Data Set Diagonalization in Global fitting

Jon Pumplin – Michigan State University

PDF4LHC meeting (CERN 29 May, 2009)

A "Data Set Diagonalization" technique lets us measure the compatibility between a subset S of the data, e.g.

- data from a single experiment
- all the data that use nuclear targets
- all the data from low Q where higher twist corrections might play a role

and the remainder of the global data set $\overline{S}$.

This technique can also be used to determine which aspects of the global fit are determined by a specific subset of the input data.

Preliminary results of a study of the internal compatibility of the CT09 data set will be shown.

Old method (2001)

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

In addition to the

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

use the stronger

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

The parameter being fitted here is the relative weight assigned to subset S. We minimize the weighted χ^2 for a series of values of the weight. We can then

- plot minimum χ_S^2 vs. $\chi_{\overline{S}}{}_{2}$

or

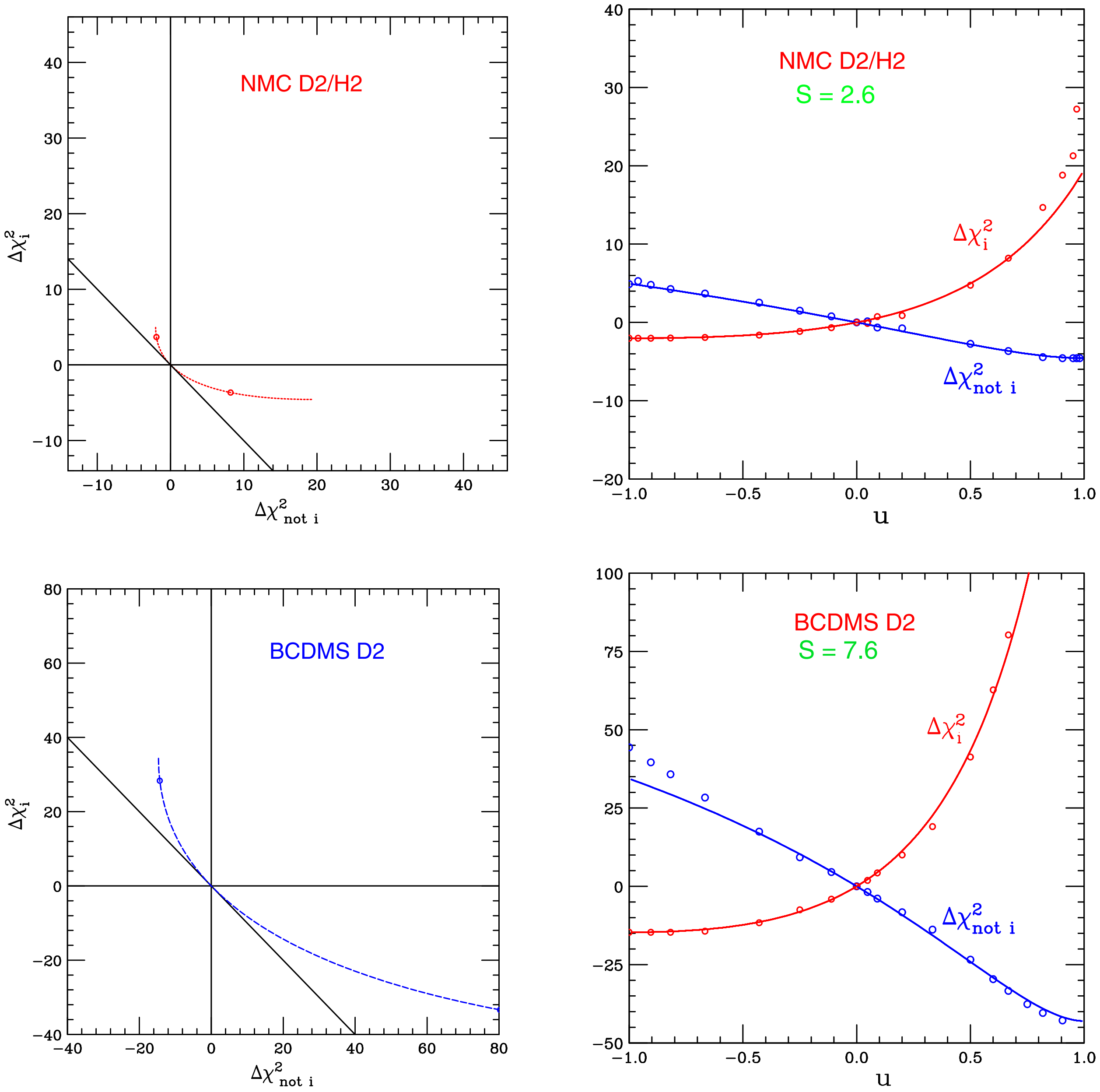
- Plot both as function of Lagrange multiplier u where $(1-u)\chi_S^2 + (1+u)(\chi_{\overline{S}}^2$ is the quantity minimized.

