PROCEEDINGS OF SCIENCE

PDFSENSE:* Mapping the sensitivity of hadronic experiments to nucleon structure

Bo-Ting Wang[a], **T. J. Hobbs**[a], **Sean Doyle**[a], **Jun Gao**[b], **Tie-Jiun Hou**[c], **Pavel M. Nadolsky**[a], **Fredrick I. Olness**[a†]

[a] *Department of Physics, Southern Methodist University, Dallas, TX 75275-0181, U.S.A.*
[b] *Shanghai Key Laboratory for Particle Physics and Cosmology, School of Physics and Astronomy, INPAC, Shanghai Jiao-Tong University, Shanghai 200240, China*
[c] *School of Physics Science and Technology, Xinjiang University, Urumqi, Xinjiang 830046 China*

Recent high precision experimental data from a variety of hadronic processes opens new opportunities for determination of the collinear parton distribution functions (PDFs) of the proton. In fact, the wealth of information from experiments such as the Large Hadron Collider (LHC) and others, makes it difficult to quickly assess the impact on the PDFs, short of performing computationally expensive global fits. As an alternative, we explore new methods for quantifying the potential impact of experimental data on the extraction of proton PDFs. Our approach relies crucially on the correlation between theory-data residuals and the PDFs themselves, as well as on a newly defined quantity — the *sensitivity* — which represents an extension of the correlation and reflects both PDF-driven and experimental uncertainties. This approach is realized in a new, publicly available analysis package PDFSENSE, which operates with these statistical measures to identify particularly sensitive experiments, weigh their relative or potential impact on PDFs, and visualize their detailed distributions in a space of the parton momentum fraction x and factorization scale μ. This tool offers a new means of understanding the influence of individual measurements in existing fits, as well as a predictive device for directing future fits toward the highest impact data and assumptions.

XXVI International Workshop on Deep-Inelastic Scattering and Related Subjects (DIS2018)
16-20 April 2018
Kobe, Japan

*The webpage for the **PDFSense** tool is: **https://metapdf.hepforge.org/PDFSense/**

We acknowledge the hospitality of CERN, DESY, and Fermilab where a portion of this work was performed. This work was also partially supported by the U.S. Department of Energy under Grant No. DE-SC0010129 and by the National Natural Science Foundation of China under the Grant No. 11465018.

†Speaker.

