Correlated errors in the global PDF analysis

Jun Gao, **Pavel M. Nadolsky**

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

August 23, 2012

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]
- An experiment may publish data with up to ~ 100 types of *asymmetric* correlated errors. Are all these errors equally important? How to select the most important ones?
- Uncertainties in **theory** values caused by scale dependence, higher twists, etc. may be viewed in some approximation as correlated systematic shifts

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 (not used by CTEQ):

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

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:

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

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

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

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

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

Fractional CSEs

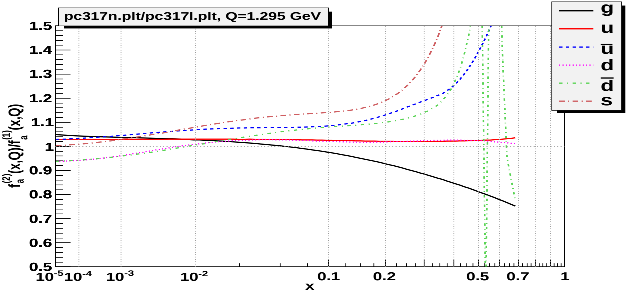
$$\chi^2 = \sum_{\{\text{exp.}\}} \left[\sum_{k=1}^{N_{pts}} \frac{1}{s_k^2} \left(D_k - T_k(\{a\}) - \sum_{\alpha=1}^{N_\lambda} \lambda_\alpha \beta_{k\alpha} \right)^2 + \sum_{\alpha=1}^{K_e} \lambda_\alpha^2 \right]$$

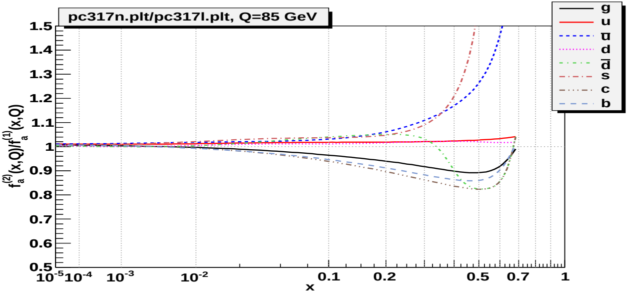
The experimental correlated systematic errors $\beta_{k\alpha}$ are often published as percentages. It can be taken to be a percentage of the theoretical prediction T_k ("truth") or the experimental datum D_k.

1. **Percentage of T_k:** results in smooth $\beta_{k\alpha}$:-); may depend on the theoretical model :-(
2. **Percentage of D_k:** $\beta_{k\alpha}$ is deduced from the measured data :-), but may be non-smooth and susceptible to d'Agostini's bias :-(

The methods are equivalent if T_k is close to D_k. In the actual CTXX fits to the Tevatron Run-2 jet data, **method 1** (used in pre-2012 CTEQ fits) results in a harder gluon at $x > 0.1$ than **in method 2**. We use **method 2** in the latest NNLO fits.

Impact on the best fit NLO PDFs



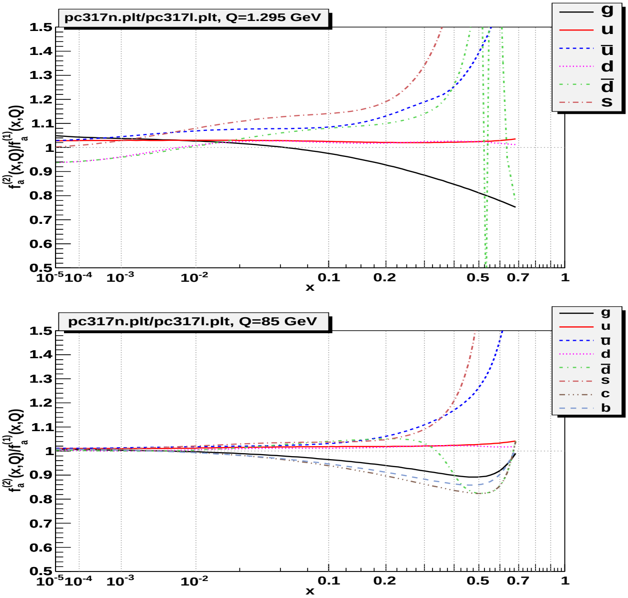


pc317l: CT12 NLO candidate obtained with method 1

pc317n: CT12 NLO candidate obtained with method 2

Notice changes in $u(x,Q)$, $d(x,Q)$, $g(x,Q)$

Impact on the best fit NLO PDFs



Neither method is 100%-proof. Data is equally happy with a harder $g(x,Q)$ in method 2 and softer $g(x,Q)$ in method 2. CT10 NNLO switched to method 2, but the alternative remains viable

Systematic uncertainties of jet cross sections

- A large number of experimental CSE's (~ 90 for ATLAS)
- Need to compare several procedures for handling of asymmetric errors. Our preliminary finding: the asymmetry of errors does not appreciably impact the PDF's.
- Significant scale dependence of NLO cross sections