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

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

$$\chi_{\overline{S}} = B + \left(\frac{p-S}{\cos\theta}\right)^2 \quad \Rightarrow \quad p = s \pm \cos\theta$$

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 $\tan\phi$	2.7 0.56	3.3 0.54	3.3 0.99	4.2 0.86	5.3 0.71	7.6 1.14	7.4 0.65	8.3 0.39

The approach John and I considered maps χ^2_S as a function of $\chi^2_{\overline{S}}$. The method has three problems:

1. Since traditional Gaussian statistics don't seem to apply to our problem because of unknown systematic errors (both in theory and in experiment), we 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 parts of the theory are affected by the tension between S and $\overline{S}$.

3. Discrepancies between experiments don't matter if they are along parameter directions that are well-constrained by other experiments.

A new "Data Set Diagonalization" method, which expands upon the Hessian method, appears to solve problems 2 and 3.

The new method works in multiple dimensions: it finds all of the directions in parameter space that are controlled by the particular experiment that is under study.

DSD method

The quality of the fit of a PDF set to the data is measured by

$$\chi^2 = \sum_{i=1}^{M} \left(\frac{D_i - T_i}{E_i} \right)^2$$

where D_i and E_i represent a data point and its uncertainty. T_i is the theoretical prediction, which depends on “shape parameters” a_1,. . . ,a_N which describe the PDFs.

Near the Best Fit minimum, Taylor series implies that χ^2 is a quadratic function of the shape parameters:

$$\chi^2 = f + \sum_{i=1}^{N} g_i \, a_i + \sum_{i=1}^{N} \sum_{j=1}^{N} h_{ij} \, a_i \, a_j$$

Using the eigenvectors of the Hessian matrix $\mathbf{h}$, we can make a linear transformation to obtain new shape coordinates such that

$$\chi^2 = \chi^2_{\mathrm{min}} + \sum_{i=1}^{N} z_i^2 \, .$$

(Owing to numerical instabilities associated with flat directions, it is generally necessary to calculate that transformation by an iterative method.)

The contribution to χ^2 from subset S of the data can also be expanded to quadratic accuracy:

$$\chi_S^2 = a + \sum_{i=1}^{N} b_i\, z_i + \sum_{i=1}^{N} \sum_{j=1}^{N} c_{ij}\, z_i\, z_j$$

The key step of the DSD method occurs now: a further orthogonal transformation defined by the eigenvectors of c_{ij} makes the matrix $\mathbf{c}$ diagonal. Thus

$$\chi_{\mathrm{S}}^2 = \alpha + \sum_{i=1}^{N} (2\,\beta_i\, z_i + \gamma_i\, z_i{}^2)$$

while preserving

$$\chi^2 = \chi^2_{\mathrm{min}} + \sum_{i=1}^{N} z_i{}^2 \, .$$

Assume for the moment that all of the γ_i parameters lie between 0 and 1. Then by simple algebra, we obtain

$$\begin{aligned}\chi^2 &= \chi^2_{\mathbf{S}} + \chi^2_{\overline{S}} + \text{const} \\ \chi^2_{\mathbf{S}} &= \sum_{i=1}^{N} \left(\frac{z_i - A_i}{B_i} \right)^2 \\ \chi^2_{\overline{S}} &= \sum_{i=1}^{N} \left(\frac{z_i - C_i}{D_i} \right)^2 \quad . \end{aligned} \tag{1}$$

This has a simple interpretation: S and its complement take the form of independent measurements of the quantities $z_1, \ldots, z_N$:

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

The difference between these is

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

The incompatibility between S and the rest of the global fit along direction i is thus given by

$$|A_i - C_i| / \sqrt{B_i^2 + D_i^2}$$

in standard deviations.

