PDF reweighting in global fits

Contribution to LPC workshop
"Confronting Theory with Experiment: Puzzles, Challenges and Opportunities in the LHC Era"

(Fermilab Nov 17–18, 2011)

Jon Pumplin - Michigan State University
John Collins - Pennsylvania State University

The goal of this talk is to explain how PDF reweighting would be done in the traditional Hessian method for PDF fitting. The result appears to contradict the method used in recent NNPDF work.

1. Review of the Hessian method
2. Influence of a new data set
3. Monte Carlo implementation
4. Contrast with NNPDF prescription
5. Explicit example

Hessian method: Best Fit

The PDFs are characterized by "shape parameters" $a_1, \ldots, a_N$ that parametrize the x-dependence for each flavor at a chosen evolution starting scale μ_0.

($N \sim 25$ in current CTEQ fits; but N as large as 80 has been used to study parametrization dependence effects.)

Best-Fit values $a_1^{(0)}, \ldots, a_N^{(0)}$ are found by minimizing χ^2 for a fit to a large "global" set of data (~ 3000 data points).

Hessian method: Uncertainty range

In the neighborhood of the minimum, 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 H introduces new coordinates $z_1, \ldots, z_N$ 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 PDF determination can be thought of as a measurement of N uncorrelated variables with result

$$z_i = 0 \pm 1 \quad (i = 1, \ldots, N)$$

New experiment with one new data point

Prior to including a new experiment, our knowledge of the PDFs is described by the probability distribution

$$\frac{dP}{dz_1 \cdots dz_N} = \text{const} \times \exp[-(z_1^2 + \ldots + z_N^2)/2] \quad (1)$$

i.e.,

$$z_1 = 0 \pm 1, \ldots, z_N = 0 \pm 1 \, .$$

A single new data point will measure some linear combination of the parameters z_1,...,z_N. But Eq.(1) is invariant to an arbitrary orthogonal transformation of the $\{z_i\}$. We can use that freedom to redefine the $\{z_i\}$ so that the new measurement is sensitive only to z_1. Hence without loss of generality, the new measurement can be assumed to have the form

$$z_1 = A \pm B \, .$$

Including one new data point

Assuming Gaussian statistics, we can combine the new measurement with the old using the standard formula from freshman physics lab:

$$z_1 = 0 \pm 1, \quad z_1 = A \pm B$$

leads to

$$z_1 = \frac{\frac{1\cdot 0}{1^2} + \frac{1\cdot A}{B^2}}{\frac{1}{1^2} + \frac{1}{B^2}} \pm \sqrt{\frac{1}{\frac{1}{1^2} + \frac{1}{B^2}}}\,.$$

This result corresponds to

$$\left(\frac{dP}{dz_1 \cdots dz_N}\right)_{\text{new}} = \text{const} \times \exp(-\chi^2/2) \times \left(\frac{dP}{dz_1 \cdots dz_N}\right)_{\text{old}}$$

where

$$\chi^2 = \left(\frac{z_1 - A}{B}\right)^2$$

with

$$\left(\frac{dP}{dz_1 \cdots dz_N}\right)_{\text{old}} = \text{const} \times \exp[-(z_1^2 + \ldots + z_N^2)/2]$$

New experiment with two data points

Prior to including a new experiment, our knowledge of the PDFs is described by the probability distribution

$$\frac{dP}{dz_1 \cdots dz_N} = \text{const} \times \exp[-(z_1^2 + \ldots + z_N^2)/2]$$

i.e.,

$$z_1 = 0 \pm 1, \ldots, z_N = 0 \pm 1 \, .$$

Using the freedom to make an orthogonal transformation of the Hessian parameters $z_1,\ldots,z_N$, we can assume without loss of generality that an experiment which measures two new data points tells us

$$z_1 = A \pm B$$

and

$$z_1 \cos\theta + z_2 \sin\theta = C \pm D$$

where the parameter θ describes the extent to which the two data points measure the same or different aspects of the PDFs.

Two data points – continued

For the first new data point, the measurements $z_1 = 0 \pm 1$ and $z_1 = A \pm B$ can be combined as done previously to yield $z_1 = E \pm F$.

Defining $\tilde{z}_1 = (z_1 - E)/F$ with $\tilde{z}_2 = z_2$, $\tilde{z}_3 = z_3$, ... restores the symmetric form

$$\frac{dP}{d\tilde{z}_1 \cdots d\tilde{z}_N} = \text{const} \times \exp[-(\tilde{z}_1^2 + \ldots + \tilde{z}_N^2)/2] \, .$$

The information from the other new data point can then be included by the same elementary means. The result is that the two new points

$$z_1 = A \pm B, \quad z_1 \cos\theta + z_2 \sin\theta = C \pm D$$

are found to modify the original probability distribution to

$$\left(\frac{dP}{dz_1 \cdots dz_N}\right)_{\text{new}} = \text{const} \times \exp(-\chi^2/2) \times \left(\frac{dP}{dz_1 \cdots dz_N}\right)_{\text{old}}$$

where

$$\chi^2 = \left(\frac{z_1 - A}{B}\right)^2 + \left(\frac{z_1 \cos\theta + z_2 \sin\theta - C}{D}\right)^2$$

New experiment with n data points

The result derived here explicitly for a new experiment with 1 or 2 data points can be generalized to the case of a new experiment with n data points. The result is that the original probability distribution

$$\left(\frac{dP}{dz_1 \cdots dz_N}\right)_{\text{old}} = \text{const} \times \exp[-(z_1^2 + \ldots + z_N^2)/2]$$

gets multiplied by $e^{-\chi^2_{\text{new}}/2}$, where χ^2_{new} is just the usual chisquared measure of agreement with the new data set for the PDFs defined by $z_1,\ldots,z_N$.

