PDF Calculations for the LHC

HiggsTools Second Annual Meeting

Zahari Kassabov

April 13th 2016

University of Turin, University of Milan

Introduction

Need to calculate some PDF dependent quantity (PDF+α_S uncertainty).
What should I do?

Which PDF set should I use?

Introduction

Need to calculate some PDF dependent quantity (PDF+α_S uncertainty). What should I do?

Which PDF set should I use?

The 2015 PDFLHC15 Recommendation [Butterworth et al, 1510.0386] provides some guidelines.

Introduction

Need to calculate some PDF dependent quantity (PDF+α_S uncertainty). What should I do?

Which PDF set should I use?

The 2015 PDFLHC15 Recommendation [Butterworth et al, 1510.0386] provides some guidelines.

What is beyond it?

PDF4LHC Recommendation

Criteria for inclusion

The set of conditions for a PDF determination to be included are:

- Be based on a *global* dataset.

Criteria for inclusion

The set of conditions for a PDF determination to be included are:

- Be based on a *global* dataset.
- Provide NNLO and predictions in a GM-VFNS.

Criteria for inclusion

The set of conditions for a PDF determination to be included are:

- Be based on a *global* dataset.
- Provide NNLO and predictions in a GM-VFNS.
- α_S set to the PDG average (and consistent quark masses).

Criteria for inclusion

The set of conditions for a PDF determination to be included are:

- Be based on a *global* dataset.
- Provide NNLO and predictions in a GM-VFNS.
- α_S set to the PDG average (and consistent quark masses).
- Account for all known sources of uncertainty.

Criteria for inclusion

The set of conditions for a PDF determination to be included are:

- Be based on a *global* dataset.
- Provide NNLO and predictions in a GM-VFNS.
- α_S set to the PDG average (and consistent quark masses).
- Account for all known sources of uncertainty.

Benchmarks have shown that these features account for most of the differences with other determinations.

Although not everyone agrees [Accardi et al, 1603.08906].

The PDF4LHC recommendation

Recent progress has improved the agreement between the PDFs by the CT, MMHT and NNPDF groups.

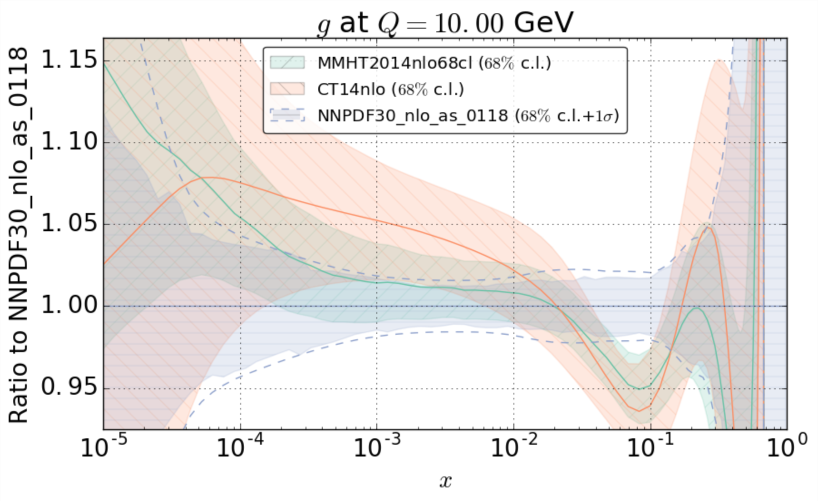
- Added new data which constrains the PDFs.

The PDF4LHC recommendation

Recent progress has improved the agreement between the PDFs by the CT, MMHT and NNPDF groups.

- Added new data which constrains the PDFs.
- Better understanding of parametrization and fitting issues.
  - More general parameterization for MMHT and CT
  - Closure tests for NNPDF.

Agreement in the gluon PDF



Differences between included sets

The agreement between these sets suggests that the spread could be interpreted in a statistical way, including a crude *theoretical uncertainty* estimate. Some differences are:

- The parametrization and fitting procedure.
- Some theoretical parameters (quark masses).
- The precise implementation of the GM-VFNS.
- The way in which the DGLAP evolution equations are solved.

Therefore a statistical combination is deemed desirable.

Statistical combination

- The combination is based on sets of 300 Monte Carlo replicas obtained from MMHT14, CT14 and NNPDF3.0.
- This makes 900 computations necessary to estimate the PDF uncertainty of a given observable, which is *impractical*.

