < u(x)$ at large x. However, when weight=20 is used, gluon at large x becomes comparable to $u(x)$.
- The presence or absence of the "bump" is really a false issue – although the central cteq6.6 fit has one and the new central fit doesn't, acceptable fits of either type appear among the eigenvector sets of cteq6.6 and of the new fits.

Solid: CTEQ6.6.

Dashed: fit with weight 2 for both CDF and D0

Dotted: fit with weight 20 for both CDF and D0

GREEN=$\bar{u}$, BLUE=$\bar{d}$, RED=$s=\bar{s}$ at $Q_0 = 1.3\,\mathrm{GeV}$.

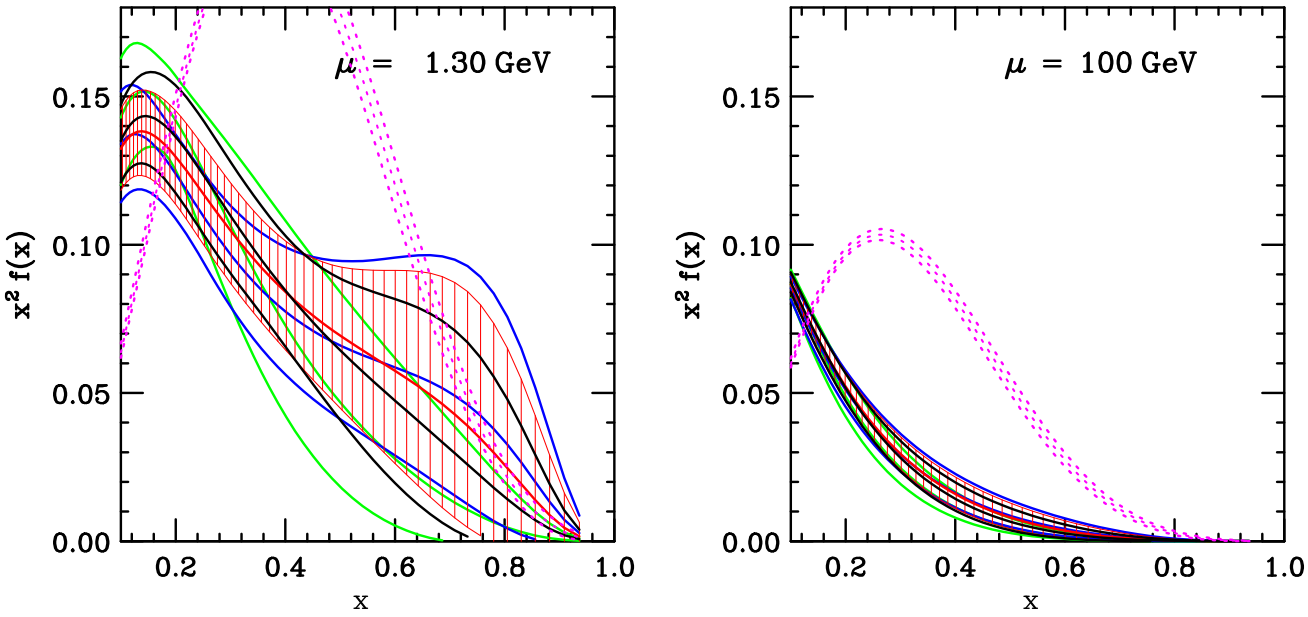


- No significant change in the sea quarks

Further fits with Run 2 jet data

The addition of the run 2 jet data does not reduce the gluon uncertainty very much, given the large increase in integrated luminosity. The reason for this is not that the run 1 and run 2 data disagree with each other, since fits that use only run 2 data have approximately the same uncertainty.

BLUE = CTEQ6.6 (run 1 jet data only)
RED = new fit (run 1 and run 2 jet data)
BLACK = new fit with run 2 jet data only
GREEN = new fit (*NO* jet data)
MAGENTA = up quark



Puzzling feature: large increase in quantity and quality of inclusive jet data has not particularly reduced the gluon uncertainty. Indeed, fit with no jet data at all (green) is quite narrow.