χ^2/N_{pts}	CT10NLO			MSTW08NLO			HERA1.5NLO		
$\mu_{F(R)}/\mu_0$	0.5	1	2	0.5	1	2	0.5	1	2
CDF inc.	1.79	1.78	1.72	1.54	1.38	1.27	2.86	3.08	2.94
D0 inc.	1.43	1.23	1.19	1.42	1.09	0.98	1.73	1.73	1.56
D0 dijet	4.80	4.06	2.76	3.26	2.29	1.58	1.93	2.09	2.00
ATL inc. (0.4)	0.84	0.97	0.90	0.84	0.94	0.80	0.89	0.93	0.88
ATL inc. (0.6)	0.89	0.98	0.90	0.87	0.83	0.65	0.88	0.87	0.84

χ^2/N_{pts} of jet data for FastNLO with different scale choices. The central scale μ_0 is $p_{T,ind}$ for single-incl. jet production, $p_{T,ave}$ for dijet production. *[Cf. also Thorne, Watt, arXiv:1106.5789]*

An estimate of missing higher-order corrections: basic idea

See also Olness, Soper, arXiv:0907.5052; Cacciari, Houdeau, arXiv:1105.5152

For arbitrary $\mu_{R,F}$, the NLO cross sections in the experimental bins i can be written as

$$\sigma_{bin}^{NLO}(\mu_F,\,\mu_R,\,i) = \sigma_{bin}^{NLO}(\mu_F^{(0)},\,\mu_R^{(0)},\,i)\Big\{1 + \sum_{j=1}^{5} e_j(\mu_F^{(0)},\mu_R^{(0)},i)x_j + \mathcal{O}(\alpha_s^3(\mu_R^{(0)}))\Big\}$$

with

$$x_1 = \ln(\frac{\mu_F}{\mu_F^{(0)}}), \qquad x_2 = \ln(\frac{\mu_R}{\mu_R^{(0)}}), \qquad x_3 = \ln^2(\frac{\mu_F}{\mu_F^{(0)}}),$$
$$x_4 = \ln^2(\frac{\mu_R}{\mu_R^{(0)}}), \qquad x_5 = \ln(\frac{\mu_F}{\mu_F^{(0)}})\ln(\frac{\mu_R}{\mu_R^{(0)}}),$$

where $\mu_F^{(0)}$ and $\mu_R^{(0)}$ are the reference scales.

An estimate of missing higher-order corrections: basic idea

See also Olness, Soper, arXiv:0907.5052; Cacciari, Houdeau, arXiv:1105.5152

For arbitrary $\mu_{R,F}$, the NLO cross sections in the experimental bins i can be written as

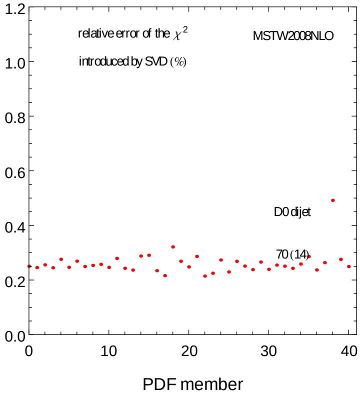
$$\sigma_{bin}^{NLO}(\mu_F,\,\mu_R,\,i) = \sigma_{bin}^{NLO}(\mu_F^{(0)},\,\mu_R^{(0)},\,i)\Big\{1 + \sum_{j=1}^{5} e_j(\mu_F^{(0)},\mu_R^{(0)},i)x_j + \mathcal{O}(\alpha_s^3(\mu_R^{(0)}))\Big\}$$

Treat x_i as independent corr. sources with quasi-Gaussian distributions (plausible, but not necessarily true). Assign your favorite confidence level (68% c.l.) to the range $1/2 < \mu_{F,R}/\mu_{F,R}^{(0)} < 2$. Evaluate the variation of $\sigma_{bin}^{NLO}(\mu_F,\,\mu_R,\,i)$ in this scale range. Find $e_j(i)$ numerically and use them to construct the correlation matrix. Reduce the number of principal components to eliminate x_i combinations that have vanishing effect on theory cross sections.

A new tool for the Correlation Error Matrix Analysis (CEMA)