PoS(DIS2018)024

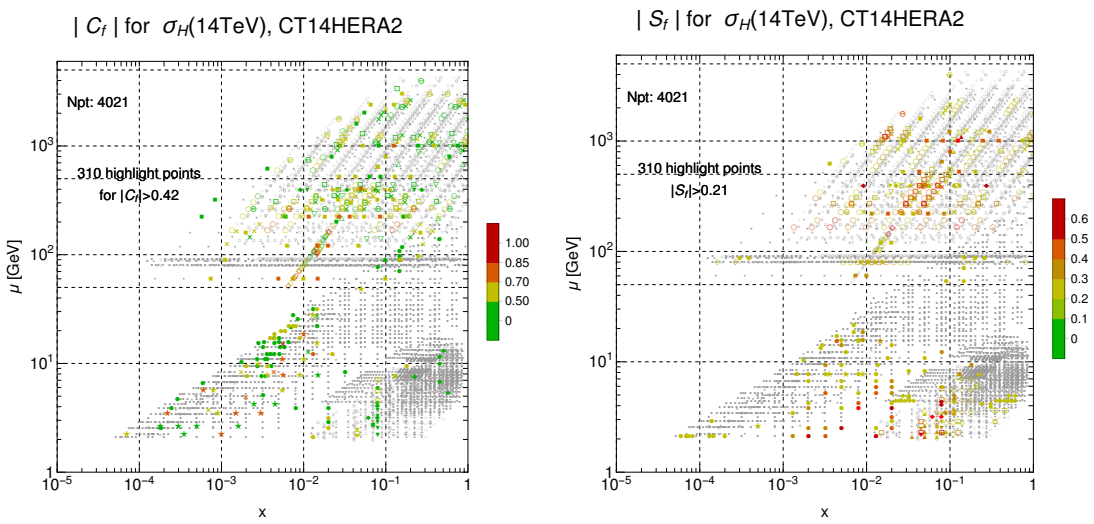

Figure 1: For the full CTEQ-TEA data-set (CT14HERA2), we show the absolute correlation $|C_f|$ and sensitivity $|S_f|$ associated with the 14 TeV Higgs production cross section $\sigma_{H^0}(14\,\text{TeV})$. Points with significant magnitudes of $|C_f|$ and $|S_f|$ are highlighted with color; the thresholds are chosen to highlight the same effective number of points (310) in both plots. When the $|C_f|$ plot at left is considered, only a very small sub-population of jet production data (diagonal open circles and closed squares with $\mu \gtrsim 100$ GeV) exhibits significant correlations, as well as some HERA DIS and $t\bar{t}$ production data points. Conversely, the sensitivity in the right panel reveals a broader range of points constraining the Higgs cross section. Here, a larger fraction of jet production points are important (especially CMS measurements), as well as processes at smaller μ, particularly DIS data from HERA and fixed-target experiments (*e.g.*, BCDMS, NMC, CDHSW, and CCFR). Although its cumulative impact is comparatively modest, ATLAS $t\bar{t}$ production data register significant per-point sensitivities, as do CCFR F_p^2 measurements and CMS 7 TeV A_μ data; similarly, some of the high-p_T Z production information from ATLAS provide modest constraints.

1. Introduction

The determination of the nucleon's collinear parton distribution functions (PDFs) is becoming an increasingly precise discipline with the advent of high-luminosity experiments at both colliders and fixed-target facilities, and several research groups are involved in the rich research domain of modern PDF analysis. PDFs provide a description of hadronic structure, and are an essential ingredient in perturbative QCD computations.[1] Since the start of the Large Hadron Collider Run II (LHC Run II), the volume of experimental data pertinent to the PDFs is growing with such speed that isolating measurements of greatest impact presents a significant challenge for PDF fitters. To help address this challenge, we present a new analysis method for identifying high-impact experiments which constrain the PDFs and the resulting Standard Model (SM) predictions that depend on them; this method is complementary to other frameworks like Hessian profiling techniques [2] and Bayesian reweighting [3, 4].

2. Correlations

The notion of using correlations between the PDF uncertainties of two physical observables was proposed in Refs. [5, 6] as a means of quantifying the degree to which these quantities were

[1] See Ref. [1] for additional details and a complete set of references.

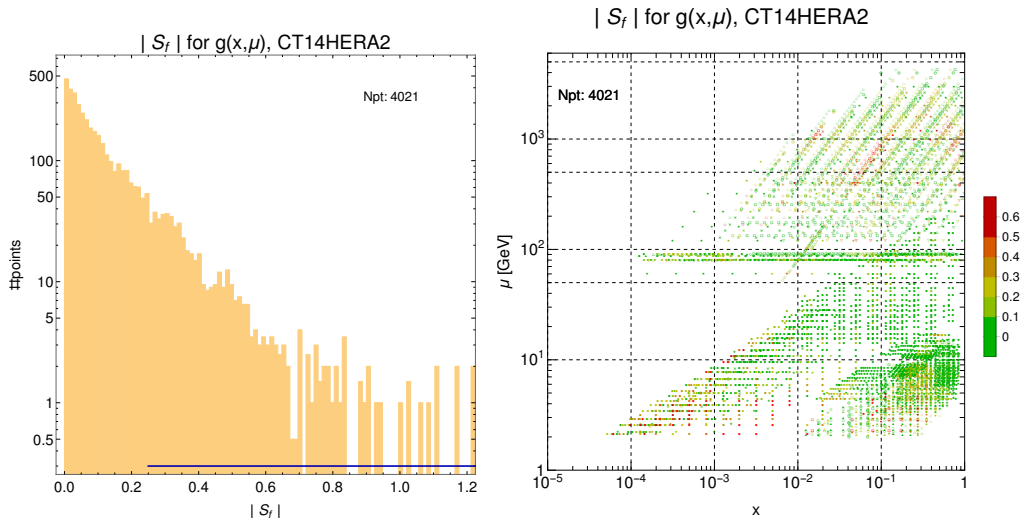

Figure 2: Two representations of the sensitivity $|S_g|(x_i,\mu_i)$ to the gluon PDF $g(x,\mu)$ of the experimental measurements making up the augmented CTEQ-TEA data-set; in the left panel we plot a histogram showing the distribution of sensitivities for 4021 physical measurements. In the right panel we show the $\{x_i,\mu_i\}$ map corresponding to these data within the full data-set.