Perhaps this is simply a parametrization effect: All of these fits use fixed a_2 in $(1-x)^{a_2}$ for gluon — as done in CTEQ6.6 and all of our previous fits.

"valence-like" gluon? Question can't be answered at this time: good fits are possible with gluon

considerably larger or considerably smaller than up quark at large x for scale Q=1.3.

At Q=100, the quark dominates no matter what.

Unfinished business

We assume an additional 1% point-to-point error in the run 2 inclusive jet data. If this is supposed to represent error in the theory, e.g. NLO approximation, it should instead be a smooth parametrization.

Near-term data sets to be included in PDF fit

1. W rapidity asymmetry and lepton-from-W-decay rapidity asymmetry (with p_T cuts)

2. "Combined fitting:" PDFs and nonperturbative ResBos parameters. (Preliminary fits and eigenvector sets have already been made, but need to include the p_T distribution data for Z^0 production at the Tevatron.)

3. HERA run 2 data, when it becomes available.

Theoretical prejudices and Neural Nets

The NNPDF collaboration has become a significant player in the PDF game.

Obvious difference between that approach and ours:

1. We enforce smoothness in the input PDFs – limited number of parameters.
2. We make mild Regge-based assumptions on functional forms at $x \to 0$
3. We make mild Spectator Counting assumptions on functional forms and parameter ranges at $x \to 1$.
4. In finalizing CTEQ6.6, we went to considerable effort to reduce

$$\frac{s(x) + \bar{s}(x)}{\bar{d}(x) + \bar{u}(x)}$$

at small x. It is extremely easy to get fits where this ratio is as large as 2 at moderately small x. *Should we have a problem with that??*

How to measure consistency of global fit

(Work with John Collins in 2001:)

J. C. Collins and J. Pumplin, "Tests of goodness of fit to multiple data sets" [hep-ph/0105207].

Key idea: In addition to the

Hypothesis-testing criterion: $\Delta\chi^2 \sim \sqrt{2N}$

use the stronger

Parameter-fitting criterion: $\Delta\chi^2 \sim 1$

Parameters here are relative weights assigned to various experiments, or to results obtained using various experimental methods. Examples:

- Plot minimum χ^2_i vs. $\chi^2_{\rm tot} - \chi^2_i$, where χ^2_i is one of the experiments, or all data on nuclei, or all data at low Q^2,...

or

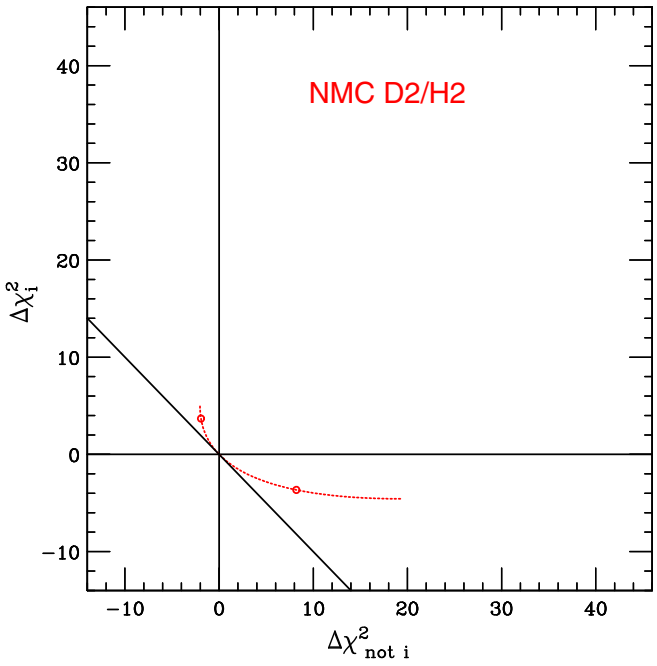
- Plot both as function of Lagrange multiplier u where $(1-u)\chi^2_i + (1+u)(\chi^2_{\rm tot} - \chi^2_i)$ is the quantity minimized.