Reduced representations

Three methods were implemented for producing compressed sets:

PDF4LHC15_mc [Carrazza et al, 1504.06469] Monte Carlo set to account for Non-Gaussian effects.

Reduced representations

Three methods were implemented for producing compressed sets:

PDF4LHC15_mc [Carrazza et al, 1504.06469] Monte Carlo set to account for Non-Gaussian effects.

PDF4LHC15_30 [Gao, Nadolsky, 1401.0013] Lower accuracy Hessian set

Reduced representations

Three methods were implemented for producing compressed sets:

PDF4LHC15_mc [Carrazza et al, 1504.06469] Monte Carlo set to account for Non-Gaussian effects.

PDF4LHC15_30 [Gao, Nadolsky, 1401.0013] Lower accuracy Hessian set

PDF4LHC_100 [Carrazza et al, 1505.06736] Higher accuracy Hessian set (see last year's talk).

Reduced representations

Three methods were implemented for producing compressed sets:

PDF4LHC15_mc [Carrazza et al, 1504.06469] Monte Carlo set to account for Non-Gaussian effects.

PDF4LHC15_30 [Gao, Nadolsky, 1401.0013] Lower accuracy Hessian set

PDF4LHC_100 [Carrazza et al, 1505.06736] Higher accuracy Hessian set (see last year's talk).

These sets have been benchmarked extensively in the Les Houches 2015 Proceedings.

Accuracy of the uncertainty band

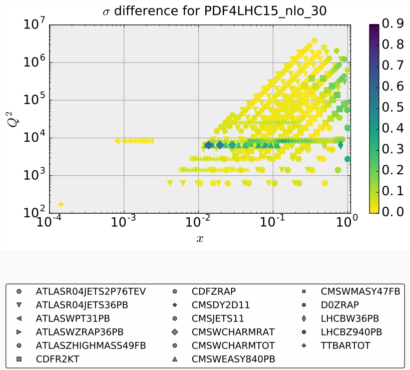
We calculated the PDF uncertainties for each reduced set and the prior, for all the hadronic observables in the NNDPF dataset:

$$\Delta_i =: \frac{\left| s_i^{(\text{prior})} - s_i^{(\text{reduced})} \right|}{s_i^{(\text{prior})}}, \; i = 1, \ldots, N_{\text{data}}$$

Accuracy of the uncertainty band

We calculated the PDF uncertainties for each reduced set and the prior, for all the hadronic observables in the NNDPF dataset:

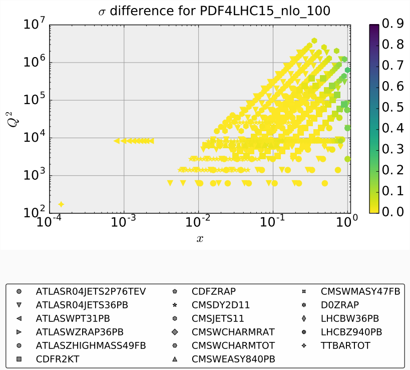
$$\Delta_i =: \frac{\left| s_i^{(\text{prior})} - s_i^{(\text{reduced})} \right|}{s_i^{(\text{prior})}}, \; i = 1, \ldots, N_{\text{data}}$$



Accuracy of the uncertainty band

We calculated the PDF uncertainties for each reduced set and the prior, for all the hadronic observables in the NNDPF dataset:

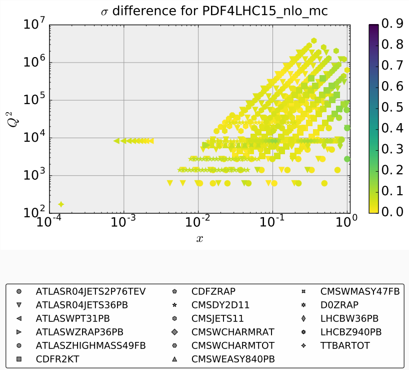
$$\Delta_i =: \frac{\left|s_i^{(\text{prior})} - s_i^{(\text{reduced})}\right|}{s_i^{(\text{prior})}},\ i = 1, \ldots, N_{\text{data}}$$



Accuracy of the uncertainty band

We calculated the PDF uncertainties for each reduced set and the prior, for all the hadronic observables in the NNDPF dataset:

$$\Delta_i =: \frac{\left| s_i^{(\text{prior})} - s_i^{(\text{reduced})} \right|}{s_i^{(\text{prior})}},\ i = 1, \ldots, N_{\text{data}}$$



Non-Gaussianities in the PDF4LHC combination

We investigated how well can the predictions of the PDF4LHC set be approximated by a Gaussian.

