Parton Distribution Functions

PDF updates and tools for LHC Run II

Stefano Carrazza

PP @ LHC 2016, Università di Pisa, 18 May 2016

Theoretical Physics Department, CERN, Geneva.

Outline

1 PDF update summary

2 PDF tools for LHC Run II

(Butterworth et al., arXiv:1510.03865)

3 CMC-PDFs

(S.C. et al., arXiv:1504.06459)

4 MC2H & Meta PDFs

(S.C. et al., arXiv:1504.06736)

5 Accuracy and Gaussianity of the PDF4LHC15 combinations

(S.C., Forte, Kassabov, Rojo, LH15 proceedings)

6 SMPDF

(S.C., Forte, Kassabov, Rojo, arXiv:1601.00005)

7 Summary

PDF update summary

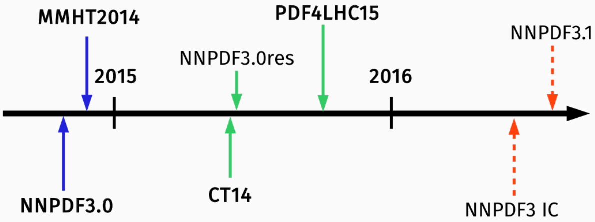
PDF current and future releases

Current results:

- Global PDF determinations: NNPDF3.0, MMHT2014 and CT14.
- NNPDF3.0res with threshold resummation
- The PDF4LHC15 recommendation

Future results:

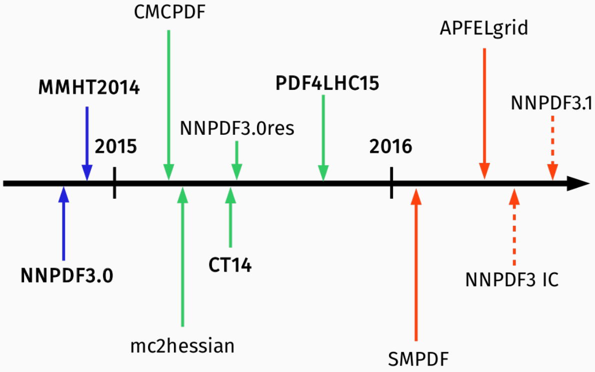
- NNPDF3 IC with intrinsic charm
- NNPDF3.1 with additional datasets from LHC



PDF releases and new tools

PDF tools for the PDF4LHC15:

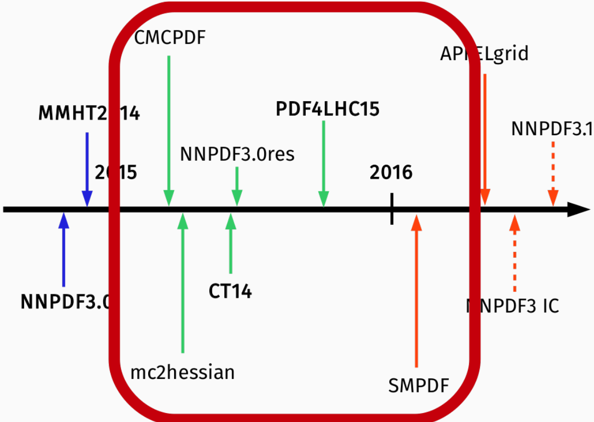
- CMC-PDFs: compression algorithm for MC PDFs.
- mc2hessian: MC to hessian conversion tool for PDFs.
- SMPDF: Specialized Minimal PDFs.



PDF releases and new tools

PDF tools for the PDF4LHC15:

- CMC-PDFs: compression algorithm for MC PDFs.
- mc2hessian: MC to hessian conversion tool for PDFs.
- SMPDF: Specialized Minimal PDFs.



PDF tools for LHC Run II

(Butterworth et al., arXiv:1510.03865)

PDF tools for LHC Run II

Problem: Need to calculate some PDF dependent quantity

$\Downarrow$

Which PDF set should I use?