This decomposition answers the question "What is measured by data subset S?" — it is those parameters z_i for which the $B_i \lesssim D_i$.

Compatibility between S and its complement only matters for the directions in which the two measurements have comparable errors. In practice, that is roughly the range $0.2 < \gamma_i < 0.8$:

γ_i	B_i/D_i
0.1	3
0.2	2
0.5	1
0.8	1/2
0.9	1/3

Relation between $B_i =$ uncertainty from $\mathbf{S}$ and $D_i =$ uncertainty from $\overline{S}$, for various γ_i .

In particular, directions for which γ_i is outside the range 0 to 1 are irrelevant for the comparison.

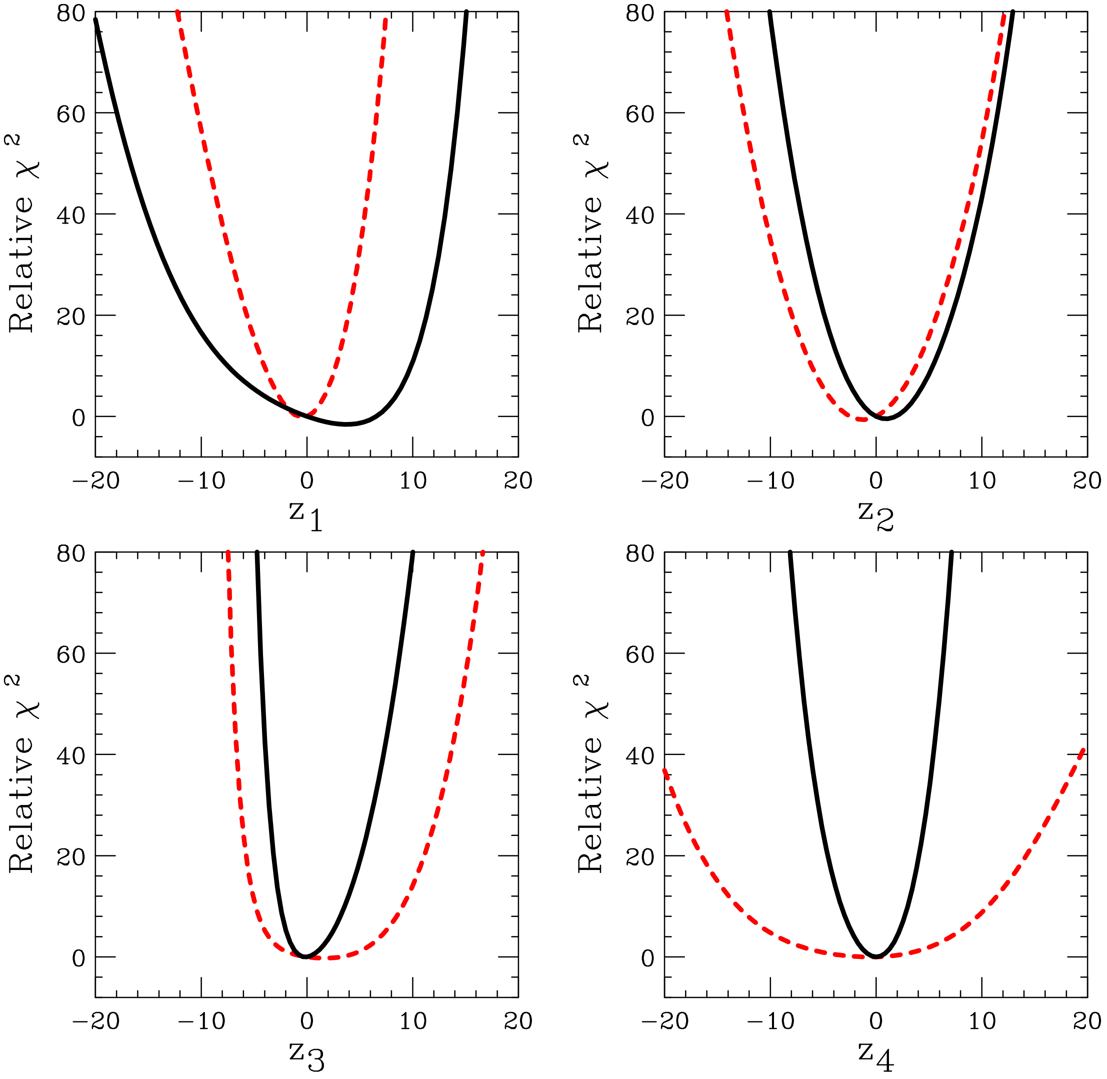
Example: E605 (DY pair production in p-Cu)