- We computed predictions for a representative set of LHC processses.

Non-Gaussianities in the PDF4LHC combination

We investigated how well can the predictions of the PDF4LHC set be approximated by a Gaussian.

- We computed predictions for a representative set of LHC processses.
- We investigate:
    - Effect of non-Gaussianity on the compression quality.
    - Check that compressed Monte Carlo reproduces non-Gaussianities well.
    - When it should be favored over the Hessian.

Kernel Density Estimate

First we construct a continuous probability density from a Monte Carlo sample (*Kernel Density Estimate*):

$$P(\sigma) = \frac{1}{N_{\text{rep}}} \sum_{i=1}^{N_{\text{rep}}} K\left(\sigma - \sigma_i\right) . \tag{1}$$

We choose the function K to be a normal distribution

$$K(\sigma - \sigma_i) = \frac{1}{h\sqrt{2\pi}} e^{\frac{-(\sigma - \sigma_i)^2}{h}} , \tag{2}$$

here we set the parameter h (*bandwidth*) so that it is the optimal choice if the underlying data was Gaussian

Comparing probability distributions

We use the Kullback–Leibler divergence to measure how much information we are loosing by approximating the prior $P(\sigma)$ with the distribution spanned form each of the optimized representations $Q(\sigma)$.

$$D_{KL}(P|Q) = \int_{-\infty}^{+\infty} P(\sigma) \left(\frac{\log P(\sigma)}{\log Q(\sigma)} \right) d\sigma$$

Comparing probability distributions

We use the Kullback–Leibler divergence to measure how much information we are loosing by approximating the prior $P(\sigma)$ with the distribution spanned form each of the optimized representations $Q(\sigma)$.

$$D_{KL}(P|Q) = \int_{-\infty}^{+\infty} P(\sigma) \left(\frac{\log P(\sigma)}{\log Q(\sigma)} \right) d\sigma$$

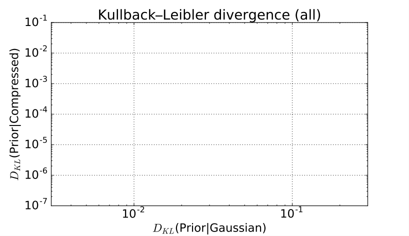
- A Gaussian given by $(\mu = \langle \sigma_i \rangle_i$, $\sigma = \frac{1}{N-1} \left(\sum (\sigma_i - \mu)^2 \right)^{1/2})$.
- The MCH Gaussian.
- The CMC KDE.

Comparing probability distributions

We use the Kullback–Leibler divergence to measure how much information we are loosing by approximating the prior $P(\sigma)$ with the distribution spanned form each of the optimized representations $Q(\sigma)$.

$$D_{KL}(P|Q) = \int_{-\infty}^{+\infty} P(\sigma) \left(\frac{\log P(\sigma)}{\log Q(\sigma)} \right) d\sigma$$

- A Gaussian given by $(\mu = \langle \sigma_i \rangle_i$, $\sigma = \frac{1}{N-1} \left(\sum (\sigma_i - \mu)^2 \right)^{1/2})$.
- The MCH Gaussian.
- The CMC KDE.

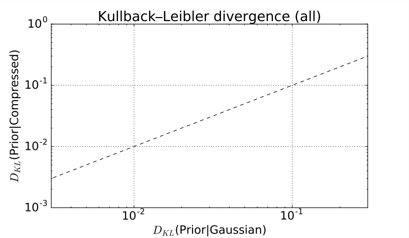


Comparing probability distributions

We use the Kullback–Leibler divergence to measure how much information we are loosing by approximating the prior $P(\sigma)$ with the distribution spanned form each of the optimized representations $Q(\sigma)$.

$$D_{KL}(P|Q) = \int_{-\infty}^{+\infty} P(\sigma) \left(\frac{\log P(\sigma)}{\log Q(\sigma)} \right) d\sigma$$

- A Gaussian given by $(\mu = \langle \sigma_i \rangle_i$, $\sigma = \frac{1}{N-1} \left(\sum (\sigma_i - \mu)^2 \right)^{1/2})$.
- The MCH Gaussian.
- The CMC KDE.