$\Downarrow$

The **PDF4LHC 2015 Recommendation** provides some guidelines

Introduction

Challenge

Determine the **best combined PDF uncertainty** from **individual PDF sets**.

From 2010, the PDF4LHC WG released recommendations, updated several times to include newer versions and bug fixes.

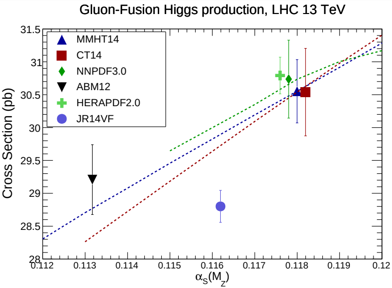
Challenge

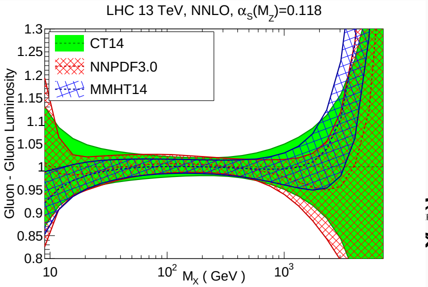
Determine the **best combined PDF uncertainty** from **individual PDF sets**.

From 2010, the PDF4LHC WG released recommendations, updated several times to include newer versions and bug fixes.

Towards the PDF4LHC15 recommendation

In 2014/2015 MMHT, CT and NNPDF **improve significantly agreement** due to new data, better theory treatment and better understanding of fitting issues.



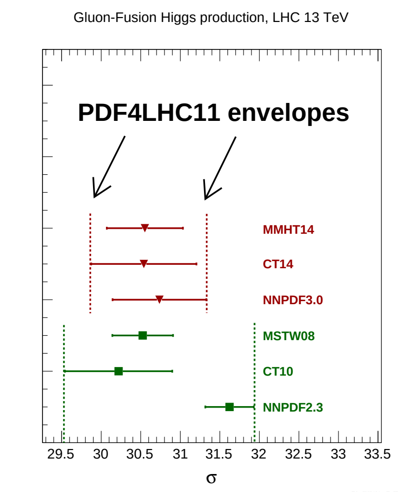


PDF4LHC recommendations

PDF4LHC11 recommendation

1. Use MSTW, CT and NNPDF PDFs
2. Take the **envelope** of uncertainties as uncertainty

- agreement was not so good, e.g. *ggH* cross section uncertainty was $>$2x the given by any individual set.
- over-conservative: **no proper statistical meaning**



PDF4LHC recommendations

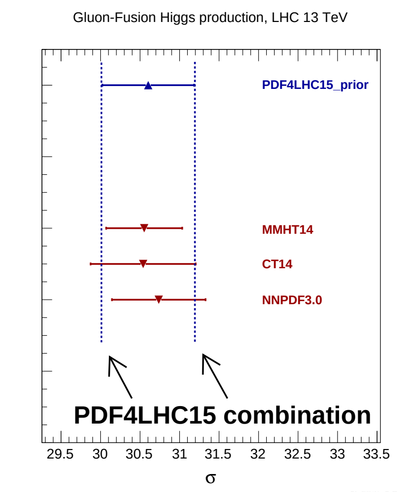
PDF4LHC11 recommendation

1. Use MSTW, CT and NNPDF PDFs
2. Take the **envelope** of uncertainties as uncertainty

- agreement was not so good, e.g. *ggH* cross section uncertainty was >2x the given by any individual set.
- over-conservative: **no proper statistical meaning**

PDF4LHC15 possibility

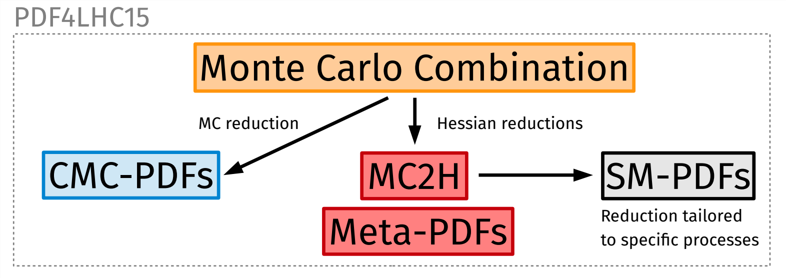
- Provide a clear **statistical interpretation**
- Deliver MC & Hessian **representations**