related based upon their underlying PDFs. The PDF-mediated correlation C_f can determine whether there *may* exist a predictive relationship between the PDF f and goodness of fit to the i^{th} data point; it is defined as:[2]

$$C_f \equiv \text{Corr}[f,r_i] = \frac{\vec{\nabla} f \cdot \vec{\nabla} r_i}{\Delta f\, \Delta r_i} \,. \tag{2.1}$$

We have suggestively inserted f and r_i as arguments of the correlation function where f is a PDF and r_i is the residual constructed as $r_i = [T_i - D_i^{sh}]/s_i$. We take T_i as the theory prediction, D_i^{sh} is the datum shifted by the systematic uncertainties and s_i is the uncorrelated uncertainty; see Ref. [1] for a compete definition including the details of correlated uncertainties.

The Hessian correlation was deployed extensively in Ref. [7] to explore implications of the CTEQ6.6 PDFs for envisioned LHC observables; in this context it proved to be instrumental for identifying the specific PDF flavors and x ranges most tied to the PDF uncertainties for W, Z, H, and $t\bar{t}$ production cross sections as well as other processes. At the same time, C_f alone does not fully encode the potential impact of measurements on improving PDF determinations in terms of uncertainty reduction, particularly since the correlation of Eq. (2.1) does not significantly depend on the *size* of experimental errors.

3. Sensitivity

As a remedy to these limitations, we introduce a generalization of the PDF-mediated correlations called the *sensitivity* S_f; this object better identifies those experimental data points that tightly constrain PDFs both by merit of their inherent precision and their ability to discriminate among PDF error fluctuations. Such an approach can aid in identifying regions of $\{x,\mu\}$ in which the PDFs are particularly constrained by physical observables.

[2]Here, the gradients $\vec{\nabla} f$ and r.m.s. sums Δf are computed in the PDF parameter space.

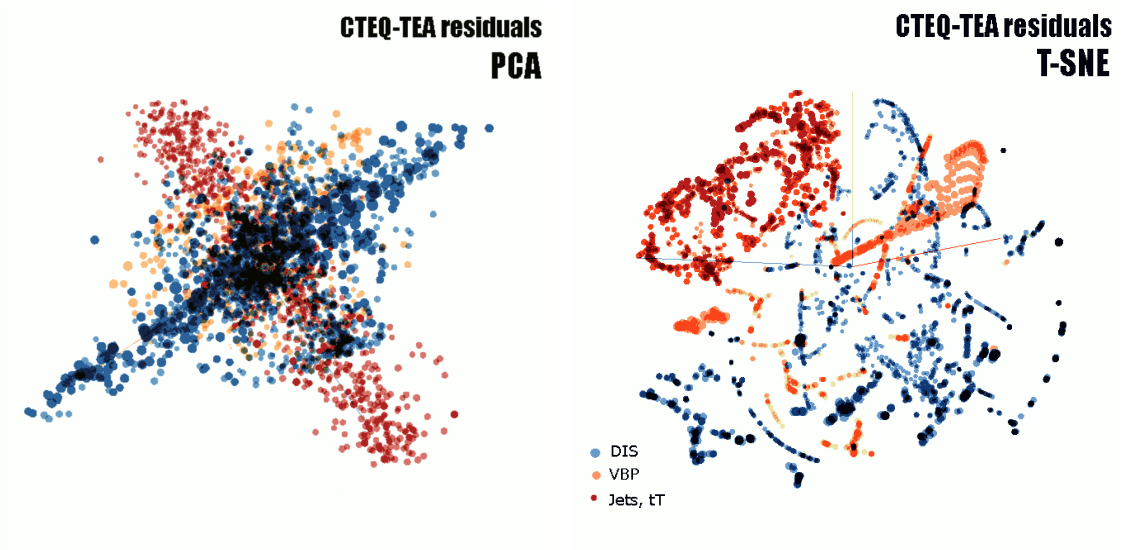

Figure 3: Distributions of displaced residuals $\vec{\delta}_i$ from the CTEQ-TEA analysis obtained by dimensionality reduction methods. Left: a 3-dimensional projection of a 10-dimensional manifold constructed by principal component analysis (PCA). Right: a distribution from the 3-dimensional t-SNE clustering method. Blue, orange, and red colors indicate data points from DIS, vector boson production, and jet/$t\bar{t}$ production processes, respectively.