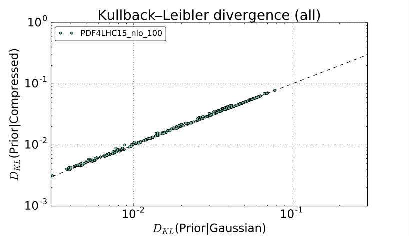


Comparing probability distributions

We use the Kullback–Leibler divergence to measure how much information we are loosing by approximating the prior $P(\sigma)$ with the distribution spanned form each of the optimized representations $Q(\sigma)$.

$$D_{KL}(P|Q) = \int_{-\infty}^{+\infty} P(\sigma) \left(\frac{\log P(\sigma)}{\log Q(\sigma)} \right) d\sigma$$

- A Gaussian given by $(\mu = \langle \sigma_i \rangle_i$, $\sigma = \frac{1}{N-1} \left(\sum (\sigma_i - \mu)^2 \right)^{1/2})$.
- The MCH Gaussian.
- The CMC KDE.

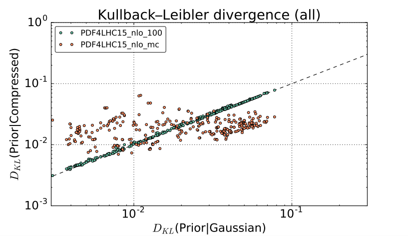


Comparing probability distributions

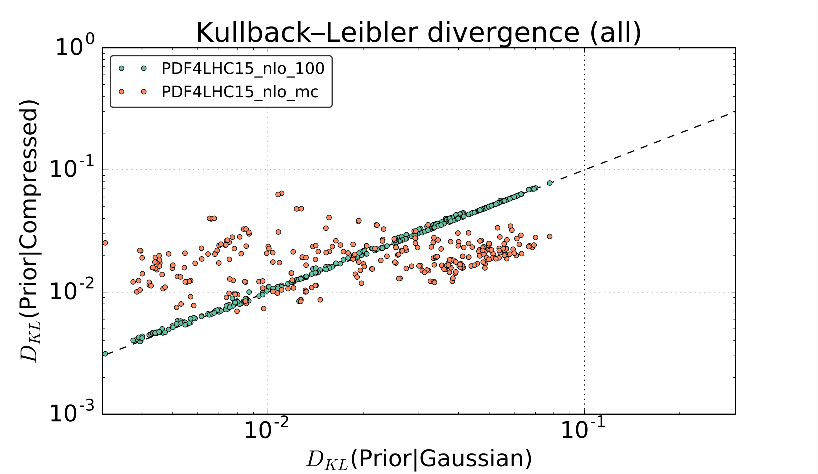
We use the Kullback–Leibler divergence to measure how much information we are loosing by approximating the prior $P(\sigma)$ with the distribution spanned form each of the optimized representations $Q(\sigma)$.

$$D_{KL}(P|Q) = \int_{-\infty}^{+\infty} P(\sigma) \left(\frac{\log P(\sigma)}{\log Q(\sigma)} \right) d\sigma$$

- A Gaussian given by $(\mu = \langle \sigma_i \rangle_i$, $\sigma = \frac{1}{N-1} \left(\sum (\sigma_i - \mu)^2 \right)^{1/2})$.
- The MCH Gaussian.
- The CMC KDE.

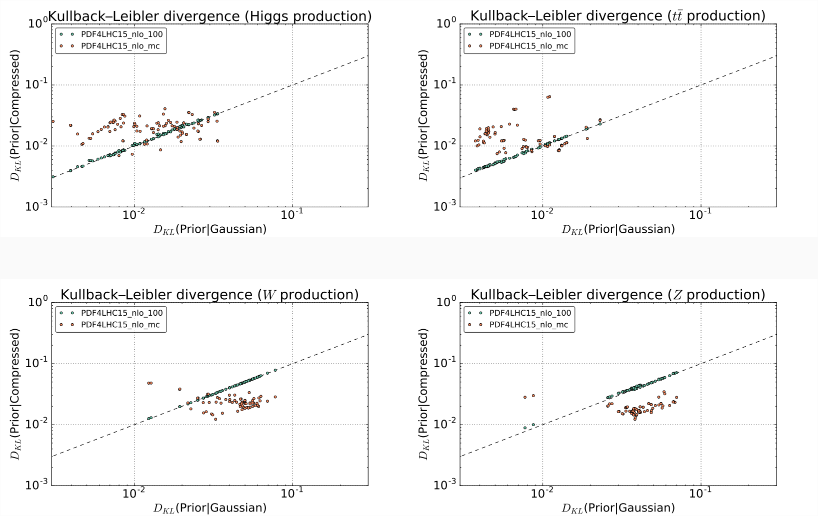


Results



- MCH as close to the result of the prior as a Gaussian can be.
- CMC $\sim$ independent of the degree of Gaussianity.