The PDF4LHC15 strategy

The new PDF4LHC15 prescription

1. Construct a **Monte Carlo combined** set from global PDF determinations
   - sets entering into the combination must satisfy requirements, e.g.: global datasets, use the GM-VFNS, α_s set to the PDG average.
2. Reduce **redundant information**
3. Deliver a single combined PDF set - either **Monte Carlo** or **Hessian** form.

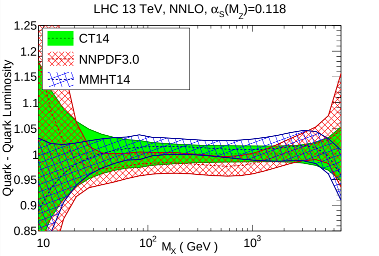


Monte Carlo combination

The combination strategy

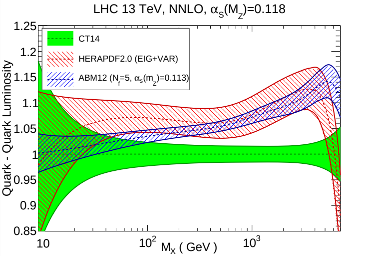
We select the PDF sets that enter the combination
⇒ must be **reasonably consistent** among them.

Global sets:



- Sets are compatible.
- Good candidate for combination.

Non global + global sets:



- Clear incompatibility.
- Little data, different evolution, characterization of uncertainty.

The PDF4LHC15 implementation

The combined sets are based on a statistical combination of:

PDF4LHC15_prior: CT14, MMHT2014 and NNPDF3.0 (MC set, $N_{rep} = 900$)

The PDF4LHC15 implementation

The combined sets are based on a statistical combination of:

PDF4LHC15_prior: CT14, MMHT2014 and NNPDF3.0 (MC set, $N_{rep} = 900$)

Reduced sets:

PDF4LHC15_*_mc: A compressed Monte Carlo set with $N_{rep} = 100$.
(*CMC-PDFs approach*, *arXiv:1504.06469*)

PDF4LHC15_*_100: A symmetric Hessian set with $N_{eig} = 100$.
(*MC2H approach*, *arXiv:1505.06736*)

PDF4LHC15_*_30: A symmetric Hessian set with $N_{eig} = 30$.
(*Meta-PDF approach*, *arXiv:1404.0013*)

Monte Carlo: contains **non-Gaussian** features important for **searches** at high masses (high x).

Hessian: useful for many experimental needs and when using **nuisance** parameters. 100 eigenvectors when **optimal precision** is needed.

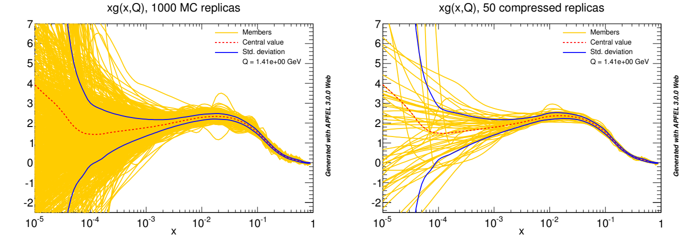
CMC-PDFs

(S.C. et al., arXiv:1504.06459)

The compression idea

Compression idea:

Reduce the size of a PDF set of Monte Carlo replicas with no/minimal **loss of information**, e.g.:



Problem: Preserve as much as possible *the underlying statistical distribution* of the prior MC PDF set.

- avoid bias in the extrapolation region.
- conserve physical requirements, e.g. positivity, correlations, etc.