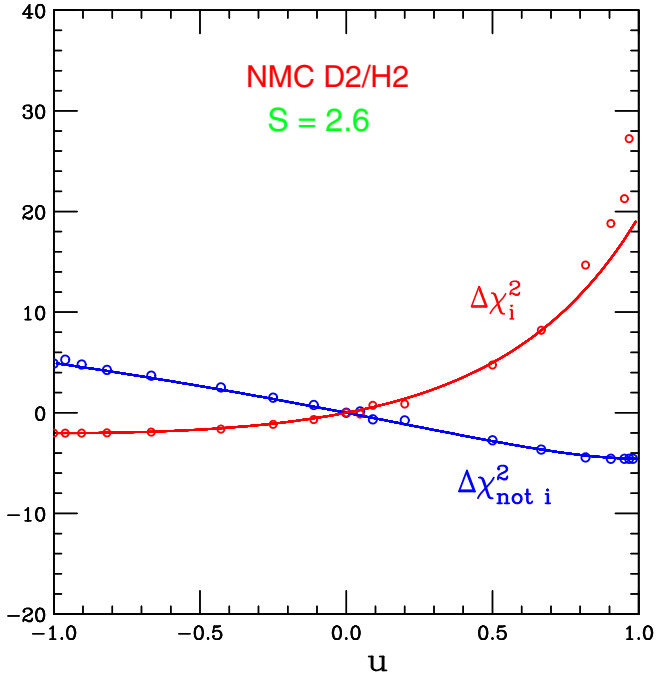
Can obtain quantitative results by fitting to a model with a single common parameter p:

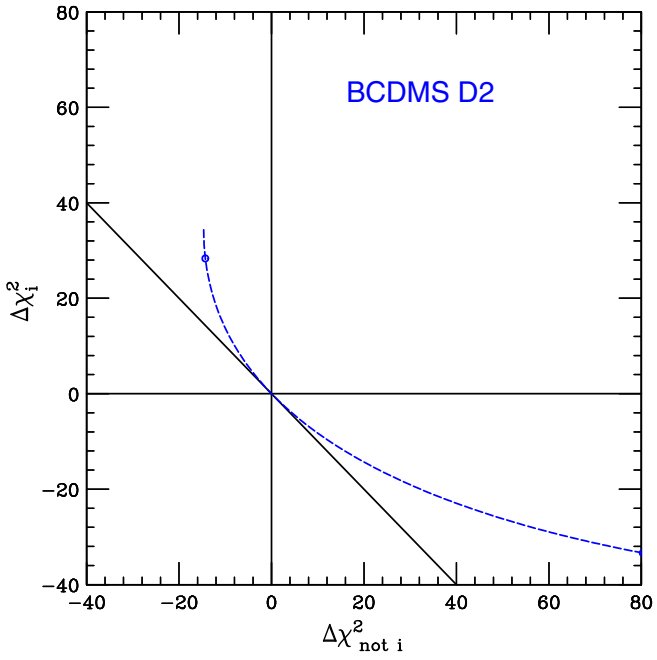
$$\chi^2_i = A + \left(\frac{p}{\sin\theta}\right)^2 \quad \Rightarrow \quad p = 0 \pm \sin\theta$$

