Statistical Issues in the Measurement of Parton Distribution Functions

Jon Pumplin - Michigan State University

CTEQ-TEA collaborators: J. Huston, H.-L. Lai, P. Nadolsky, C.-P. Yuan

BIRS Workshop on Statistical Issues relevant to the Significance of Discovery Claims (Banff July 11–16, 2010)

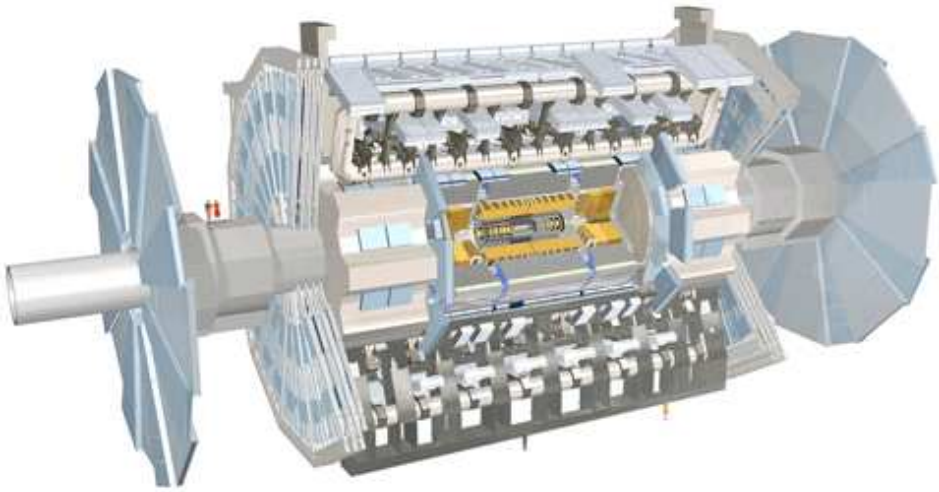
Outline:

1. Abstract view of PDF measurement
2. Tools for estimating PDF uncertainties
   - Lagrange Multiplier method
   - Hessian method
3. Major problem: enhanced uncertainties
4. Current practical methods
5. Cry for help: can the creative environment of BIRS invent a better procedure?

Proton Collider experiments

Experiments in particle physics use collisions of high energy protons (and/or antiprotons). New champion is the Large Hadron Collider (near Mont Blanc):

Protons with an energy of $7000\,\mathrm{m\,c^2}$ travel in both directions around the 27 km tunnel, and collide inside elaborate particle detectors:



Parton Distribution Functions

Protons are a complicated quantum mechanical system made up of a large (indeterminant) number of "partons": the quarks u, d, s, c, b, t; their corresponding antiquarks; and gluons.

(You may have thought that the proton is composed of just 3 quarks. But consider the hydrogen atom which you may imagine as made of just a proton and an electron. The Lamb Shift taught us that there are also photons and e^+e^- pairs in it, with probabilities that are small because $\alpha_{\rm EM}$ is small. The α_s of the quark and gluon interactions is large, so multiparton states in the proton have large probabilities.)

Theorems from the theory of partons (Quantum Chromodynamics) show that the most interesting "hard" interactions between two colliding protons are dominated by collisions in which a single parton from one proton strikes a a single parton from the other.

The calculations needed to interpret data from collider experiments thus depend on knowing the probability density functions (PDFs) $u(x)$, $\bar{u}(x)$, $d(x)$,... that describe the probability for finding a u, $\bar{u}$, d,... in the proton with fraction $0 < x < 1$ of the proton's momentum.

Technical detail: Parton evolution

The parton distributions actually depend on two variables: the fraction x of the momentum of the proton that the parton is carrying, and the QCD factorization scale μ which is the quantum-mechanical conjugate to the distance scale of the interaction.

But the variation with μ can be calculated by perturbation theory. Also c, b, t can be assumed negligible at a low scale; so only 6 or 7 functions of the single variable x need to be determined.

As an added bonus, the evolution to large μ – where the most immediately interesting LHC physics lies – tends to smooth out fluctuations in the PDFs at the scale where the parametrizations are made, so any artificial structure there is not so important.

Parametrizations

To extract functions of x from a finite set of measurements with finite errors would appear to be an ill-posed mathematical problem.

However, there is good reason to think that the PDFs should be very smooth functions, so it is possible to parametrize them. Quantum Chromodynamics helps this because any fine structure tends to get washed out under the evolution to hard scales.