Thus, we define the *sensitivity* S_f to the PDF f of the i^{th} point in experiment E to be:

$$S_f \equiv \frac{\vec{\nabla} f \cdot \vec{\nabla} r_i}{\Delta f \, \langle r_0 \rangle_E} = \frac{\Delta r_i}{\langle r_0 \rangle_E} C_f \, , \tag{3.1}$$

where Δr_i represents the variation of the residuals across the set of Hessian error PDFs, and we normalize it to the r.m.s. residual for the experiment E data-set, $\langle r_0 \rangle_E$, to reduce the impact of random fluctuations in the data values.

This definition has the benefit of encoding not only the correlated relationship of f with r_i, but also the comparative size of the experimental uncertainty with respect to the PDF uncertainty. For example, if new experimental data have reported uncertainties that are much tighter than the present PDF errors, these data would then register as high-sensitivity points.

In fact, in the numerical approach the user can quantify the sensitivity of data not only to individual PDF flavors, but even to specific physical observables, including the modifications due to correlated systematic uncertainties in every experiment of the expanded CTEQ-TEA analysis. For example, for 14 TeV Higgs boson production via gluon fusion ($gg \to H$) at the LHC, the short-distance cross sections are known up to N^{3}LO with a scale uncertainty of about 3% [8]. It has been suggested that $t\bar{t}$ production and high-p_T Z boson production already provide comparable constraints on the gluon PDF in the x region sensitive to LHC production, and that these are comparable to the constraints from LHC and Tevatron data [9, 10]. Verifying the degree to which this hypothesis is true has been difficult without actually including all these data in a fit.

As an alternative to doing a full global fit, we can critically assess this supposition using the Hessian correlations and sensitivities, $|C_f|$ and $|S_f|$, associated with the Higgs production cross section σ_{H^0}, in the context of the updated CTEQ-TEA set that includes the CT14HERA2 [11]

points and newer LHC Run I data.

Fig. 1 shows $|C_f|$ and $|S_f|$ distributions that we obtain in $\{x, \mu\}$ space. The data points that have large values of $|C_f|$ and $|S_f|$, and hence constrain the PDF dependence of σ_{H^0}, are highlighted with color according to the conventions described in Ref. [1]. The sensitivity measure generally identifies a different outlay of data providing constraints on σ_{H^0} than the correlation, as can be seen by comparing highlighted data points in the left and right panels. Note, the thresholds are chosen to highlight the same effective number of points in both plots. In the left panel, only a very small subgroup of the inclusive jet production points, select HERA Neutral Current (NC) DIS measurements, and several ATLAS p_T^Z data, show the most significant correlations, taken to have $|C_f| > 0.42$ in this comparison. With our improved definition for the sensitivity, however, the corresponding plot in the right panel demonstrates that a different collection of points has large sensitivity to σ_{H^0}, with $|S_f| > 0.21$. These data include most of the analyzed jet production data, $t\bar{t}$ and high-p_T Z production, as well as various DIS experiments. From this comparison, one would conclude that efforts to constrain PDF-based SM predictions for Higgs production relying only on a few points of $t\bar{t}$ data would be significantly handicapped by the neglect of high-energy jet production points.

4. Manifold learning and dimensionality reduction

Lastly, we illustrate a possible analysis technique carried out with the help of the TensorFlow Embedding Projector software for the visualization of high-dimensional data [1, 12]. We operate on a table of 4021 vectors $\vec{\delta}_i$ defined from theory-data residuals according to $\vec{\delta}_i = \delta_{i,k} = [r_i(f_k) - r_i(f_0)]/\langle r_0 \rangle_E$ with $\langle r_0 \rangle_E$ representing the *rms*-averaged residual of the central fit evaluated over experiment E. This quantity can be computed for the CTEQ-TEA data-set (corresponding to our total number of raw data points) and is generated by our package PDFSENSE and uploaded to the Embedding Projector website. As variations along many eigenvector directions result only in small changes to the PDFs, the 56-dimensional $\vec{\delta}_i$ vectors can in fact be projected onto an effective manifold spanned by fewer dimensions. Specifically, the Embedding Projector approximates the 56-dimensional manifold by a 10-dimensional manifold using principal component analysis (PCA). In practice, this 10-dimensional manifold is constructed out of the 10 components of greatest variance in the effective space, such that the most variable combinations of $\delta_{i,l}$ are retained, while the remaining 46 components needed to fully reconstruct the original 56-dimensional $\vec{\delta}_i$ are discarded. However, because the 10 PCA-selected components describe the bulk of the variance of $\delta_{i,l}$, the loss of these 46 components results in only a minimal relinquishment of information, and in fact provides a more efficient basis to study $\delta_{i,l}$ variations.