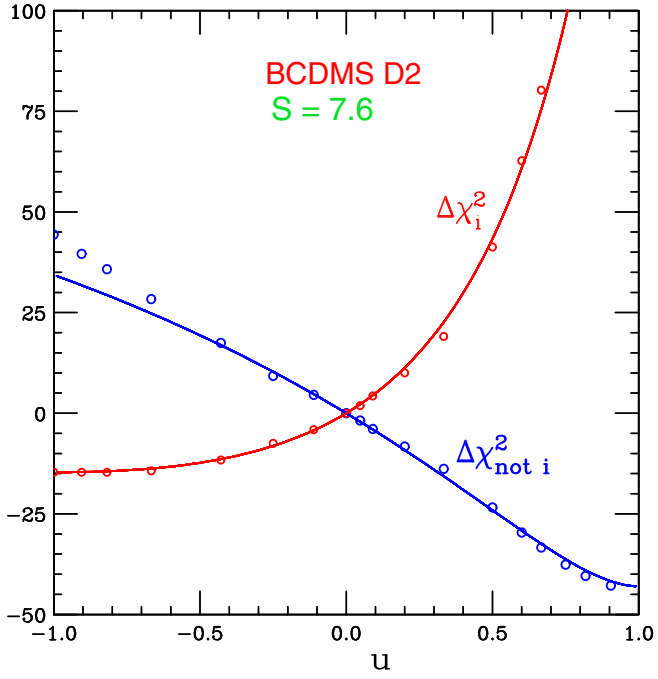
$$\chi^2_{{\rm not}\ i} = B + \left(\frac{p-S}{\cos\theta}\right)^2 \quad \Rightarrow \quad p = S \pm \cos\theta$$

These differ by $S \pm 1$, i.e., by S "standard deviations"









Fits to 8 of the experiments in the CTEQ5 analysis

Expt	1	2	3	4	5	6	7	8
S	2.7	3.3	3.3	4.2	5.3	7.6	7.4	8.3
$\tan\phi$	0.56	0.54	0.99	0.86	0.71	1.14	0.65	0.39

The situation John and I considered was to consider a subset S of the global data set — such as the data from a single experiment, or data of a single type (e.g. inclusive jets) — and its complement $\overline{S}$ which consists of the rest of the data.

By fitting to minimize $w\,\chi^2_S + \chi^2_{\overline{S}}$ for a number of values of the relative weight w, one can map out χ^2_S as a function of $\chi^2_{\overline{S}}$.

Two problems:

1. Since traditional Gaussian statistics don't apply to our problem because of unknown systematic errors (both in theory and in experiment), we still don't know how to decide whether a particular χ^2_S vs. $\chi^2_{\overline{S}}$ curve shows compatibility or incompatibility.
2. The method doesn't directly show what aspects of the theory are affected by the tension between S and $\overline{S}$ (although the PDF fits that were obtained to make the curve can be used to explore the "pull" associated with S).

A new method that I am currently studying appears to solve the second problem.

New method for studying compatibility

Near the minimum, χ^2 is an approximately quadratic function of the PDF shape parameters $A_1, \ldots, A_N$. By calculating the Hessian matrix

$$H_{i,j} = \frac{\partial^2 \chi^2}{\partial A_i \, \partial A_j}$$

which defines the quadratic form, and using its eigenvectors as new basis vectors in the N-dimensional space, we obtain a linear transformation that puts χ^2 into a very simple form:

$$\chi^2 = \chi_0^2 + \sum_{i=1}^{N} z_i^2 \qquad (1)$$

(In practice, it is necessary to carry this out by an iterative method, because the large range of eigenvalues of the Hessian — corresponding to a spectrum from steep to flat directions — requires different step sizes in different directions to avoid non-quadratic behavior when calculating the derivatives numerically.)

The linear transformation that leads to

$$\chi^2 = \chi_0^2 + \sum_{i=1}^{N} z_i^2$$

is not unique, because any further orthogonal transform of the z_i will preserve it. Such an orthogonal transformation can be defined using the eigenvectors of any symmetric matrix. After this second linear transformation of the coordinates, the chosen symmetric matrix will then be diagonal in the resulting new coordinates. Thus there is a freedom to diagonalize an additional matrix while preserving the simple form for χ^2. (In the standard Hessian method, and in my usual iterative procedure, the coordinate space distance measure

$$\sum_{i}^{N} (A_i - A_i^{(0)})^2$$

is made diagonal along with χ^2.)