Typical form: $a_0\, x^{a_1}\, (1-x)^{a_2}\, \exp(a_3 x + a_4 x^2 + a_5 \sqrt{x})$
This is sufficiently flexible to handle the accuracy of current data, while building in positivity and known or expected behaviors at $x \to 0$ and $x \to 1$.

Modern PDF analyses use a total of ~ 25 parameters to describe the full set of functions. The number of parameters that can be used increases with time, as new data become available to constrain them. Hence the analysis always has some parameters that are poorly constrained, which causes non-quadratic behavior of χ^2 close to the minimum.

Alternative parametrizations

To study and/or reduce the bias caused by choosing a form of parametrization, I have also recently been studying fits based on Chebyshev polynomials. These fits use 70 free parameters, with a penalty added to χ^2 to enforce smoothness.

Another alternative is to represent the PDF functions by Neural Networks, which allow a very large number of effective parameters — see Robert Thorne's talk.

Experimental input

The PDFs are "nonperturbative" features of proton structure, which cannot in practice be calculated from first principles (even though the widely accepted "standard model" should in principle be sufficient to determine them).

Hence the goal is to determine the PDF parameters by simultaneously fitting data from the wide variety of experiments that are sensitive to them. These include experiments using $e^{\pm} p$, $\mu^{\pm} p$, νp, $\bar{p} p$, and $p p$ scattering, with a wide variety of final states that are calculable by available techniques.

Currently, there are ~ 35 useful data sets with a total of ~ 3000 data points. The quality of fit to each data set is measured by a χ^2_j defined by

$$\chi^2_j = \sum_{i=1}^{N_j} \left(\frac{\text{data}_i - \text{theory}_i}{\text{error}_i} \right)^2$$

(except for refinements to incorporate published correlated systematic errors).

Further explanation of χ^2 definition

The goodness-of-fit to experiment j is measured by

$$\chi^2_j = \sum_{i=1}^{N_j} \left(\frac{\text{data}_i - \text{theory}_i}{\text{error}_i} \right)^2$$

as shown on the previous page.

data_i is an experimentally reported data point, which is based on on the experimenters counting the number of events that fall into a particular kinematic bin.

error_i is the experimentally reported error, which may come from Poisson statistics; but more often involves estimates of systematic effects, since experimenters tend to choose running time and bin sizes to get the statistical counting errors down to where they are comparable to estimates of unknown systematic errors.

theory_i is the theory prediction – which is based on a complicated QCD calculation, and hence is a non-linear function of the parameters $a_1, \ldots, a_{25}$ which parametrize the parton distributions that we are trying to determine by the fitting procedure.

Other choices for goodness of fit?

Our traditional measure of goodness-of-fit is given by

$$\chi_j^2 = \sum_{i=1}^{N_j} \Delta_i^2$$

where

$$\Delta_i = \frac{\text{data}_i - \text{theory}_i}{\text{error}_i}$$

(We call this quantitity χ^2 – which of course is not sufficient to guarantee that it obeys a chi-squared distribution.)

One could imagine preferring some other measure of Gof, such as for example

$$\chi_j^2 = \sum_{i=1}^{N_j} \log\left(1 + \Delta_i^2\right)$$

which is similar for $|\Delta_i| \lesssim 1$, but which more-or-less gives up on points that disagree by a lot. Is there any experience with this??

The PDF "Global Fitting" paradigm

For any choice of the PDF parameters

$$\boxed{a_1, \ldots, a_{25}}$$

we can compute the quality of fit to the data sets

$$\boxed{\chi^2_1, \ldots, \chi^2_{35}}$$

The goal is to find the PDF parameters $\{a_i\}$ that yield acceptable fits to the data, as characterized by $\{\chi^2_j\}$ at various desired levels of confidence.

The "obvious" procedure is to define

$$\chi^2 = \sum_{j=1}^{35} \chi^2_j$$

A Best Fit can be estimated by minimizing this χ^2 with respect to $a_1, \ldots, a_{25}$.

An Uncertainty Range can be found by accepting all PDF sets in $(a_1, \ldots, a_{25})$ space for which

$$\chi^2 < \chi^2_{\text{BestFit}} + \Delta\chi^2$$

with $\Delta\chi^2 = 1$ for 68.3% confidence, 2.71 for 90% confidence, 6.63 for 99% confidence, etc.

Trouble!

Up to this point, statisticians should be reasonably happy with our procedures.

