Parton Distribution Functions

- PDFs describe a "simple" 1-body aspect of the nonperturbative quark and gluon structure of the proton.
- PDFs are a necessary input to make predictions for experiments at LHC.
- Because the PDFs are universal, they can be measured by a QCD Global Analysis in which many different types of experiment contribute.

The focus of this talk will be on details of the Global Analysis procedure that should be examined carefully during the workshop. Because of the short time available, current results on the PDFs themselves will not be discussed.

The PDF fitting paradigm

1. Parameterize the x-dependence of each flavor at a fixed small Q_0
2. Compute PDFs $f_a(x, Q_0)$ at all $Q > Q_0$ by DGLAP
3. Compute cross sections for DIS(e,μ,ν), Drell-Yan, Inclusive Jets,... using QCD perturbation theory
4. Compute "χ^2" measure of agreement between predictions and measurements:

$$\chi^2 = \sum_i \left(\frac{\text{data}_i - \text{theory}_i}{\text{error}_i} \right)^2$$

5. Varying the shape parameters $\{A_i\}$ to minimize χ^2 yields Best Fit PDFs: CTEQ6.1, MRST,...
6. Define a PDF Uncertainty Range as the region in $\{A_i\}$ space for which χ^2 is sufficiently close to its minimum value.
7. Make results on Best Fit and representative uncertainty sets, available to consumers.

Parametrization at Q_0

We parameterize the x-dependence of each flavor at a fixed small Q_0:

$$x f_a(x, Q_0) = A_0\, x^{A_1}\, (1-x)^{A_2}\, e^{A_3 x} (1 + e^{A_4} x)^{A_5}$$

This form was chosen as a simple generalization of the traditional $A_0\, x^{A_1}\, (1-x)^{A_2}$ parametrization, which in turn is based on Regge and quark counting rule ideas. The generalization consists of adding a 1:1 Padé approximation:

$$\frac{d}{dx} \ln (xf) = \frac{A_1}{x} - \frac{A_2}{1-x} + \frac{c_3 + c_4 x}{1 + c_5 x}\,.$$

We use $Q_0 = 1.3\,\text{GeV}$ and parametrize g, u_v, d_v, $\bar{d} + \bar{u}$, $\bar{d}/\bar{u}$, $s + \bar{s}$, $(s - \bar{s})/(s + \bar{s})$. Some of the "shape parameters" $\{A_i\}$ are frozen at arbitrary values to leave ~ 20 free fitting parameters. More free parameters would lead to unstable fits even with the extensions to Minuit that will be discussed below.

Issues for study:

- Does parametrization dependence influence any of the predictions of interest?
- Other functional forms?
- Negative gluon distribution at low Q?
- Intrinsic b or c?

PDF Evolution in Q

The parton distributions $f_a(x, Q)$ at all $Q > Q_0$ are derived from the parametrized $f_a(x, Q_0)$ using the DGLAP renormalization group evolution equations. Issue for study: Is Non-DGLAP behavior important at small x?

One approach, already started by MRST, is to examine the consistency of small-x data with the rest of the global fit.

Another approach to measuring the internal consistency will be described at the end of this talk.

Compute Cross Sections in QCD

Issue for study: Importance and feasibility of upgrading to NNLO.

(For charm production, just upgrading to full NLO is still in progress.)

χ^2 with systematic errors

Known experimental systematic errors can be included in a straightforward way. The systematic error parameters are determined as part of the fitting process.

The traditional definition

$$\chi^2 = \sum_{i=1}^{N} \frac{(D_i - T_i)^2}{\sigma_i^2} \qquad \left\{ \begin{array}{l} D_i = \text{ data} \\ T_i = \text{ theory} \\ \sigma_i = \text{ ``expt. error''} \end{array} \right.$$

is optimal for random Gaussian errors,

$$D_i = T_i + \sigma_i r_i \qquad \text{with} \qquad P(r) = \frac{e^{-r^2/2}}{\sqrt{2\pi}}.$$

With systematic errors,

$$D_i = T_i(A_1, \ldots) + \alpha_i r_{\text{stat},i} + \sum_{k=1}^{K} r_k \beta_{ki} \, .$$

The fitting parameters are $\{A_i\}$ (theoretical model) and $\{r_k\}$ (corrections for systematic errors).