In principle, this result could be used to facilitate the inclusion of new data sets in traditional PDF analysis. There is no strong incentive for the PDF collaborations like CTEQ to do that, because it is easy enough to redo the minimization and Hessian error analysis from scratch for the enlarged data set; and that method is superior to the extent that χ^2 depends somewhat non-quadratically on the fitting parameters.

However, the result becomes important theoretically when we consider a Monte Carlo method for implementing the error analysis.

Monte Carlo Method

PDF uncertainties on a prediction such as a Higgs cross section could be obtained by an obvious Monte Carlo method: one could generate a large number of PDF sets from the probability distribution

$$\left(\frac{dP}{dz_1 \cdots dz_N}\right) = \text{const} \times \exp[-(z_1^2 + \ldots + z_N^2)/2] \, .$$

Then compute the Higgs cross section for each of the generated configurations. The mean and standard deviation of these computed cross sections would directly yield the central value and 1-sigma limits of the prediction.

The effect of including a new data set in the global fit could be found in principle by computing χ^2_{new} for the new data set for each of the generated configurations, and keeping that configuration with probability $\exp(-\chi^2_{\text{new}}/2)$. That procedure is not promising as a practical method, however, since χ^2_{new} will generally be on the order of the number of new data points, so if there are many new data points, only a tiny fraction of the original PDF samples would be retained.

NNPDF paradox

The NNPDF method also produces a large sample of PDFs whose distribution is supposed to model the probability distribution corresponding to the knowledge that can be extracted from the input data sets. That sample can be used to predict central values and uncertainties in the same manner as the Hessian Monte Carlo method sketched here.

However, the rule for modifying the Monte Carlo sample to incorporate a new experiment is claimed to be to keep configurations according to a probability proportional to

$$(\chi^2_{\rm new})^{n-1}\ \exp(-\chi^2_{\rm new}/2)$$

where n is the number of points in the new data set.

The extra factor $(\chi^2_{\rm new})^{n-1}$ would be wonderful if it is correct, because it circumvents the practical difficulty that without it, almost all of the original PDF samples are rejected.

However, we do not see how this factor can be correct, in view of its conflict with the simple calculation described here in the context of the Hessian method.

Explicit example

Suppose the "old" data measures two uncorrelated variables:

$$z_1 = 0 \pm 1, \quad z_2 = 0 \pm 1 \,.$$

Hence the "old" probability distribution is

$$\frac{dP}{dz_1\, dz_2} = \text{const} \times \exp[-(z_1^2 + z_2^2)/2] \,.$$

Suppose the "new" data consists of two measurements:

$$z_1 = 1 \pm 1, \quad z_1 + z_2 = 1 \pm 1 \,.$$

The fit to all four constraints has

$$\chi^2 = \left(\frac{z_1 - 0}{1}\right)^2 + \left(\frac{z_2 - 0}{1}\right)^2 + \left(\frac{z_1 - 1}{1}\right)^2 + \left(\frac{z_1 + z_2 - 1}{1}\right)^2$$

The Hessian approach leads to the transformation

$$z_1 = \frac{3}{5} + \sqrt{\frac{1}{5}} u_1 + \sqrt{\frac{1}{5}} u_2$$

$$z_2 = \frac{1}{5} + \frac{1}{2}\left(1 - \sqrt{\frac{1}{5}}\right) u_1 + \frac{1}{2}\left(-1 - \sqrt{\frac{1}{5}}\right) u_2$$

After the transformation (whose Jacobian is a constant), we have

$$\chi^2 = \frac{3}{5} + u_1^2 + u_2^2$$

Hence

$$u_1 = 0 \pm 1 \quad \text{and} \quad u_2 = 0 \pm 1$$

$$\frac{dP}{du_1\, du_2} = \text{const} \times \exp[-(u_1^2 + u_2^2)/2] \,.$$

The ratio of the new probability density to the old is proportional to

$$\frac{\exp[-(u_1^2 + u_2^2)/2]}{\exp[-(z_1^2 + z_2^2)/2]}$$

By explicit calculation, this is equal to $\exp[-\chi^2_{\text{new}}/2]$ where χ^2_{new} is just the expected contribution from the two new data points:

$$\chi^2_{\text{new}} = \left(\frac{z_1 - 1}{1}\right)^2 + \left(\frac{z_1 + z_2 - 1}{1}\right)^2 \,.$$

Conclusion

Hence in this simple example, refitting the combined data set is explicitly found to be equivalent to reweighting the probability distribution from the original data set by

$$\exp(-\chi^2_{\rm new}/2)$$

without any additional factor of

$$(\chi^2_{\rm new})^{n-1} \, .$$

The ensemble of PDFs created in the NNPDF method is constructed in a different way than the Hessian method: each member of the ensemble can be thought of as a best fit to a "fake" data set whose error bars are the same as the real data, with the central values shifted randomly according to those errors. But – even after a number of clarifying e-mail exchanges – we do not see how the extra weight factor $(\chi^2_{\rm new})^{n-1}$ can be correct.