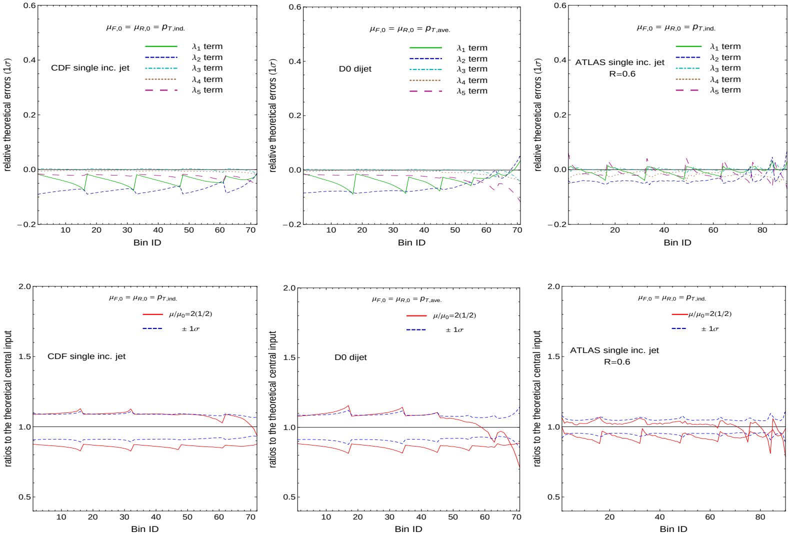
- A Mathematica package for general purpose comparisons of theory with correlated data (not only jet production).
- Calculation of correlated theoretical errors based on the above prescription.
- Standalone computation of χ^2 given experimental data with corr. syst. errors, and theoretical predictions with uncertainties (from FASTNLO, etc.).
- Principal component analysis (PCA) = elimination of corr. error combinations that have negligible effect on the PDFs (can be done outside of the fit).
- Symmetrization of asymmetric errors according to 3 prescriptions
- Various plots: histograms of best-fit syst. shifts, plots of theory vs. shifted data, plots of singular values of the correlation matrix.

Application of CEMA: reduction of non-essential CSE's from D0 Run-2 dijet data *(arXiv:1002.4594)*



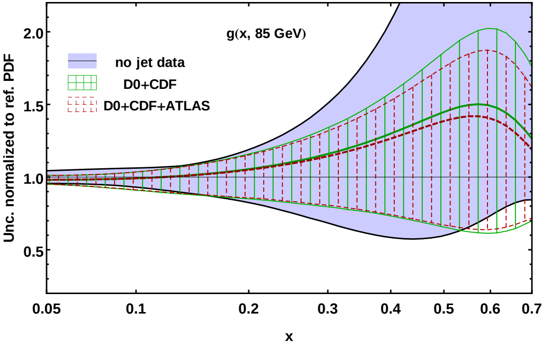
Only 14 combinations of CSE's out of 70 make non-negligible contributions to the calculations of χ^2. The rest may be removed from the PDF analysis safely

Application of CEMA: correlated theoretical errors for jet cross sections at the Tevatron and LHC



Fit with(out) jet data

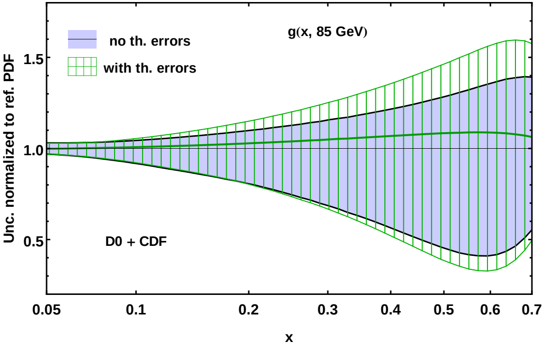
- Tevatron inclusive jet data does impose constraints on the CT10 gluon PDF. The 2010 ATLAS jet data does not strengthen the constraints yet because of large exp. errors.



PRELIMINARY

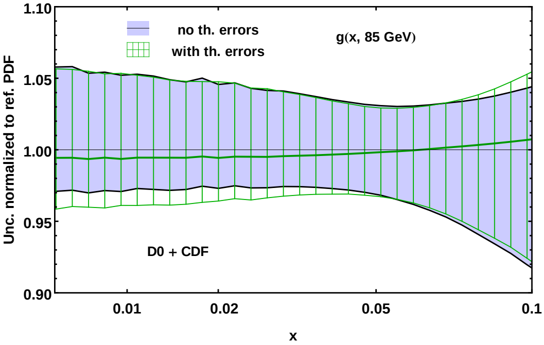
Effects of theoretical errors of jet data

- Gluon PDF uncertainties at 90% C.L. for the fits with and without theoretical errors. Scale dependence of jet cross sections increases the net gluon PDF uncertainty at $x > 0.1$ by about 20%.



PRELIMINARY

- The gluon PDFs in the moderate x region is also affected by the scale dependence errors, as a result of the anti-correlation with the gluon PDF at large x



PRELIMINARY

Conclusions

- Full analysis of the impact of hundreds of experimental and theoretical correlated errors on the PDFs is daunting. Insights about the bulk of this effect on the PDFs can be gained by algebraic minimization with respect to nuisance parameters, in the Gaussian approximation, assuming that many random independent influences and asymmetric effects cancel.

- We explored a prescription for treating theoretical uncertainties in NLO jet cross sections caused by scale variations as correlated systematic errors. In this approach, scale dependence can increase the gluon PDF uncertainties in the large x region by about 20%, also in the moderate x region by about 10% indirectly.