The problem we hope BIRS can help with arises from the fact that the actual PDF uncertainties are much larger than what is given by the prescription

$$\text{“}\Delta\chi^2 = 1\text{”}$$

Before showing how we know that, I will introduce two useful tools:

- Lagrange Multiplier method
- Hessian method

Lagrange Multiplier method

To find the PDF uncertainty on predictions for some physical process, e.g., the cross section σ for Higgs production at the LHC, you vary $\{a_i\}$ to minimize

$$F = \chi^2 + \lambda\sigma$$

at a variety of values of the Lagrange Multiplier parameter λ.

In this way, you map out χ^2 as a function of σ.

That determines the uncertainty distribution for the predicted σ — if you can decide on the range of $\Delta\chi^2$ to accept.

Hessian method

In the neighborhood of its minimum at $(a_1^{(0)}, \ldots, a_{25}^{(0)})$, χ^2 has a quadratic form

$$\chi^2 = \chi^2_{\text{min}} + \sum_{ij} H_{ij} \, (a_i - a_i^{(0)}) \, (a_j - a_j^{(0)})$$

where H is the Hessian matrix (inverse error matrix).

Expressing the displacements $a_i - a_i^{(0)}$ as linear combinations of the eigenvectors of the real symmetric matrix $\mathbf{H}$ introduces new coordinates $z_1, \ldots, z_{25}$ such that

$$a_i = a_i^{(0)} + \sum_j w_{ij} \, z_j \, .$$

with χ^2 taking the very simple form

$$\chi^2 = \chi^2_{\text{min}} + \sum_i z_i^2$$

Using these coordinates, the region of PDF parameter space allowed by a given $\Delta\chi^2$ is simply the interior of a hypersphere.

Eigenvector Uncertainty sets

$$\chi^2 = \chi^2_{\rm min} + \sum_i z_i^2$$

implies that the region of PDF parameter space allowed at a given $\Delta\chi^2$ can be characterized by a collection of Eigenvector sets:

$$(z_1, z_2, z_3, \dots) = \begin{cases} (+U_1, 0, 0, \dots) \\ (-D_1, 0, 0, \dots) \\ (0, +U_2, 0, \dots) \\ (0, -D_2, 0, \dots) \\ \dots \end{cases} .$$

According to the quadratic approximation, $U_i = D_i = \sqrt{\Delta\chi^2}$. In practice, the individual values U_i and D_i are adjusted separately for each eigenvector set to compensate for non-quadratic behaviors and produce the desired $\Delta\chi^2$ exactly.

Eigenvector Uncertainty Sets are extremely useful, because they permit a simple calculation of the expected uncertainty range for any observable. The method is to compute the allowed deviation from the best fit by adding the deviations allowed along each eigenvector direction in quadrature—separately for positive and negative deviations to compute asymmetric errors.

How we know that $\Delta\chi^2$ must be large

1. Inconsistency between individual subsets of data and the rest of the data: DSD method.

2. Inconsistency between individual subsets of data and the best fit: Distribution of S_j.

3. Uncertainties caused by parametrization choice.

Each of these points will now be discussed in detail.

Data Set Diagonalization

Partition the data into two subsets:

$$\chi^2 = \chi^2_S + \chi^2_{\overline{S}} \, .$$

Subset S can be any one of the experiments, or all experiments of a particular type that might be suspected of an untreated systematic error; while $\overline{S}$ is all the rest of the data.

The DSD method will answer the questions

1. What does subset S measure?

2. Is subset S consistent with the rest of the data?

The essential trick is that in the Hessian method, the linear transformation that leads to

$$\chi^2 = \chi^2_0 + \sum_{i=1}^{N} z_i^2$$

is not unique, because any further orthogonal transform of the z_i will preserve it. Such a transformation can be defined using the eigenvectors of the quadratic form corresponding to χ^2_S. Then . . .

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

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

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

This answers "What is measured by subset S?" — it is the parameters z_i for which the $B_i \lesssim D_i$. The fraction of the measurement of z_i contributed by S is

