DETECTING UNKNOWN SYSTEMATIC EFFECTS: DIAGNOSIS OF A BAD FIT IN MULTIPLE DATA SETS

M.J. Wang
Institute of Physics, Academia Sinica,Taipei 11529, Taiwan , R.O.C.

Abstract

Error estimates on parton distribution functions using global fits to multiple data sets are becoming important due to the improving precision data of hadron collider experiments. For estimating a set of reliable parameter uncertainties, one needs to make sure that an uniformly consistent data set has been established beforehand. For this reason, a more stringent criterion for testing goodness of fit had been proposed and inconsistent data sets in the CTEQ5 global fits were indicated. In this paper, we propose an idea of examining the inconsistent experimental data set using correlations among pull quantities for the purpose of locating an unknown systematic shift in the data.

1 INTRODUCTION

The demand for Parton Distribution Function (PDF) uncertainties was clearly demonstrated in the interpretation of Tevatron data for the inclusive jet transverse energy[1]. PDF uncertainties could be the dominant component of uncertainty in RunII for the W mass measurement[2] with which the mass of Standard Model Higgs is constrained. Therefore, error estimates on PDFs are becoming very important for the upcoming precision hadron collider analyses.

1.1 Global fits and its goodness of fit

Reliable PDF parameter and uncertainty estimates require passing goodness of fit criteria. The conventional way is to use total χ^2 as the test statistic. Since there are usually thousands of data points in a global fit, the value of total χ^2 is insensitive to a small data set with bad fit. Clearly, a more stringent criterion for goodness of fit is needed for the purpose of fitting multiple data sets.

1.2 Parameter fitting criterion

With an idea motivated by L. Lyons's goodness-of-fit paradox at ACAT2000 , J.C. Collins and J. Pumplin applied the parameter-fitting criterion[3] to global fits. The authors plotted the subset χ^2 against the total χ^2 and found inconsistent data sets in the CTEQ5 global fits. Even though the inconsistent data sets could be identified they still could not tell which one is right or wrong. In order to make an exclusion decision, we need to look into the identified data sets in more detail.

2 DIAGNOSIS OF A BAD FIT IN MULTIPLE DATA SETS

To diagnose the bad fit in detail is very important, since it could provide hints leading to answers of the following three questions: (1) Is the inconsistent data set free of systematic effects? (2) Is the theoretical prediction adequate? (3) Is there any hint for new physics? The first two are important for selecting the right data sets for global fits. The last one, needless to say, is one of the major goals of our field.

2.1 Main idea: correlation among pulls suggests existence of unknown systematic effect

The pull[4] of each subset of data could be used to identify inconsistent data sets within the least square method framework. The central question of this study, however, is to examine whether the identified data set is free of unknown systematic effects. In the following, we will assume that the inconsistent experiment has been singled out by the parameter-fitting criteria.

2.11 Pull independence property in the naive case

In the naive case where there are no systematic errors, the pulls of the individual data points should be independent of one another, with a Gaussian distribution. This statement is valid when the statistical errors of the data points are indeed Gaussian, and when the number of degrees of freedom is sufficiently large. With over 1000 data points and high precision data in the global fits, this pull independence property is applicable.

2.12 Experiment pull distribution

In the global fits, all subsets of the data should have the same pull distribution. Therefore the pull distribution of all data points should have the same distribution as the individual pull distribution of each data point. In this paper, we will call the pull distribution of all data points in an individual experiment as the experiment pull distribution .

Any unknown systematic effects could introduce visible correlations among pulls, and distort this experiment pull distribution from its expected Gaussian shape. Therefore, any visible distortion of this experiment pull distribution in a particular data set could suggest the existence of an unknown systematic effect in that data set. This is the main idea of this paper.

2.2 Test of the pull-correlation idea

For simplicity, we will construct pull using the theoretical true values instead of the global fit values. There are two reasons for this simplification. First, we are not interested in investigating theoretical systematic effects, therefore we could ignore the difference between global fit results and the true values. Second, we are not interested in the biases and fluctuations introduced by the fitting procedure, therefore we could ignore the difference between the global fit results and the true values.