Process by process



Absolute scale

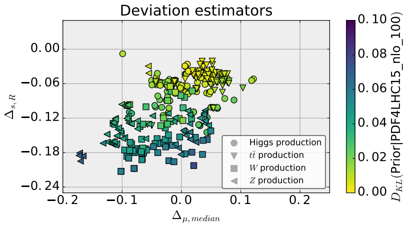
We define the estimators:

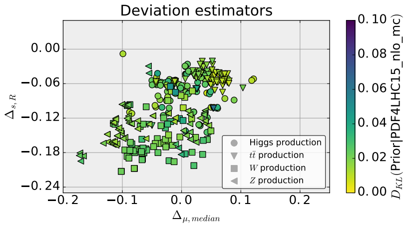
$$\Delta_\mu = \frac{\mathrm{median} - \mu}{s}$$
$$\Delta_s = \frac{R - s}{s}$$

with

$$R = \frac{1}{2}\min\{[x_{\mathrm{min}}, x_{\mathrm{max}}]; \qquad \int_{x_{\mathrm{min}}}^{x_{\mathrm{max}}} P(x) = 0.683\,.$$

Results


Deviation estimators
0.00
−0.06
−0.12
−0.18
−0.24
−0.2
−0.1
0.0
0.1
0.2
$\Delta_{s,R}$
$\Delta_{\mu,median}$
Higgs production
$t\bar{t}$ production
W production
Z production
0.10
0.08
0.06
0.04
0.02
0.00
D_{KL}(Prior|PDF4LHC15_nlo_100)


Deviation estimators
0.00
−0.06
−0.12
−0.18
−0.24
−0.2
−0.1
0.0
0.1
0.2
$\Delta_{s,R}$
$\Delta_{\mu,median}$
Higgs production
$t\bar{t}$ production
W production
Z production
0.10
0.08
0.06
0.04
0.02
0.00
D_{KL}(Prior|PDF4LHC15_nlo_mc)

Improved reduction methods

- General purpose reduction require ~ 100 error sets to reach a percent level accuracy (PDF4LHC_100, PDF4LHC15_mc).

Improved reduction methods

- General purpose reduction require ~ 100 error sets to reach a percent level accuracy (PDF4LHC_100, PDF4LHC15_mc).
- Further compression (PDF4LHC15_30) can induce rather large ($> 50\%$) deviations for certain observables.

Improved reduction methods

- General purpose reduction require ~ 100 error sets to reach a percent level accuracy (PDF4LHC_100, PDF4LHC15_mc).
- Further compression (PDF4LHC15_30) can induce rather large ($> 50\%$) deviations for certain observables.
- It is possible to reduce the number of error sets when we only want to reproduce a specific set of processes [Pumplin, 0904.2425] while retaining accuracy.

Improved reduction methods

- General purpose reduction require ~ 100 error sets to reach a percent level accuracy (PDF4LHC_100, PDF4LHC15_mc).
- Further compression (PDF4LHC15_30) can induce rather large ($> 50\%$) deviations for certain observables.
- It is possible to reduce the number of error sets when we only want to reproduce a specific set of processes [Pumplin, 0904.2425] while retaining accuracy.
- We developed the SMPDF [Carraza et al, 1602.00005] methodology to realize this.

SMPDF

SMPDF

- Efficient and accurate PDF process-specific Hessian reduction algorithm:

SMPDF

- Efficient and accurate PDF process-specific Hessian reduction algorithm:

$$\text{(Prior PDF, list of observables)} \longrightarrow \text{Reduced representation (SMPDF)}$$

SMPDF

- Efficient and accurate PDF process-specific Hessian reduction algorithm:

$$(\text{Prior PDF, list of observables}) \longrightarrow \text{Reduced representation } (\texttt{SMPDF})$$

- Suitable for use in experimental analysis:

SMPDF

- Efficient and accurate PDF process-specific Hessian reduction algorithm:

$$(\text{Prior PDF, list of observables}) \longrightarrow \text{Reduced representation } (\texttt{SMPDF})$$

- Suitable for use in experimental analysis:
    - Easy to combine independent SMPDFs

SMPDF

- Efficient and accurate PDF process-specific Hessian reduction algorithm:

$$\text{(Prior PDF, list of observables)} \quad \longrightarrow \quad \text{Reduced representation (SMPDF)}$$

- Suitable for use in experimental analysis:
  - Easy to combine independent SMPDFs
  - Stable against varying kinematical cuts.