$$\gamma_i = D_i^2/(B_i^2 \,+\, D_i^2)\,.$$

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

$$\sigma_i = |A_i - C_i|/\sqrt{(B_i^2 + C_i^2)}\,.$$

Experiments that provide at least one measurement with $\gamma_i > 0.1$

Process	Expt	N	$\sum_i \gamma_i$
$e^+ p \to e^+ X$	H1 NC	115	2.10
$e^- p \to e^- X$	H1 NC	126	0.30
$e^+ p \to e^+ X$	H1 NC	147	0.37
$e^+ p \to e^+ X$	H1 CC	25	0.24
$e^- p \to \nu X$	H1 CC	28	0.13
$e^+ p \to e^+ X$	ZEUS NC	227	1.69
$e^+ p \to e^+ X$	ZEUS NC	90	0.36
$e^+ p \to \nu X$	ZEUS CC	29	0.55
$e^+ p \to \bar{\nu} X$	ZEUS CC	30	0.32
$e^- p \to \nu X$	ZEUS CC	26	0.12
$\mu p \to \mu X$	BCDMS F_2p	339	2.21
$\mu d \to \mu X$	BCDMS F_2d	251	0.90
$\mu p \to \mu X$	NMC F_2p	201	0.49
$\mu p/d \to \mu X$	NMC F_2p/d	123	2.17
$p\,\mathrm{Cu} \to \mu^+\mu^- X$	E605	119	1.52
$pp, pd \to \mu^+\mu^- X$	E866 pp/pd	15	1.92
$pp \to \mu^+\mu^- X$	E866 pp	184	1.52
$\bar{p}p \to (W \to \ell\nu)X$	CDF I Wasy	11	0.91
$\bar{p}p \to (W \to \ell\nu)X$	CDF II Wasy	11	0.16
$\bar{p}p \to \mathrm{jet}\,X$	CDF II Jet	72	0.92
$\bar{p}p \to \mathrm{jet}\,X$	D0 II Jet	110	0.68
$\nu Fe \to \mu X$	NuTeV F_2	69	0.84
$\nu Fe \to \mu X$	NuTeV F_3	86	0.61
$\nu Fe \to \mu X$	CDHSW	96	0.13
$\nu Fe \to \mu X$	CDHSW	85	0.11
$\nu\,\mathrm{Fe} \to \mu^+\mu^- \mathrm{X}$	NuTeV	38	0.68
$\bar{\nu}\,\mathrm{Fe} \to \mu^+\mu^- \mathrm{X}$	NuTeV	33	0.56
$\nu\,\mathrm{Fe} \to \mu^+\mu^- \mathrm{X}$	CCFR	40	0.41
$\bar{\nu}\,\mathrm{Fe} \to \mu^+\mu^- \mathrm{X}$	CCFR	38	0.14

Total of $\sum \gamma_i = 23$ is close to the actual number of fit parameters.

Consistency tests

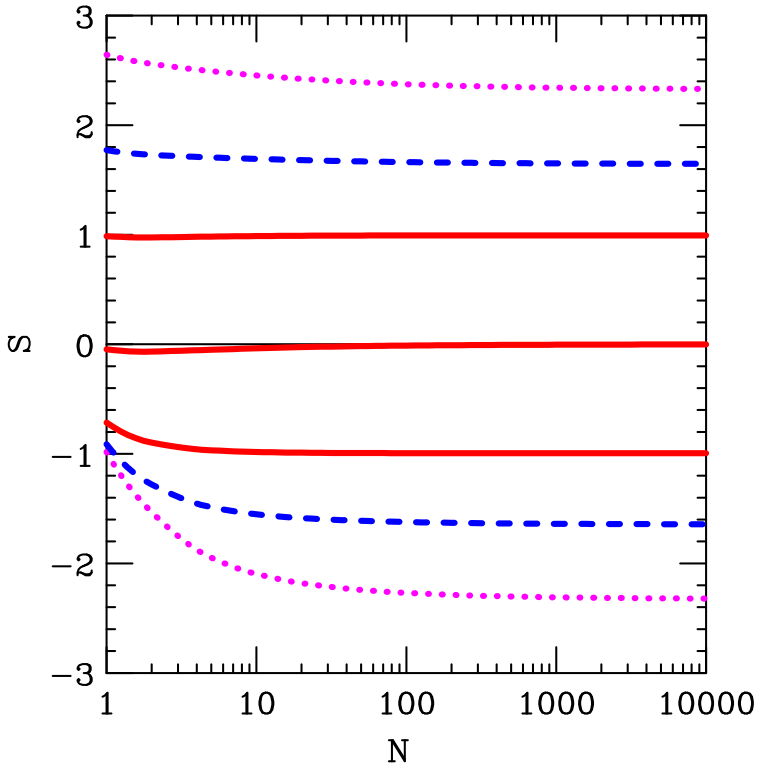
Measurements that conflict strongly with the others ($\sigma_i > 3$) are shown in red. There are lots of them!

