Nuisance parameters in the PDF analysis

Pavel Nadolsky
in collaboration with
J. Huston, H.-L. Lai, J. Pumplin, D. Stump, and C.-P. Yuan

Southern Methodist University
Dallas, TX, U.S.A.

October 23, 2009

Correlated systematic errors (CSE) in CTEQ fits

- CSE provided by experiments are important. PDF errors are underestimated without them.

- CTEQ takes them into account since 2000, by applying **algebraic minimization** (AM) of χ^2 with respect to systematic (nuisance) parameters λ_α *(D. R. Stump et al., PR D65 (2002) 014012)*

  - Nuisance parameter = a parameter that does not appear explicitly in the PDF parametrization, but must be accounted for in the fit

- In the course of the CT09 analysis, we re-examined the role of CSE's, partly because they affect determination of the gluon PDF from Tevatron jet production cross sections

New ideas about algebraic minimization (AM)

AM can be applied in new ways to resolve several issues:

1. reduce complexity of correlated systematic errors published by experiments
2. check validity of published experimental CSE's (e.g. covariance matrix for DO Run-1 jet data)
3. allow experimental normalizations to float when producing Hessian PDF eigenvector sets
4. evaluate correlated shifts in **theory** values caused by scale dependence, higher twists, etc.
5. propagate correlated PDF uncertainties into third-party fits (e.g., into M_W measurements)

Common representations for CSE

1. $N_{pt} \times N_\lambda$ correlation matrix $\beta_{k\alpha}$ for N_λ random nuisance parameters λ_α

$$\chi^2 = \sum_{e=\{\text{expt.}\}} \left[\sum_{k=1}^{N_{pt}} \frac{1}{s_k^2} \left(D_k - T_k(\{z\}) - \sum_{\alpha=1}^{N_\lambda} \lambda_\alpha \beta_{k\alpha} \right)^2 + \sum_{\alpha=1}^{N_\lambda} \lambda_\alpha^2 \right]$$

- D_k and T_k are data and theory values $(k = 1, ..., N_{pt})$;
- s_k is the stat.+syst. uncorrelated error;
- $\{z\}$ are PDF parameters;$\{z = 0\}$ in the best fit

2. $N_{pt} \times N_{pt}$ covariance matrix C (less common than β):

$$\chi^2 = \sum_{k,k'} (D_k - T_k) C^{-1}_{kk'} (D_{k'} - T_{k'})$$

Algebraic solution for CSE parameters λ_a

β and C are related by **algebraic minimization** of χ^2 with respect to λ_α. **If** $d_i \equiv D_i - T_i$**;** d_i**,** $\beta_{i\alpha}$ **are given in units of** s_i **for each** $i = 1, ..., N_{pt}$; and **for Gaussian** λ_α:

$$\lambda_\alpha\left(\{z\}\right) = \sum_{\alpha'=1}^{N_\lambda} (\mathcal{A}^{-1})_{\alpha\alpha'} B_{\alpha'}(\{z\})$$

$$\mathcal{A}_{\alpha\alpha'} = \delta_{\alpha\alpha'} + \sum_{i=1}^{N_{pt}} \beta_{\alpha i}\beta_{\alpha' i}; \qquad B_\alpha(\{z\}) = \sum_{i=1}^{N_{pt}} \beta_{\alpha i}(D_i - T_i)$$

$$\chi^2(z, \lambda(z)) = \sum_{k,k'} d_k \left[I - \beta\mathcal{A}^{-1}\beta^T\right]_{kk'} d_{k'} \equiv d^T \left[I - \beta\mathcal{A}^{-1}\beta^T\right] d$$

$$\therefore C = \left(I - \beta\mathcal{A}^{-1}\beta^T\right)^{-1} = I + \beta\beta^T$$

Numerical minimization of $\chi^2(z, \lambda(z))$ establishes the region of acceptable $\{z\}$, which includes the largest possible variations of $\{z\}$ allowed by the systematic effects