SMPDF

- Efficient and accurate PDF process-specific Hessian reduction algorithm:

$$\text{(Prior PDF, list of observables)} \longrightarrow \text{Reduced representation } (\texttt{SMPDF})$$

- Suitable for use in experimental analysis:
  - Easy to combine independent SMPDFs
  - Stable against varying kinematical cuts.
  - The accuracy vs N_{eig} balance can be tuned by users.

SMPDF

- Efficient and accurate PDF process-specific Hessian reduction algorithm:

  (Prior PDF, list of observables) $\longrightarrow$ Reduced representation (SMPDF)

- Suitable for use in experimental analysis:
  - Easy to combine independent SMPDFs
  - Stable against varying kinematical cuts.
  - The accuracy vs N_{eig} balance can be tuned by users.
- Public code: https://github.com/scarrazza/smpdf

Methodology

Based on MCH:

- Write Hessian parameters in terms of a linear combination of MC replicas.

$$\boxed{X_{lk}(Q) \equiv f_\alpha^{(k)}(x_i, Q) - f_\alpha^{(0)}(x_i, Q)}$$

- Select the most relevant ones (at some scale).
- Trivially apply DGLAP evolution to the linear combination.

Suplemented with:

- Measure of the relevance

Methodology

Based on MCH:

- Write Hessian parameters in terms of a linear combination of MC replicas.

$$\boxed{X_{lk}(Q) \equiv f_\alpha^{(k)}(x_i, Q) - f_\alpha^{(0)}(x_i, Q)}$$

- Select the most relevant ones (at some scale).
- Trivially apply DGLAP evolution to the linear combination.

Suplemented with:

- Measure of the relevance

$$\rho\left(x_i, Q_\sigma, \alpha, \sigma\right) \equiv \frac{N_{\mathrm{rep}}}{N_{\mathrm{rep}}-1} \frac{\left\langle X(Q_\sigma)_{lk} \cdot \left(\sigma^{(k)} - \sigma^{(0)}\right)\right\rangle_{\mathrm{rep}} - \left\langle X(Q_{\sigma_1})_{lk}\right\rangle_{\mathrm{rep}} \cdot \left\langle \sigma^{(k)} - \sigma^{(0)}\right\rangle_{\mathrm{rep}}}{s_\alpha^{\mathrm{PDF}}(x_i, Q_\sigma) \cdot s_\sigma}$$

$$\Xi = \left\{(x_i, \alpha) : \rho\left(x_i, Q_\sigma, \alpha, \sigma\right) \geq t \cdot \rho_{\mathrm{max}}\right\}$$

$$X \longrightarrow X_\Xi\left(Q_\sigma\right)$$

Methodology

Based on MCH:

- Write Hessian parameters in terms of a linear combination of MC replicas.

$$\boxed{X_{lk}(Q) \equiv f_{\alpha}^{(k)}(x_i, Q) - f_{\alpha}^{(0)}(x_i, Q)}$$

- Select the most relevant ones (at some scale).
- Trivially apply DGLAP evolution to the linear combination.

Suplemented with:

- Measure of the relevance
- Measure of the accuracy goal

Methodology

Based on MCH:

- Write Hessian parameters in terms of a linear combination of MC replicas.

$$\boxed{X_{lk}(Q) \equiv f_\alpha^{(k)}(x_i, Q) - f_\alpha^{(0)}(x_i, Q)}$$

- Select the most relevant ones (at some scale).
- Trivially apply DGLAP evolution to the linear combination.

Suplemented with:

- Measure of the relevance
- Measure of the accuracy goal

$$T_R < \max_{i \in (1, N_\sigma)} \left| 1 - \frac{\tilde{s}_{\sigma_i}}{s_{\sigma_i}} \right|$$

Methodology

Based on MCH:

- Write Hessian parameters in terms of a linear combination of MC replicas.

$$\boxed{X_{lk}(Q) \equiv f_\alpha^{(k)}(x_i, Q) - f_\alpha^{(0)}(x_i, Q)}$$

- Select the most relevant ones (at some scale).
- Trivially apply DGLAP evolution to the linear combination.

Suplemented with:

- Measure of the relevance
- Measure of the accuracy goal
- Procedure to retain information

Methodology

Based on MCH:

- Write Hessian parameters in terms of a linear combination of MC replicas.

$$\boxed{X_{lk}(Q) \equiv f_\alpha^{(k)}(x_i, Q) - f_\alpha^{(0)}(x_i, Q)}$$

- Select the most relevant ones (at some scale).
- Trivially apply DGLAP evolution to the linear combination.