i	γ_i	z_i from $\mathbf{S}$	z_i from $\overline{S}$	Difference	
1	0.93	-0.40 ± 1.06	3.78 ± 2.55	4.18 ± 2.76	$1.51\,\sigma$
2	0.42	-1.22 ± 1.52	0.93 ± 1.34	2.15 ± 2.03	$1.06\,\sigma$
3	0.09	1.71 ± 3.34	-0.15 ± 1.01	1.86 ± 3.48	$0.53\,\sigma$
4	0.05	-0.57 ± 4.54	0.03 ± 1.03	0.60 ± 4.65	$0.13\,\sigma$
...	...	...	...	...	...

Parameter z_1 is determined almost entirely by this experiment.

Parameter z_2 is determined about equally by E605 and its complement.

E605 doesn't give any useful information about any of the other parameters.



χ^2 for fit to E605 (dashed curves) and to the rest of the data (solid curves) along the four leading eigenvector directions in descending order of γ_i. The overall best fit is at $z_i = 0$ in each case.

These figures confirm the results of the previous table: E605 dominates the measurement of z_1, and is mildly in conflict with the other experiments along that direction; it plays an important role in determining z_2; and it has nothing to say about $z_3, \ldots, z_{24}$.

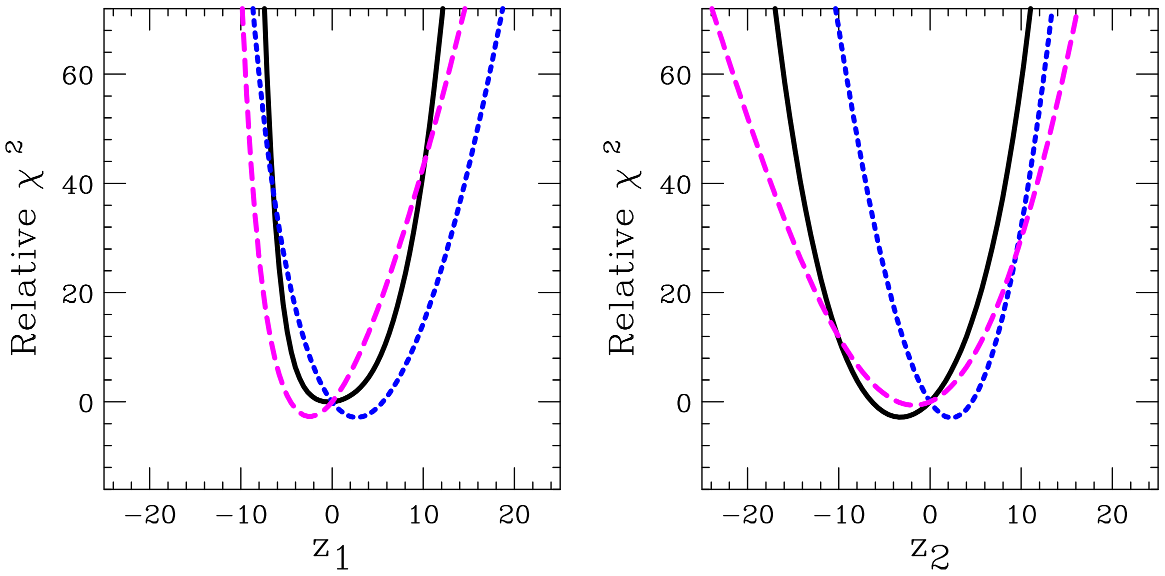
Example: Tevatron Run II Jet experiments

i	γ_i	z_i from $\mathbf{S}$	z_i from $\overline{S}$	Difference	
1	0.79	0.09 ± 1.13	-0.34 ± 2.10	0.43 ± 2.39	$0.2\,\sigma$
2	0.72	1.19 ± 1.16	-3.24 ± 1.92	4.42 ± 2.24	$2.0\,\sigma$
3	0.10	0.41 ± 3.08	-0.05 ± 1.06	0.45 ± 3.25	$0.1\,\sigma$
4	0.03	-6.16 ± 6.47	0.18 ± 1.07	6.34 ± 6.55	$1.0\,\sigma$

S = CDF+D0 Run II; Run I jet data removed for simplicity.