The compression strategy

We define statistical estimator for the MC prior set:

1. **moments:** central value, variance, skewness and kurtosis
2. **statistical distance:** the Kolmogorov distance
3. **correlations:** between flavors at multiple x points

The compression strategy

We define statistical estimator for the MC prior set:

1. **moments:** central value, variance, skewness and kurtosis
2. **statistical distance:** the Kolmogorov distance
3. **correlations:** between flavors at multiple x points

These estimators are them **compared** to subsets of replicas **interactively** driven by an *error function*, i.e.:

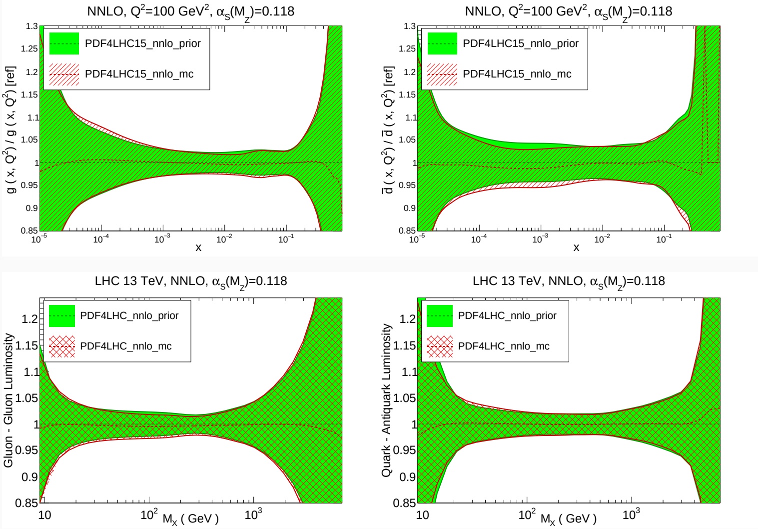
$$\mathrm{ERF} = \sum_k \frac{1}{N_k} \sum_i \left(\frac{C_i^{(k)} - O_i^{(k)}}{O_i^{(k)}} \right)^2$$

where k runs over the number of statistical estimators and

- N_k is a normalization factor extracted from random realizations
- $O_i^{(k)}$ is the value of the estimator for the prior
- $C_i^{(k)}$ is the corresponding value for the compressed set

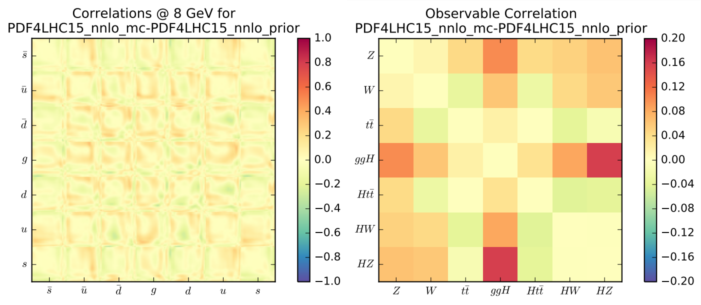
CMC-PDFs (aka PDF4LHC15_*_mc)

Good agreement between the **PDF4LHC15_prior** and **CMC-PDFs** from a number of compressed replicas $N_{rep} > 50$, e.g.:



CMC-PDFs (aka PDF4LHC15_*_mc)

Reasonable agreement as well for the **correlations** between different PDF flavours and inclusive cross-sections.



A similar number of replicas from each of the three sets is automatically selected by the compression algorithm

⇒NNPDF3.0: 23 replicas; CT14: 36 replicas, MMHT14: 32 replicas

MC2H & Meta PDFs

(S.C. et al., arXiv:1504.06736)

mc²hessian

Problem addressed here:

Determine an **unbiased Hessian representation** for **MC** PDFs.