Suplemented with:

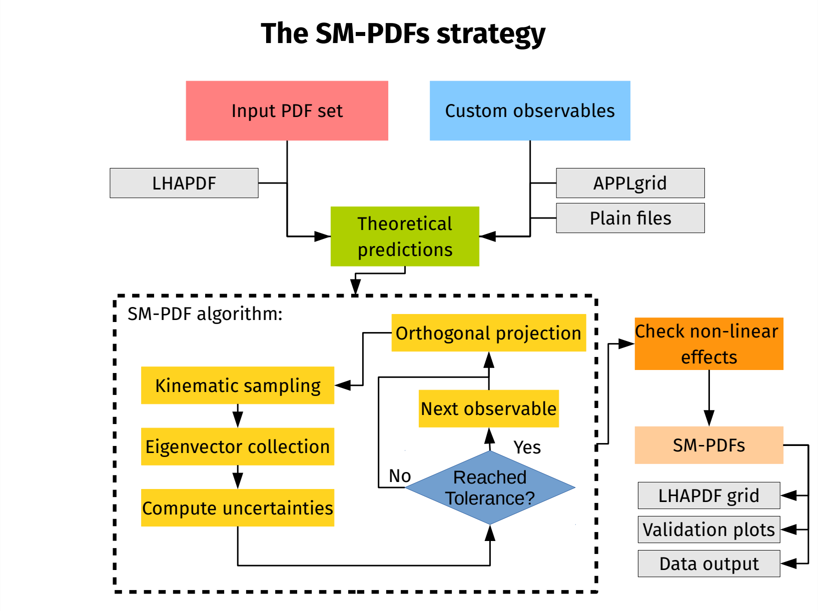
- Measure of the relevance
- Measure of the accuracy goal
- Procedure to retain information

$$X = USV^t$$

$$\begin{pmatrix} P & R \end{pmatrix} = V \in \mathbb{R}^{N_{\mathrm{rep}}} \times \begin{pmatrix} \mathbb{R}^{N_{\mathrm{eig}}} & \mathbb{R}^{N_{\mathrm{rep}} - N_{\mathrm{eig}}} \end{pmatrix}$$

$$X \longrightarrow XR$$

Iterative procedure



Example cases

- We have generated SMPDFs for the most important Higgs production processes:
    - Gluon fusion, VBF,hW, hZ, $ht\bar{t}$
- and the main backgrounds:
    - $t\bar{t}$, W and Z production
- We have considered total cross sections and various differential distributions.

Results

Process	MC900		NNPDF3.0		MMHT14	
	$T_R = 5\%$	$T_R = 10\%$	$T_R = 5\%$	$T_R = 10\%$	$T_R = 5\%$	$T_R = 10\%$
h	15	11	13	8	8	7
$t\bar{t}$	4	4	5	4	3	3
W, Z	14	11	13	8	10	9
Ladder	17	14	18	11	10	10

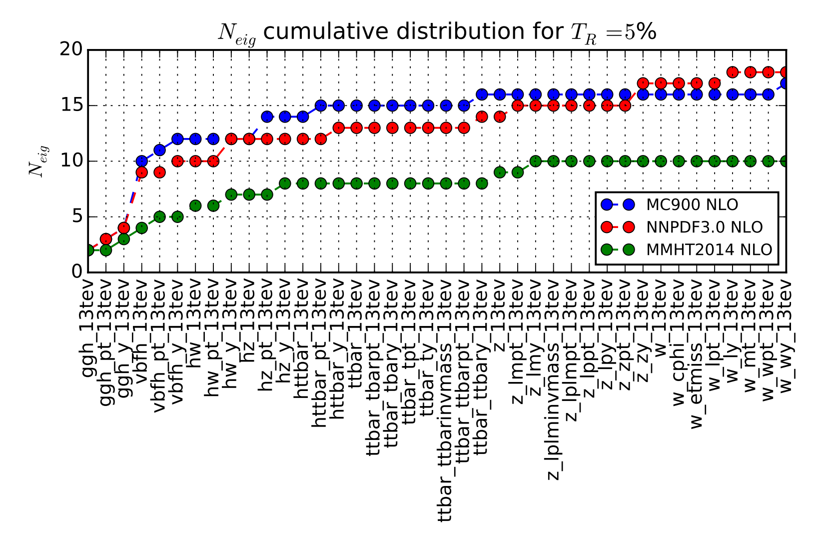
- T_R (set by user) is the maximum allowed deviation from the prior for any bin.
    - Typical difference is much smaller.