The jet experiments dominate along two directions, showing a mild incompatibility along one of them.

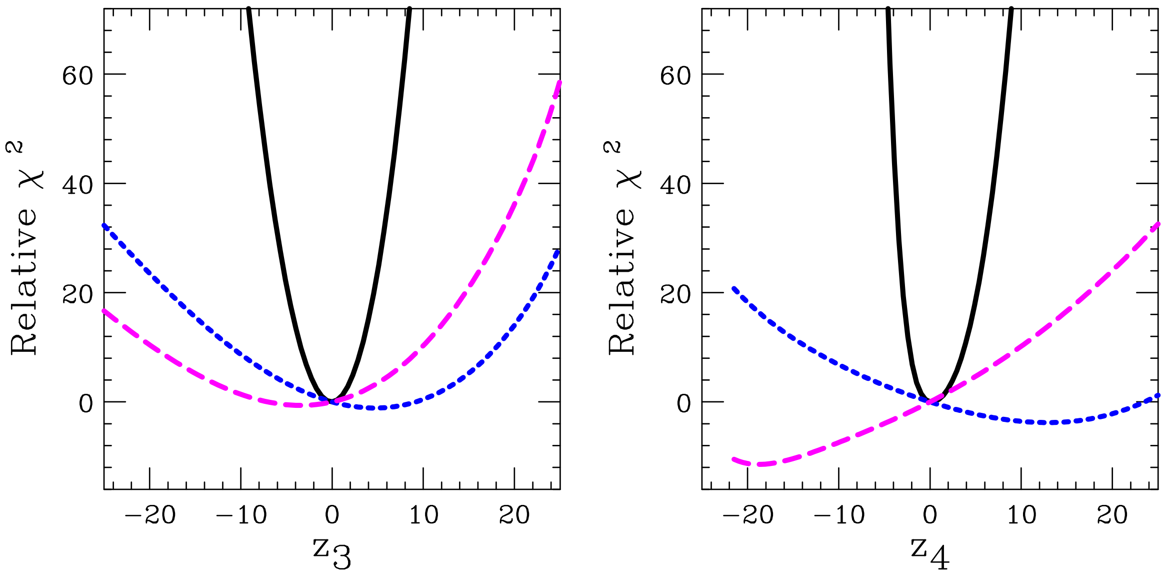
(Note, these z_i are the result of diagonalizing the *jet* contribution to χ^2: they are *not* the same parameters as the z_i used in the study of E605!)



χ^2 for CDF (blue), D0 (magenta), and rest of the data (black) along z_1 and z_2.

- Good agreement between the average of CDF and D0 with the rest of the global fit along z_1, but there is clearly some difference between CDF and D0 along that direction.

- Also some difference between CDF and D0 with regard to z_2; but this time D0 agrees with the non-jet data, while CDF has a bit of tension with it (2σ).



χ^2 for fit to CDF (blue), D0 (magenta), and the rest of the data (black) along z_3 and z_4.

- Parameters z_3, z_4,. . . ,z_{24} don't matter because the non-jet data determine those parameters. The apparent incompatibility between CDF and D0 along the z_4 axis, for example, is no cause for concern because the non-jet data determine that parameter very well, as shown by the very narrow black parabola.

The discrepancy between the two run II jet experiments (run I experiments were not used in this study) is given by

i	z_i from CDF	z_i from D0	Difference	
1	2.70 ± 1.65	-2.45 ± 1.38	5.15 ± 2.15	$2.40\,\sigma$
2	2.33 ± 1.35	-1.74 ± 2.22	4.07 ± 2.60	$1.57\,\sigma$

The Error Puzzle

Applying the Data Diagonalization Method to every experiment in the CT09 fit finds lots of discrepancies on the order of 1 or 2 σ; but not much larger than that. Quantitatively, 68 % of the discrepancies are less than 1.7 σ — Within a factor of 2 of the expectation from Gaussian Statistics!

Furthermore, the global fit generally has χ^2/Npt close to 1 for all experiments.

Nevertheless, the apparently overly conservative $\Delta\chi^2 \sim 100$ criterion in the context of these global fits is necessary to satisfy the basic requirements that the uncertainty should not be small compared to the differences between the results from groups using a similar approach; or differences from one published PDF set to the next when those differences are not explained by qualitatively new data or theory.