MC2H Strategy:

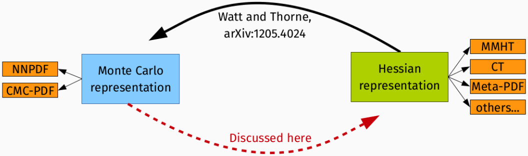
use MC replicas **themselves** as the **basis** of the linear representation.

use **Principal Component Analysis** (PCA) to reproduce PDF covariance matrix with arbitrary precision.

Meta-PDF Strategy:

each MC replica is **re-fitted** using a flexible "meta-parametrization".

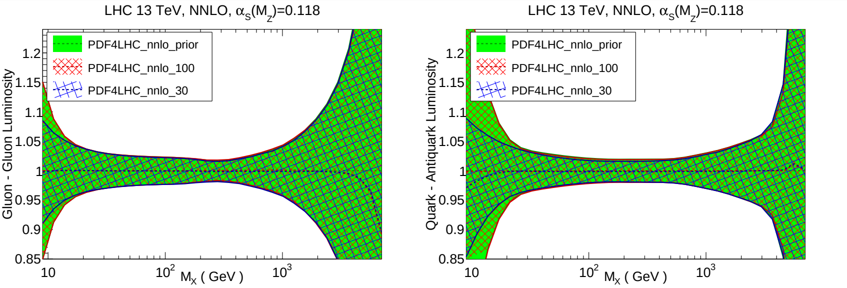
the best constrained combination are found by **diagonalization** of the covariance matrix on the PDF space



MC2H (aka PDF4LHC15_*_100)

A Hessian representation of the PDF4LHC15_prior has been constructed using

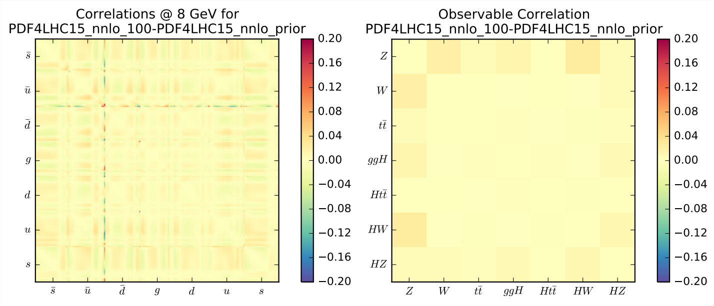
- MC2H ⇒ PDF4LHC15_*_100 with $N_{\mathrm{eig}} = 100$ (high accuracy)
- Meta-PDF ⇒ PDF4LHC15_*_30 with $N_{\mathrm{eig}} = 30$



Excellent level of agreement for PDFs and luminosities as compared with the prior.

MC2H (aka PDF4LHC15_*_100) correlations

Excellent agreement with the prior for PDF correlation and observable correlation.



Tiny residual differences at the level of few percent, irrelevant for LHC phenomenology.

Accuracy and Gaussianity of the PDF4LHC15 combinations

(S.C., Forte, Kassabov, Rojo, LH15 proceedings)

Robustness and gaussianity of the PDF4LHC15 combinations

Idea

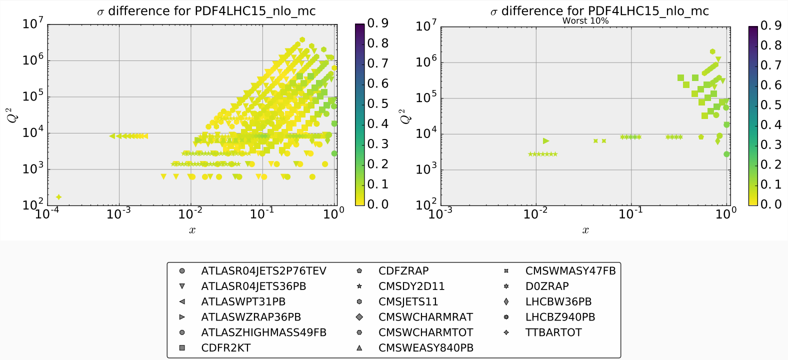
Test the **accuracy** and **Gaussianity** of the PDF4LHC15 sets.