Expt	$\sum_i \gamma_i$	(γ_1, σ_1), (γ_2, σ_2), ...
H1 NC	2.10	(0.72, 0.01) (0.59, 3.02) (0.43, 0.20) (0.36, 1.37)
H1 NC	0.30	(0.30, 0.02)
H1 NC	0.37	(0.21, 0.06) (0.16, 0.83)
H1 CC	0.24	(0.24, 0.00)
H1 CC	0.13	(0.13, 0.00)
ZEUS NC	1.69	(0.45, 3.13) (0.42, 0.32) (0.35, 3.20) (0.29, 0.80) (0.18, 0.64)
ZEUS NC	0.36	(0.22, 0.01) (0.14, 1.61)
ZEUS CC	0.55	(0.55, 0.04)
ZEUS CC	0.32	(0.32, 0.10)
ZEUS CC	0.12	(0.12, 0.02)
BCDMS F_2p	2.21	(0.68, 0.50) (0.63, 1.63) (0.43, 0.80) (0.34, 4.93) (0.13, 0.94)
BCDMS F_2d	0.90	(0.32, 0.67) (0.24, 2.49) (0.19, 2.09) (0.16, 5.22)
NMC F_2p	0.49	(0.20, 4.56) (0.17, 4.76) (0.12, 0.50)
NMC F_2p/d	2.17	(0.61, 1.11) (0.56, 3.60) (0.43, 0.90) (0.36, 0.79) (0.21, 1.41)
E605 DY	1.52	(0.91, 1.29) (0.38, 1.12) (0.23, 0.31)
E866 pp/pd	1.92	(0.88, 0.57) (0.69, 1.15) (0.35, 1.80)
E866 pp	1.52	(0.75, 0.04) (0.39, 1.79) (0.23, 1.94) (0.14, 3.57)
CDF Wasy	0.91	(0.57, 0.33) (0.34, 0.51)
CDF Wasy	0.16	(0.16, 2.84)
CDF Jet	0.92	(0.48, 0.47) (0.44, 3.86)
D0 Jet	0.68	(0.39, 1.70) (0.29, 0.76)
NuTeV F_2	0.84	(0.37, 2.75) (0.29, 0.42) (0.18, 0.97)
NuTeV F_3	0.61	(0.30, 0.50) (0.16, 1.35) (0.15, 0.30)
CDHSW	0.13	(0.13, 0.04)
CDHSW	0.11	(0.11, 1.32)
NuTeV	0.68	(0.39, 0.31) (0.29, 0.66)
NuTeV	0.56	(0.32, 0.18) (0.24, 2.56)
CCFR	0.41	(0.24, 1.37) (0.17, 0.12)
CCFR	0.14	(0.14, 0.79)

Only measurements that play a significant role ($\gamma_i > 0.1$) are listed.

Quality of fits to individual data sets

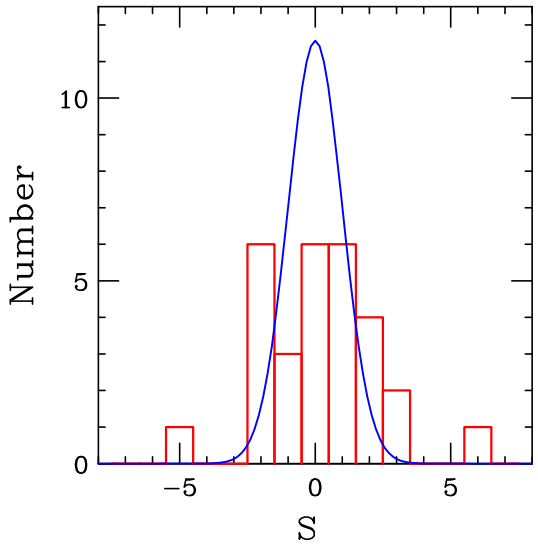
Look at fit quality using equivalent normal distributions. Simple way (R. Fisher, 1925), $S = \sqrt{2\,\chi^2} - \sqrt{2N-1}$ has Mean ≈ 0 and SD ≈ 1, and is nearly gaussian-distributed down to N as small as 10, which is the smallest number of points in any of our data sets:



Values of S corresponding to central probability 68% (solid), 90% (dashed), and 98% (dotted).

S values can be used to conveniently compare fit quality among experiments with different numbers of data points.

S-values for the 29 data sets in CT10



Smooth curve is gaussian with mean 0 and standard deviation 1 – essentially the statistical prediction. The observed histogram is broader than this prediction, which demonstrates that the uncertainties are not strictly gaussian.

The extreme outlier at negative S is CCFR F3, for which we improperly added systematic errors in quadrature, artificially reducing χ^2.