Higgs production in individual channels

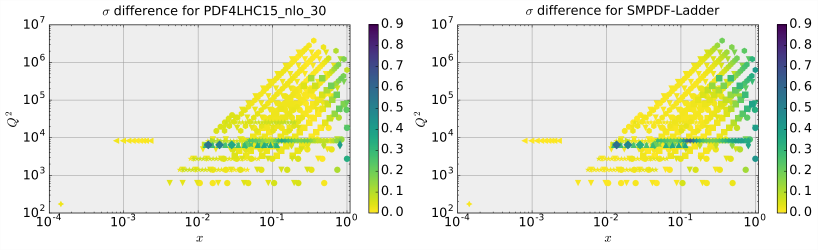
Process	MC900		NNPDF3.0		MMHT14	
	$T_R = 5\%$	$T_R = 10\%$	$T_R = 5\%$	$T_R = 10\%$	$T_R = 5\%$	$T_R = 10\%$
$gg \to h$	4	5	4	4	3	3
VBF hjj	7	5	10	5	4	3
hW	6	5	6	4	6	3
hZ	11	7	6	4	8	5
$ht\bar{t}$	3	2	4	4	3	2
Total h	15	11	13	8	8	7

Ladder SMPDF

Multiple processes can be efficiently stacked together:



Comparison of Ladder SMPDF and PDF4LHC15_30



PDFs with scale variations

Motivation

Currently theoretical (MHO) uncertainties in PDFs are:

- Not estimated at all (usually).

Motivation

Currently theoretical (MHO) uncertainties in PDFs are:

- Not estimated at all (usually).
- Estimated with crude recipes.

Motivation

Currently theoretical (MHO) uncertainties in PDFs are:

- Not estimated at all (usually).
- Estimated with crude recipes.
  - N^3LO uncertainty $\longrightarrow \frac{1}{2}(NNLO - NLO)$ result [Anastasiou et al, 1602.00695]

Motivation

Currently theoretical (MHO) uncertainties in PDFs are:

- Not estimated at all (usually).
- Estimated with crude recipes.
  - N^3LO uncertainty $\longrightarrow \frac{1}{2}(NNLO - NLO)$ result [Anastasiou et al, 1602.00695]
  - Combination of PDFs (PDF4LHC)

Application of scale variations

Standard procedure to estimate MHU.

- Uncertainty computed trough a specified set of variations. [HXSWG, 1101.0593]

Application of scale variations

Standard procedure to estimate MHU.

- Uncertainty computed trough a specified set of variations. [HXSWG, 1101.0593]
- But other procedures exist

Application of scale variations

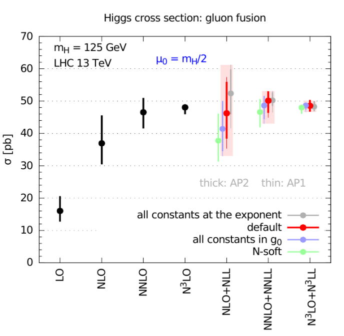
Standard procedure to estimate MHU.

- Uncertainty computed trough a specified set of variations. [HXSWG, 1101.0593]
- But other procedures exist
    - Resummation prescription variations, see [Bovini et al, 1603.08000] for a recent application.
    - Cacciari-Houdeau [Cacciari, Houdeau, 1105.5152]
    - David-Passarino [David, Passarino, 1307.1843] Estimate the sum using convergence accelaration methods.

Application of scale variations

Standard procedure to estimate MHU.

- Uncertainty computed trough a specified set of variations. [HXSWG, 1101.0593]
- But other procedures exist
- Compared for N^3LO ggH in [Bovini et al, 1603.08000].



Scale variations in PDFs

MHO determination has unique challenges for PDFs.

- Not clear how to correlate the scales of the different processes that enter in the fit.
  - e.g. μ_R for DIS and jets?
- Not clear how to relate to other sources of uncertainty.
  - e.g. we could do a set of scale variations per replica.
- Intrinsically more difficult because we have a series of functions rather than of cross-sections.

Implementing in the NNPDF framework

Roadmap:

1. Male technically possible to vary scales in the NNPDF `FastKernel` code.
2. Need to figure out how to correlate the μ_R variations among processes.
3. Need to see whether μ_R or μ_F gives the biggest contribution.
4. Figure out how to represent the resulting theoretical uncertainty.
5. Look into implementing other MHO prescriptions.

Thank you!