Class S_n of positive semi-definite symmetric $n \times n$ matrices

$C, \beta\beta^T$ belong to $S_{N_{pt}}$. All their eigenvalues σ_a are positive semi-definite:

$$\sigma_a > 0 \text{ for } \alpha \leq r, \text{ and } \sigma_a = 0 \text{ for } \alpha > r,$$

where rank $r = N_{pt}$ for C and $r = N_\lambda$ for $\beta\beta^T$.

$$\mathrm{tr}\, C \geq \mathrm{tr}\, \beta\beta^T > 0$$

One can also define a semi-norm (distance) on $S_{N_{pt}}$:

$$||A - B||^2 = \sum_{i \leq j} (A_{ij} - B_{ij})^2$$

- "a measure of how similar A and B are numerically"

1. Rank reduction on S_n

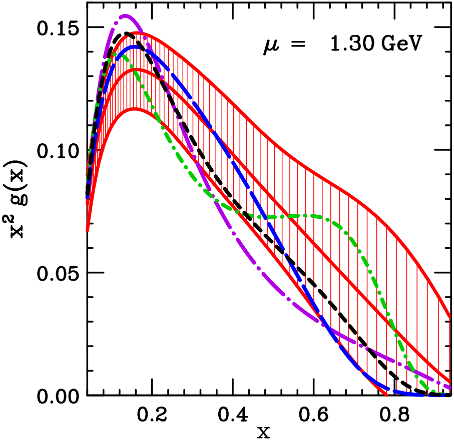
- An approximation of a matrix $A \in S_n$ with rank N by another matrix $B \in S_n$ with rank $B = M < N$ such that $||A - B||$ is as small as possible
  - The approximation B can be more useful than the full matrix A
- For example, a large CSE matrix $\beta\beta^T$ of rank N_λ can be replaced by an approximate $\beta'\beta'^T$ of rank $N'_\lambda < N_\lambda$ without appreciable precision loss (**principal component analysis**)
- For Tevatron jet production data, only $\approx N_\lambda/2$ combinations of λ_α (found by PCA) are relevant for χ^2;
  rank $\left[\beta'\beta'^T\right] \approx N_\lambda/2 \ll N_{pt}$
- PCA would simplify greatly the combined H1+ZEUS correlation matrix with tens of (small) CSE's
  - **Would HERA experimentalists be interested to provide the H1+ZEUS correlation matrix for a PCA study?**

2. Reconstruction of β from C

D0 Run-1 data on inclusive jet production prefer a very different $g(x, Q)$ as compared to 3 other Tevatron jet data sets

The D0 Run-1 measurement has

- large uncorrelated syst. errors (unspecified)
- 15 correlated systematic errors
- provides the covariance matrix C only



A fit to D0 Run-1 jets only (green dash-dots); CT09 fit (red band); fits to CDF Run-1, CDF Run-2, and D0 Run-2 jet data

Does C have a valid structure? Can β and uncorrelated errors be reconstructed from C?

D0 Run-1 jets: extraction of β from C

- An iterative algorithm can systematically extract a realistically looking matrix β and total uncorrelated error $s_i^{(r)} = \sqrt{s_{i,stat}^2 + \left(s_{i,uncor.\,syst.}^{(r)}\right)^2}$ from the published D0 Run-1 covariance matrix C

- This solution implies
  - $6-8$ large combinations of **correlated** errors (close to $N_\lambda/2 = 15/2$ expected from the D0 Run-1 publication)
  - large **uncorrelated** systematic errors (up to 6 times larger than statistical errors)

As a cross check, the method was applied to extract β from C in other three jet experiments. In all those cases, the reconstructed β agreed well with the actual β