In the 10-dimensional PCA representation, some directions result in efficient separation of residuals of different types. For example, the left panel of Fig. 3 shows a 3-dimensional projection of the $\vec{\delta}_i$ that separates clusters of DIS, vector boson production, and jet/$t\bar{t}$ production residuals. In this example, the jet/$t\bar{t}$ cluster, shown in red, is roughly orthogonal to the blue DIS cluster and intersects it. This separation is remarkable, as it is based only on numerical properties of the $\vec{\delta}_i$ vectors, and not on the meta-data about the types of experiments that is entered after the PCA is completed.

As an alternative, the Embedding Projector can organize the $\vec{\delta}_i$ vectors into clusters according to their similarity using t-distributed stochastic neighbor embedding (t-SNE). A representative

3-dimensional distribution of the vectors obtained by t-SNE is displayed in the right panel of Fig. 3. In this case, we find that such algorithms can again sort data into clusters according to the experimental process, values of x and μ, and even the experiment itself; for other examples, including animations, see Ref. [12].

5. Conclusions

We have confronted the modern challenge of a rapidly growing set of global QCD data with new statistical methodologies for quantifying and exploring the impact of this information. These novel methodologies are realized in a new analysis tool PDFSENSE[1], which allows the rapid exploration of the impact of both existing and potential data on PDF determinations. Crucial to this analysis is introduction of the sensitivity S_f which serves as a particularly powerful discriminator; both this and the correlation C_f allow us to visualize PDF constraints provided by data across a wide range in $\{x, \mu\}$. While we have demonstrated these techniques in the context of the CT14 family of global fits, they are of sufficient generality that one could readily repeat our analysis using alternative PDF sets. These various tools collectively suggest a number of possible avenues to advance PDF knowledge in the coming years.

References

[1] Bo-Ting Wang, T. J. Hobbs, S. Doyle, J. Gao, T.J. Hou, P. M. Nadolsky, and F. I. Olness. Mapping the sensitivity of hadronic experiments to nucleon structure. arXiv:1803.02777.

[2] S. Camarda et al., QCD analysis of W- and Z-boson production at Tevatron. *Eur. Phys. J.*, C75(9):458, 2015.

[3] Richard D. Ball, *et al.*, Reweighting NNPDFs: the W lepton asymmetry. *Nucl. Phys.*, B849:112–143, 2011. [Erratum: Nucl. Phys.B855,927(2012)].

[4] Richard D. Ball, *et al.*, Reweighting and Unweighting of Parton Distributions and the LHC W lepton asymmetry data. *Nucl. Phys.*, B855:608–638, 2012.

[5] J. Pumplin, *et al.*, Uncertainties of predictions from parton distribution functions. 2. The Hessian method. *Phys. Rev.*, D65:014013, 2001.

[6] Pavel M. Nadolsky and Z. Sullivan. PDF uncertainties in WH production at Tevatron. *eConf*, C010630:P510, 2001. hep-ph/0110378.

[7] Pavel M. Nadolsky, *et al.*, Implications of CTEQ global analysis for collider observables. *Phys. Rev.*, D78:013004, 2008.

[8] Charalampos Anastasiou, *et al.*, High precision determination of the gluon fusion Higgs boson cross-section at the LHC. *JHEP*, 05:058, 2016.

[9] Michal Czakon, *et al.*, Pinning down the large-x gluon with NNLO top-quark pair differential distributions. *JHEP*, 04:044, 2017.

[10] Radja Boughezal, Alberto Guffanti, Frank Petriello, and Maria Ubiali. The impact of the LHC Z-boson transverse momentum data on PDF determinations. *JHEP*, 07:130, 2017.

[11] Tie-Jiun Hou, *et al.*, CTEQ-TEA parton distribution functions and HERA Run I and II combined data. *Phys. Rev.*, D95(3):034003, 2017.

[12] Dianne Cook, Ursula Laa, and German Valencia. Dynamical projections for the visualization of PDFSense data. arXiv:1806.09742.