The extreme outlier at positive S is NMC μp data – indicates "higher twist" effects??

Even without those two outliers, the distribution appears to be broader than absolute Gaussian (though I have not made a quantitative study of how inconsistent it is).

Other indications that PDF uncertainties are larger than "$\Delta\chi^2 = 1$"

1. Space dependence: Different analysis groups obtain significantly different results

2. Time dependence: Results from a single analysis group change significantly as a result of minor changes in the procedure, such as which data sets are included.

3. Parametrization dependence: Results from a single analysis group can change significantly as a result of changes in the parametrizations, with χ^2 changing by ~ 100.

Paradox emphasized by Louis Lyons: The PDF fits have $\chi^2/N \approx 1$. In detail, a typical recent fit (CT10) has $\chi^2 = 3015$ for 2753 data points with 26 fitting parameters. That is somewhat higher than the $N \pm \sqrt{2N}$ expected range:

$$(2753 - 26) \pm \sqrt{2\,(2753 - 26)} = 2650\,\text{to}\,2800$$

But expanding all of the experimental errors uniformly by only 4% would reduce χ^2 to $3015/1.04^2 = 2788$ which lies within that range.

Current CTEQ-TEA procedure

(1) We allow up to $\Delta\chi^2 = 100$ in determining the eigenvector uncertainty sets, which we estimate to yield a 90% confidence limit.

This does **NOT** increase the uncertainty range by as much as you might think

(which would be $\sqrt{100/2.71} = 6$),

because over this range, χ^2 generally rises much faster than quadratic in most directions; and because the definition used for χ^2 also includes the penalty discussed on the next page.

The faster-than-quadratic rise of our measure of fit quality implies that our results are fairly insensitive to the choice $\Delta\chi^2 = 100$ – e.g. 50 would give very similar results.

Current CTEQ-TEA procedure

(2) As we move away from the minimum of total χ^2, we don't want to allow the fit to any individual data set to deteriorate too far. (This can especially be a problem for data sets with a small number of points.)

So we include a penalty in the effective χ^2 of

$$\sum_{i=1}^{35} S_i^A \, \theta(S_i)$$

where $A = 16$ and $S_i = \sqrt{2\,\chi_i^2} - \sqrt{2N_i - 1}$ is the standard deviation-like measure introduced earlier.

The large power $A = 16$ halts displacement away from the minimum rather sharply when any one of the data sets starts to complain.

My collaborator Liang Lai likes to make an analogy between the two components of our effective χ^2 and the House+Senate components of the US congress (Google the Great Compromise 1787 for details).

Parameter fitting vs. Hypothesis testing

The current MSTW method goes farther in this phenomenological direction, by dropping all criteria on total χ^2, and instead just halting the displacement along each eigenvector direction when the fit to any one experiment is first stretched beyond a desired confidence point.

In terminology John Collins and I introduced (arXiv:hep-ph/0106173), this is a variety of the "hypothesis testing" point of view, according to which any PDF fit that provides a satisfactory fit to all of the data sets is accepted – in contrast to the standard statistical criteria for Parameter Fitting based on an allowed tolerance in $\Delta\chi^2$.

Is there any support for such a procedure from a statistical point of view?

(The χ_i^2s discussed here both for CTEQ and MSTW are scaled if necessary to account for the fact that some data sets can never be fit within 90% confidence – not surprisingly since there are ~ 35 of them.)

Monte Carlo approach

The NNPDF collaboration use an alternative approach: generate "fake" data sets by shifting each measured value randomly according to its reported experimental error. Fitting these fake sets gives a collection of fits whose range is hoped to measure the true uncertainty.

Any comments on this validity of this procedure?

NNPDF also use neural network methods to avoid parametrization dependence; but that brings in some other problems. I am tempted to try the approach using the conventional parametrization methods.

Summary

- Workable methods to extract PDFs and their uncertainty range from data are available.

- These methods are based on intuitive approaches that might benefit from a more rigorous statistical approach.

- It is established that applying simple textbook Gaussian statistics to this task would badly underestimate the uncertainties. But it is not clear how much of that is due to

  1. Unreported systematic errors in the data
  2. Systematic errors in the theory (NLO, no nuclear corrections, choice of m_c, etc.)
  3. Effects of parametrization dependence

- Are there known methods, e.g. statistical bootstrap, that we should be trying?

- Can the Banff spirit of creativity bring us new methods to solve the important problem of PDF uncertainties in particle physics?