Algorithm for iterative extraction of β

For each desired rank r of $\beta\beta^T$, seek a solution of the form

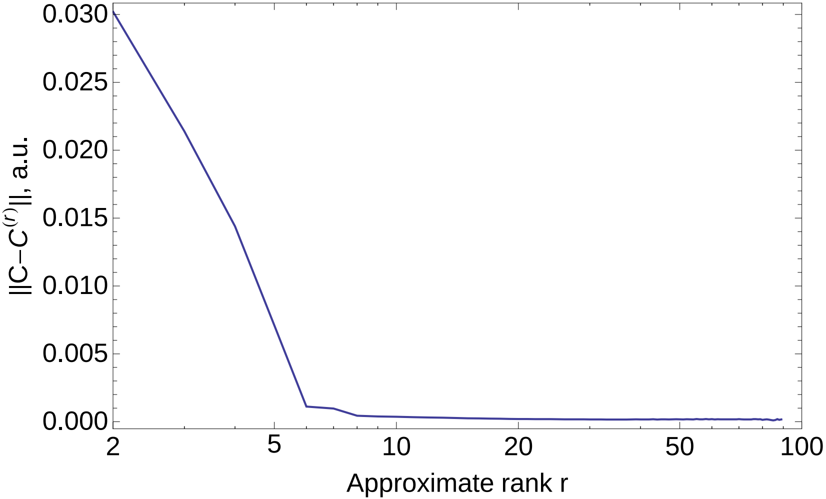
$$C_{ij}^{(r)} = s_i^{(r)} s_j^{(r)} \left[I + \left(\beta\beta^T\right)^{(r)} \right]$$

that minimizes $||C - C^{(r)}||$ (can be found recursively)

$\left(\beta\beta^T\right)^{(r)}$ is essentially the square of the correlation matrix in the PCA representation

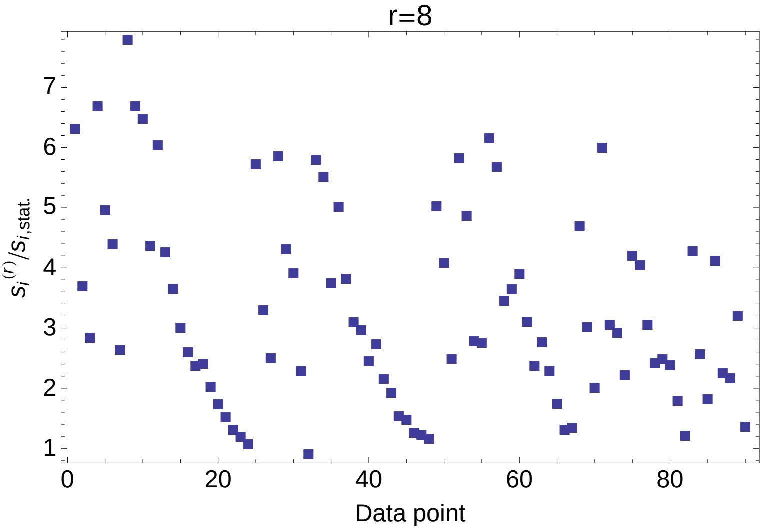
Rank r of $\beta\beta^T$ should be large enough ($r > 6 - 8$) in order to achieve small $||C - C^{(r)}||$