Certainly, in a real case, we only have the global fits predictions of each data point for constructing the pull. These effects will come in and make things more complicated. However, since global fit predictions are relatively insensitive to a single experiment, they still provide the best reference point in locating unknown systematic effects.

2.21 Test method and steps

The test method is to model a single experiment which has systematic problems, and to look at how that might show up in experiment pulls. We first picked a simple quadratic functional form $y(x)$ as a true theoretical curve.

$$y(x) = -0.1 * x^2 + 0.24 * x - 0.004 \tag{1}$$

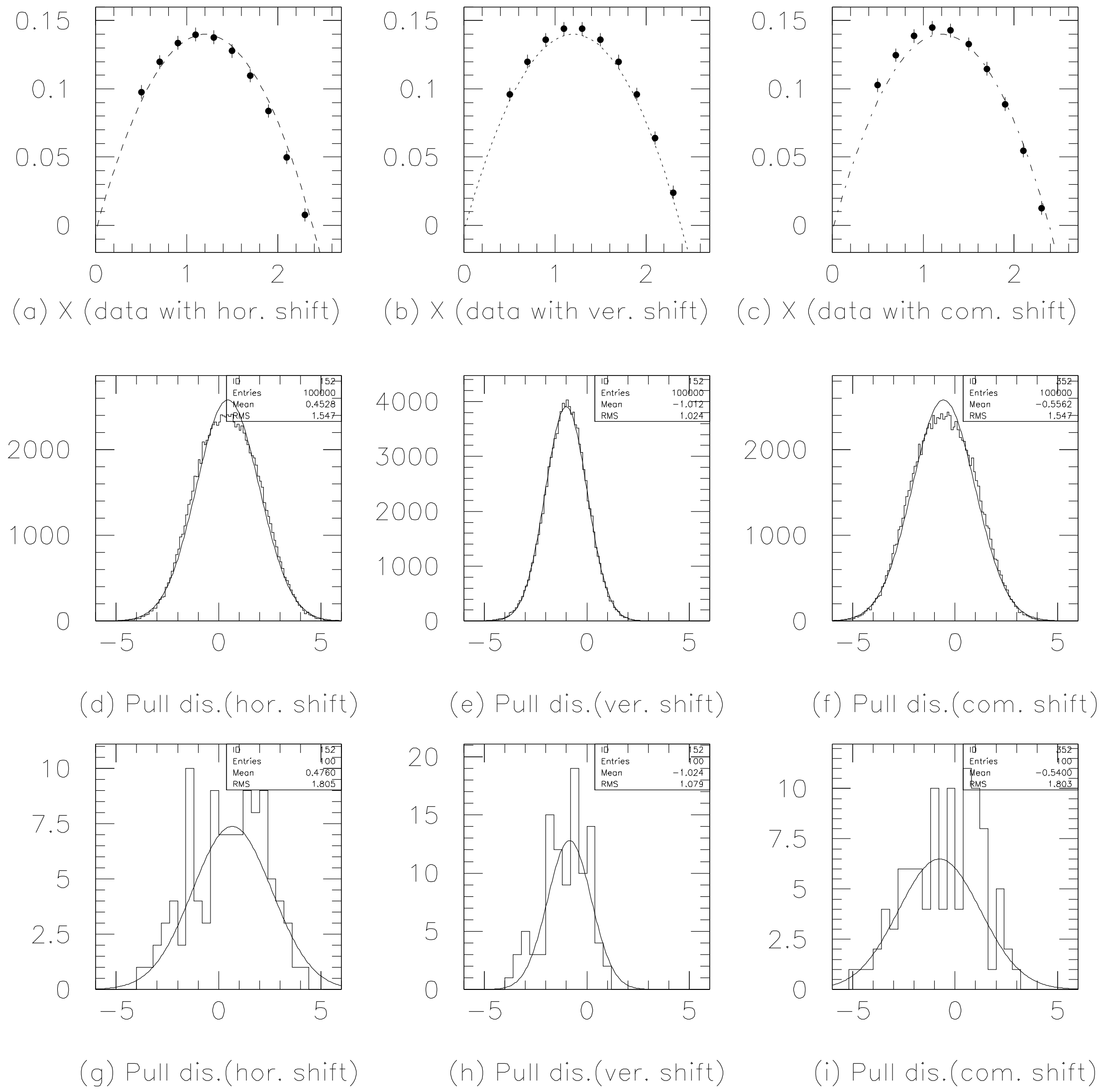
Then we suppose an overall measurement in a variable x_i, in which there are 10 bins $i = 1 : 10$. At each point in x, we generated a fake data y_i Gaussianly distributed about $y(x_i)$ with $\sigma = 0.005$. The residual and pull are defined in the following equations:

$$res_i = y(x_i) - y_i \tag{2}$$

$$pull_i = res_i / rms_i \tag{3}$$

where rms_i is the measurement error for each x_i. It was taken as a Gaussian random variable with a mean of 0.005 and a sigma of 0.0005, to simulate imperfect knowledge of the experiment's actual uncertainties.

We assumed that there are 10 sets of measurements from a single experiment with shared systematics, of a similar variable. Therefore, there are 100 data points in this experiment. Since we would like to study the asymptotic behavior of the experiment pull distribution first, we generated 1k pseudo experiments with 100 data points each. This overall experiment pull distribution has 100k pull entries



(a) X (data with hor. shift) (b) X (data with ver. shift) (c) X (data with com. shift)

(d) Pull dis.(hor. shift) (e) Pull dis.(ver. shift) (f) Pull dis.(com. shift)

(g) Pull dis.(hor. shift) (h) Pull dis.(ver. shift) (i) Pull dis.(com. shift)

Fig. 1: (a)-(c) show simulated data with horizontal, vertical, and combined shifts. Both solid and dash lines represent the true curve.(d)-(f) show experiment pull distribution with 100k entries. (g)-(i) show experiment pull distribution with 100 entries.

in total. Second, we simulated a pseudo experiment with 100 data points. This overall experiment pull distribution has 100 pull entries in total.

Next we suppose all of those experiments to have systematics, so that they measure with means not at $y(x_i)$, but either at $y(x_i$+horizontal shift), or $y(x_i)$+vertical shift, or $y(x_i$+horizontal shift)+vertical shift. Experiment pull distributions are then constructed with constant horizontal, constant vertical systematic shifts and the combination of both for the simulated data. The true curves and simulated data with all 3 shifts are shown in figure 1(a), (b), and (c).

2.3 Discussion of results

Each experiment pull distribution was fitted to a Gaussian distribution. The number of degrees of freedom varied from case to case because the number of bins over which each experiment pull distribution extends is not the same. With no systematic shift, the means of the pull distributions for both 100k and 100 cases are consistent with 0. The sigma's of the pull distributions are also quite close to 1, as expected. The shape of pull distribution is also consistent with Gaussian for the 100k case according to $\chi^2/d.o.f. = 117/91$. It is hard to distinguish the distribution shape with only 100 entries.

Table 1: Distorted pull distribution characteristics under systematic effects

Systematic shift	# of entries	Pull distribution characteristics		
		Mean	Sigma	χ^2/d.o.f.
no shift	100,000	-0.001	1.02	117/91
	100	0.12	1.04	8/10
constant horizontal shift	100,000	0.46	1.54	698/110
	100	0.67	1.95	13/18
constant vertical shift	100,000	-1.01	1.02	187/90
	100	-0.85	1.07	17/10
combination of both	100,000	-0.56	1.54	734/112
	100	-0.77	2.00	21/18

2.31 Distorted pull characteristics under a constant horizontal shift

Distorted pull characteristics could be clearly seen when a constant horizontal shift is applied to the data as shown in figures 1(a), (d), and (g). Figure 1(a) shows the shifted data and the true curve. Figure 1(d) shows the experiment pull distribution with 100k entries. Figure 1(g) shows the experiment pull distribution with 100 entries. The horizontal shift produces a corresponding vertical deviation of about 1 σ.