To take into account the systematic errors, we define

$$\chi'^2(A_\lambda, r_k) = \sum_{i=1}^{N} \frac{\left(D_i - \sum_k r_k \beta_{ki} - T_i\right)^2}{\alpha_i^2} + \sum_k r_k^2,$$

and minimize with respect to $\{r_k\}$. The result is

$$\widehat{r}_k = \sum_{k'} \left(A^{-1}\right)_{kk'} B_{k'}, \qquad \text{(systematic shift)}$$

where

$$\begin{aligned} A_{kk'} &= \delta_{kk'} + \sum_{i=1}^{N} \frac{\beta_{ki}\beta_{k'i}}{\alpha_i^2} \\ B_k &= \sum_{i=1}^{N} \frac{\beta_{ki}\left(D_i - T_i\right)}{\alpha_i^2}. \end{aligned}$$

The $\widehat{r}_k$'s depend on the PDF model parameters $\{A_\lambda\}$. We can solve for them explicitly since the dependence is quadratic.

$\{\widehat{r}_k\}$ are are the optimal corrections for systematic errors: systematic shifts to be applied to the data points to bring the data from different experiments into compatibility within the framework of the theoretical model.

Minimize χ^2

To determine the PDFs, the fitting parameters $\{A_i\}$ should minimize χ^2. This minimization is nontrivial because

1. there are lots of parameters
2. new parameters are included in the fit until it is barely stable, to reduce the dependence on parametrization assumptions. Hence the χ^2 surface is quadratic only very close to the minimum.
3. the parameters are highly correlated
4. evaluation of χ^2 for a single choice of $\{A_i\}$ takes a several seconds.

I have made some extensions to the classic Minuit to make it more effective in this situation.

Minimize χ^2 - ctd

In the neighborhood of the minimum, χ^2 can be approximated by a quadratic form

$$\chi^2 = \chi_0^2 + \sum_{ij} H_{ij} \, (A_i - A_i^{(0)}) \, (A_j - A_j^{(0)})$$

where the Hessian matrix H is the inverse of error matrix. It is convenient to express this in a diagonal form by making use of the eigenvectors of H:

$$\chi^2 = \chi_0^2 + \sum_i z_i^2$$

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

If the $\{A_i^{(0)}\}$ are not quite at the minimum, there are also linear terms in χ^2 which can easily be used to improve the estimate of the minimum. I use an iterative procedure to home in on the minimum and to achieve the diagonal form for χ^2. By the end of the iteration, χ^2 is probed in all directions at the appropriate scale of $\Delta\chi^2$.

Eigenvalues of distance

The diagonal form for χ^2 is not unique, because a further arbitrary orthogonal transformation of the z_i coordinates is allowed. My usual method is to use this freedom to diagonalize the total distance moved in the original shape parameter space:

$$\sum_i (A_i - A_i^{(0)})^2 = \sum_i d_i^2 \, z_i^2$$

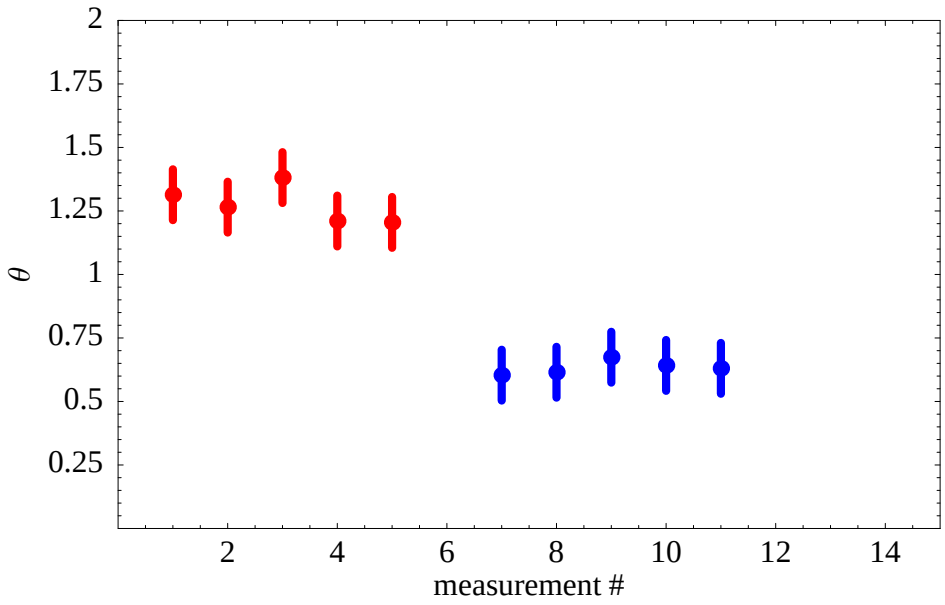
The distances d_i turn out to range from ~ 3 ("flat directions" in which χ^2 rises slowly) to $\sim 10^{-3}$ ("steep directions" in which χ^2 rises rapidly.) This difference corresponds to a $10^7 : 1$ range in the eigenvalues $(1/d_i^2)$ of the Hessian matrix, which is another way to see why the minimization is difficult. Issue for study: The code I use for finding the eigenvector directions and for exploring χ^2 along those directions while taking account of non-quadratic behavior along the flat directions could be made available—perhaps as a formal addition to the Minuit command set.