$||C - C^{(r)}||$ vs. desired rank r for D0 Run-1 jet data



Ratios of total uncorrelated to statistical errors

for D0 Run-1 jet data



3. New treatment of experimental normalizations

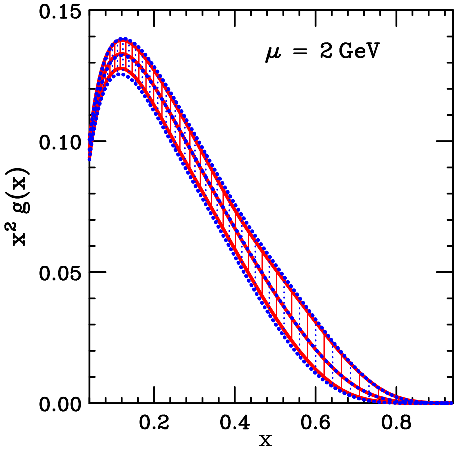
- Experimental publications commonly list the normalization (Norm) error as one of correlated systematic uncertainties
  - In the past, we fitted for Norm values numerically – Norm dependence was deleted from β matrices
- Starting with CT09, we treat **all** normalizations on the same footing as the other systematic errors
  - apply algebraic minimization, possibly with a quartic penalty on χ^2
- **Advantages**
  - Number of input free parameters reduced by a factor of 2 – much faster fits
  - Error PDFs are obtained with floating normalizations
  - No worries about D'Agostini's bias

PDF error bands with the new treatment of normalizations (Pumplin, 2009)

Results are VERY PRELIMINARY and can change

Red band: CT09 – $\Delta\chi^2 = 10$; normalizations fixed at the best-fit values when producing error PDF sets

Blue band: new fit – same $\Delta\chi^2$, but normalizations vary



The new error band is slightly wider

4. Correlated theoretical uncertainties

$$\ln \chi^2 = \sum_{e=\{\text{expt.}\}} \left[\sum_{k=1}^{N_{pt}} \frac{1}{s_k^2} \left(D_k - T_k(\{z\}) - \sum_{\alpha=1}^{N_\lambda} \lambda_\alpha \beta_{k\alpha} \right)^2 + \sum_{\alpha=1}^{N_\lambda} \lambda_\alpha^{2n} \right],$$

$\beta_{k\alpha}$ can describe shifts in theory values due to nuisance factors like scale dependence, etc.

These shifts are treated in a linear approximation, which may or may not be appropriate under realistic conditions

$$\beta_{k\alpha} = \frac{\partial T_k(\{z=0\}, \{\lambda\})}{\partial \lambda_\alpha}$$

Theoretical $\beta_{k\alpha}$matrices were published recently for single-inclusive jet production at the Tevatron and LHC
(Olness, Soper, arXiv:0907.5052)

5. Correlated PDF dependence in third-party data analyses

- W production data are employed to determine the PDFs (by us) and W boson mass M_W (by experimental template fits)
- Negligence of possible correlations between the PDFs and M_W may skew the resulting M_W values
- In general, PDF parameters $\{z\}$ behave as nuisance parameters in third-party fits by experimental collaborations
  - must be treated accordingly

5. Correlated PDF dependence in third-party data analyses

In a third-party fit, $\{z\}$-dependence around the central PDF can be parametrized by a correlation matrix:

$$\chi^2(M_W,\{z\}) = \left[\sum_{k=1}^{N_{pt}} \frac{1}{s_k^2}\left(D_k - T_k(M_W,\{z=0\}) - \sum_{\delta=1}^{N_z} z_\delta \beta_{k\delta}\right)^2 + \sum_{\delta=1}^{N_z} z_\delta^2\right],$$

which produces

$$\chi^2(M_W, z_0(M_W)) = d^T\left[I - \beta \mathcal{A}^{-1}\beta^T\right] d.$$

The M_W error from such fit is generally not the same as the "old-fashioned"

$$\delta_{PDF} M_W = \frac{1}{2}\sqrt{\sum_{\delta=1}^{N_z}\left(M_W(z_\delta^+) - M_W(z_\delta^-)\right)^2}.$$

Conclusions

Algebraic minimization with respect to nuisance parameters can advance the PDF analysis on several fronts:

1. rank reduction of experimental correlation matrices
2. reconstruction of a correlation matrix from a covariance matrix
3. improved handling of experimental normalizations
4. implementation of correlated theoretical shifts in global fits
5. account for PDF-driven correlations in third-party fits

This method relies on linear approximations for dependence on nuisance parameters; needs further tests, but opens tantalizing possibilities