In the 100k case, the mean of the pull distribution is clearly inconsistent with 0. The distribution width is widened by 50%. This widening effect is understood. Since the sign and magnitude of the shift varies with x, the net effect is to widen the distribution. The distribution shape also deviates from Gaussian normal distribution with $\chi^2/d.o.f. = 698/110$. In the 100 entries case, the mean of pull distribution is clearly inconsistent with 0. The distribution width is widened by a factor of 2. It is hard to distinguish the distribution shape with only 100 entries.

2.32 Distorted pull characteristics under a constant vertical shift

Distorted pull characteristics could be also clearly seen under constant vertical shift as shown in figures 1(b), (e), and (h). Figure 1(b) shows the shifted data and the true curve. Figure 1(e) shows the experiment pull distribution with 100k entries. Figure 1(h) shows the experiment pull distribution with 100 entries. This vertical shift on distribution produces a corresponding vertical deviation of each data point with about the size of 1 σ statistical uncertainty of data points.

In the 100k case, the mean of the pull distribution clearly is inconsistent with 0. The width of distribution is quite close to 1 as expected. The distribution shape is also deviated from Gaussian normal distribution according to $\chi^2/d.o.f. = 187/90$. In the 100 entries case, the mean of pull distribution clearly is inconsistent with 0. The width of distribution is quite close to 1 as expected. It is hard to distinguish the distribution shape with only 100 entries.

2.33 Distorted pull characteristics under a combined shift

Distorted pull characteristics could also be seen under combined constant horizontal and vertical shifts as shown in figures 1(c), (f), and (i). Figure 1(c) shows the shifted data and the true curve. Figure 1(f) shows the experiment pull distribution with 100k entries. Figure 1(i) shows the experiment pull distribution with 100 entries.

In the 100k case, the mean of the pull distribution clearly is inconsistent with 0. The distribution width is widened by 50%. The distribution shape also deviates from a Gaussian normal distribution, with

$\chi^2/d.o.f. = 734/112$. In the 100 entries case, the mean of pull distribution clearly is inconsistent with 0. The distribution width is widened by a factor of 2. It is hard to distinguish the distribution shape with only 100 entries.

2.4 New pull definition required for real case

In the real cases with known systematic corrections and therefore uncertainties, correlation among pulls have been introduced by the existing systematic uncertainties. To restore independence of the pulls, we need to subtract out the effects of known systematic uncertainty for a new definition of pull. With this new pull definition, all the above results in the naive cases will be retained and utilized as a tool to detect unknown systematic effect.

3 SUMMARY AND OUTLOOK

We propose an idea of detecting unknown systematic effect in the data using correlations among pulls. This idea was tested under constant horizontal shift, constant vertical shift and combination of them. In all the 3 cases, the distorted pull distribution characteristics could be used to indicate the existence of unknown systematic effect. The assumption adopted here is that there is no known systematic corrections in this naive case. Tests and applications in real cases in which there are known systematic corrections will be pursued next.

ACKNOWLEDGEMENTS

I would like to thank J. C. Collins for discussion and comments on this paper. I also would like to thank the conference local organizers James Stirling, Mike Whalley, and Linda Wilkinson for their warm hospitality and management efforts to make this conference successful. Last but not the least, I would like to thank J. Linnemann for his industrious efforts in editing this paper and very helpful suggestions.

References

[1] The CDF collaboration, F. Abe et al., Phys. Rev. Lett. 77(1996) 438.

[2] The CDFII collaboration, F. Abe et al., The CDFII detector technical design report, p. 2-27, FERMILAB-PUB-96/390-E.

[3] J. C. Collins and J. Pumplin, hep-ph/0105207(2001).
”Tests of goodness of fit to multiple data sets”.

[4] W.T.Eadie, D.Drijard,F.E.James, M.Roos, B.Sadoulet
”Statistical Methods in Experimental Physics”, p.278.