This freedom can be exploited by taking the symmetric matrix from the quadratic form that describes the contribution to χ^2 from any chosen subset S of the data. Then. . .

In the quadratic approximation, this choice puts the contribution from the subset of data into a diagonal form:

$$\chi^2_S = \alpha + \sum_{i=1}^{N} (\beta_i z_i + \gamma_i z_i^2) \; .$$

If all of the parameters γ_i lie in the range $0 < \gamma_i < 1$, this leads by simple algebra to

$$\begin{aligned} \chi^2 &= \chi^2_S + \chi^2_{\overline{S}} + \text{const} \\ \chi^2_S &= \sum_{i=1}^{N} [(z_i - A_i)/B_i]^2 \\ \chi^2_{\overline{S}} &= \sum_{i=1}^{N} [(z_i - C_i)/D_i]^2 \end{aligned}$$

Thus the subset S of the data and its complement $\overline{S}$ take the form of independent measurements of the N variables z_i, with results

$$\begin{aligned} \mathbf{S}: \; z_i &= A_i \pm B_i \\ \overline{S}: \; z_i &= C_i \pm D_i \end{aligned}$$

$$\chi^2 = \chi^2_S + \chi^2_{\overline{S}} + \text{const}$$

$$\chi^2_S = \sum_{i=1}^{N} [(z_i - A_i)/B_i]^2$$

$$\chi^2_{\overline{S}} = \sum_{i=1}^{N} [(z_i - C_i)/D_i]^2$$

This decomposition answers the question "What is measured by data subset S?" — it is those parameters z_i for which the $B_i \lesssim D_i$. These parameters will generally span a subspace of the full N-dimensional fitting space, with various fractions of involvement along different directions.

The decomposition also measures the compatibility between S and the rest of the data $\overline{S}$: the disagreement between the two is

$$(A_i - C_i) \pm \sqrt{B_i^2 + C_i^2}$$

along the direction of z_i. Overall chi-squared form of difference is

$$\sum_{i=1}^{N} (A_i - C_i)^2/(B_i^2 + C_i^2)$$

In our PDF application, not all of the γ_i parameters will lie in the range $0 < \gamma_i < 1$.

Directions with $\gamma_i < 0$ correspond to parameters for which the data subset S has very little to say, so the value is determined entirely by the complementary subset $\overline{S}$.

Directions with $\gamma_i > 1$ correspond to linear combinations of the original parameters that are almost entirely deterimined by S, so the complementary subset $\overline{S}$ is irrelevant.

To measure the compatibility between S and $\overline{S}$, we can simply ignore the dimensions in which $0 < \gamma_i < 1$ fails.

Work on examining the compatibility and influence of the inclusive jet data on the PDF fit using this new technique is in progress.

Notes based on discussions

1. As noted by Liang, the studies of scale choice are inconsistent, since the K-factor tables were all based on the central CTEQ6.6 fit, which included run 1 inclusive jet data that was fitted using ONLY scale ET/2. To do it consistently, need to rerun FastNLO using the various PDF sets created using different scale choices, and iterate that.

2. Could also try to implement FastNLO as it was intended, where the K-factors are calculated in each evaluation of FCN.

3. It is disquieting to have the addition of inclusive jet data *expanding* the uncertainty range associated with the high-x gluon. This is a symptom that the parametrization is too confining. It may be enough to use free a_2(gluon); but perhaps a_5(gluon) is also needed. This could be constructed as an object lesson for the HERA-only fits, which use quite restrictive parametrizations.

Final remark

The parametrizations of scale effects as a theoretical systematic error shown by Olness and Soper will be useful, once they have been extended to include the various rapidity ranges of the data. The new theory parameters induced that way will need to be taken as additional search parameters (producing additional eigenvector sets) – can't treat them the way we do experimental systematic errors by finding the best-fit analytically (quadratic form), because the same theory parameters apply to both experiments.