PDF Uncertainty ranges

The Uncertainty Range of the PDFs can be defined as the region in $\{A_i\}$ space for which χ^2 is sufficiently close to its minimum value: $\chi^2 < \chi^2_0 + T^2$.

In an ideal statistical world, the allowed range would be $\Delta\chi^2 \lesssim 1$. A variety of approaches over the last several years, based on studying conflicts between the data sets that make up the global fit, have shown that the empirical uncertainty range for present day PDFs is more like $\Delta\chi^2 \lesssim 50$.

The essential reason for this is a familiar one in phenomenology. Suppose the quantity θ is measured by two different experiments, or extracted using two different approximations to the True Theory.



What would you quote as the Best Fit and the Uncertainty? (Perhaps you would scale up the errors so the uncertainty range covers both data sets; or perhaps you would expand the uncertainty range even more, by taking the difference between these sets as a measure of the uncertainty.)

The discrepancies in the global fit are not as obvious as this because they only appear when different types of experiments are combined. But they can be studied by varying relative weights assigned to subsets of the data in the global fit; or by my new Hessian technique.

Another approach is the Statistical Bootstrap Method, in which you assign random weights to the experiments and use the variation in best fits as the measure of uncertainty.

Uncertainty Example: counting valence quarks

The number of valence up and down quarks is normally constrained in our global fits to the Standard Model values $N_u = 2$, $N_d = 1$, where

$$N_u = \int_0^1 [u(x) - \bar{u}(x)]\, dx$$

$$N_d = \int_0^1 [d(x) - \bar{d}(x)]\, dx$$

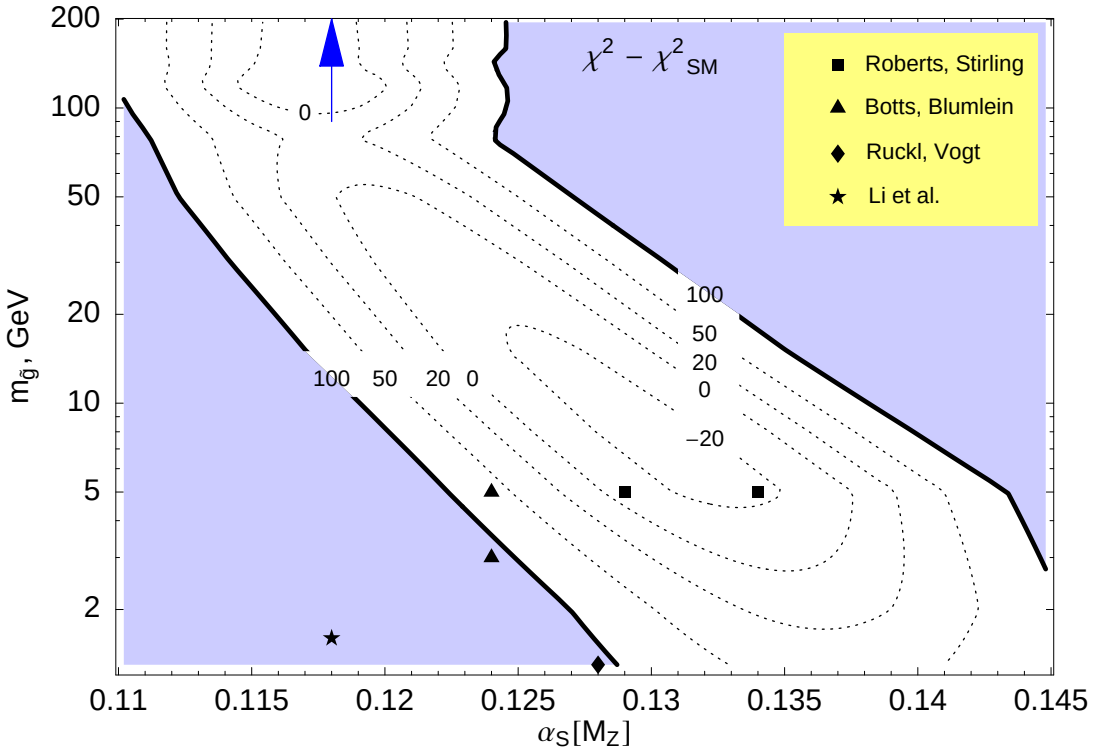
If N_u and N_d are made free parameters, the Best Fit has $N_u = 2.08$ and $N_d = 1.11$, with "improvement" in χ^2 of 4.0 .

This is to be interpreted as a nice demonstration of consistency with the standard model—not as a ($\sim 90\%$-confidence) anomaly.

Uncertainty example: Light Gluino

(Work in progress with Pavel Nadolsky, Fred Olness, Ed Berger.)

Hypothesizing a gluino of mass $\sim 10\,\mathrm{GeV}$ can improve the global fit by ~ 25 units in χ^2.
You may wish to take this as an intriguing hint of possible New Physics. But you would be crazy to consult a statistical table of χ^2 probabilities and declare it inescapable.



Dissemination of results

Representative PDF sets that explore the allowed $\Delta\chi^2 \lesssim T^2$ region can be generated by the "Lagrange Multiplier" method (wherein a specific quantity of interest such as the predicted σ_{Higgs} is minimized or maximized.

Alternatively a collection of PDF sets can be obtained from the "Hessian" method. An example are the published 40 PDF sets of CTEQ6, which are defined by $\Delta\chi^2 = 100$ along the eigenvector directions of the error matrix.

Large numbers of PDF sets that are of interest to users can be made available conveniently via the LHAPDF accord.

Issue for study: Considerations of convenience and/or speed in this protocol?

Hessian Method to Study consistency of the global fit

(Work in progress)

Partition the data into two subsets:

$$\chi^2 = \chi^2_I + \chi^2_{II}$$

where subset I might be, for example,

- any single experiment
- all of the jet experiments
- all of the low-Q data points (to look for higher twist effects)
- all of the low-x data points (to look for BFKL effects)
- all experiments that use deuteron corrections

Using the freedom to make an additional orthogonal transformation after the total χ^2 has been diagonalized, it is always possible to write

$$\chi^2 = \chi_0^2 + \sum_i z_i^2$$

$$\chi_I^2 = A + \sum_i B_i z_i + \sum_i c_i z_i^2$$

By simple algebra, this can be written as

$$\chi_I^2 = \sum_i \left(\frac{z_i - p_i}{q_i} \right)^2 + \text{const}$$

$$\Rightarrow z_i = p_i \pm q_i$$

In this way we can answer the question "How many parameters are significantly determined by any given data set?"

Similarly,

$$\chi^2_{II} = \chi^2_0 - A - \sum_i B_i z_i + \sum_i (1 - c_i) z_i^2$$

can be written as

$$\chi^2_{II} = \sum_i \left(\frac{z_i - r_i}{s_i} \right)^2 + \text{const}$$

$$\Rightarrow z_i = r_i \pm s_i$$

By comparing this with the result $z_i = p_i \pm q_i$ from subset I, we can awswer the question "How consistent are the data points I with the remainder II of the global fit?"

We just need to see if $(p_i - q_i) \pm \sqrt{r_i^2 + s_i^2)}$ is consistent with 0.