- Verify the range of validity of prediction using data included in PDF fits.
- Discriminate Gaussianity of predictions $\Rightarrow$ verify MC vs Hessian representations.

Results elaborated for the Les Houches 2015 proceedings.

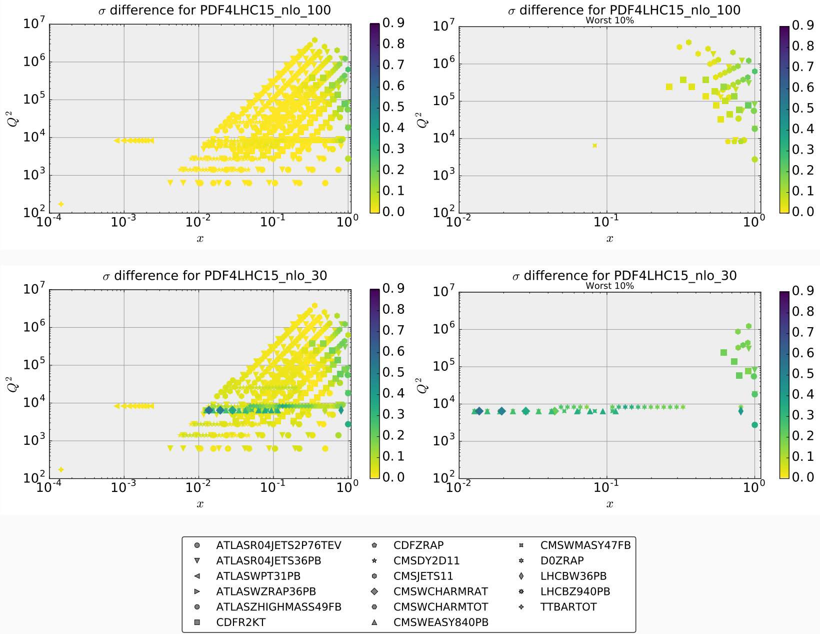
Robustness of the PDF4LHC15 combinations

We have computed predictions with **PDF4LHC15_prior** and the three reduced sets, for all data hadronic data included in the NNPDF3.0 dataset.



Deviations are generally small, and concentrated in regions in which experimental information is scarce and PDF uncertainties are largest ⇒ large x and large Q.

Robustness of the PDF4LHC15 combinations



Gaussianity of the PDF4LHC15 combinations

In order to estimate the gaussianity of predictions we construct a continuous probability density from a Monte Carlo sample (Kernel Density Estimate):

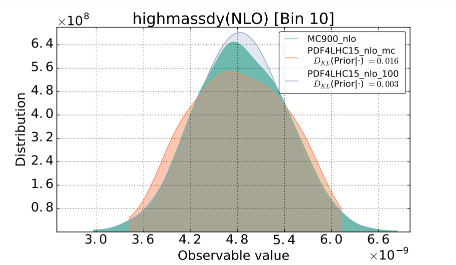
$$P(\sigma_i) = \frac{1}{N_{\rm rep}} \sum_{k=1}^{N_{\rm rep}} K(\sigma_i - \sigma_i^{(k)}), \quad i = 1, \ldots, N_{\rm dat}$$

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 $Q(\sigma)$.

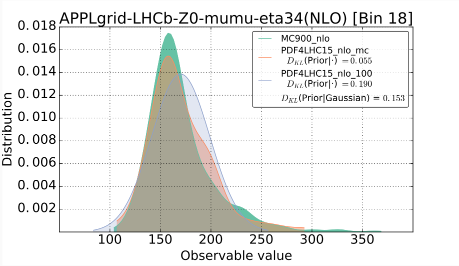
$$D_{KL}^{(i)}(P|Q) = \int_{-\infty}^{\infty} \left(P(\sigma_i) \cdot \frac{\log P(\sigma_i)}{\log Q(\sigma_i)} \right) d\sigma_i$$

Gaussianity of the PDF4LHC15 combinations

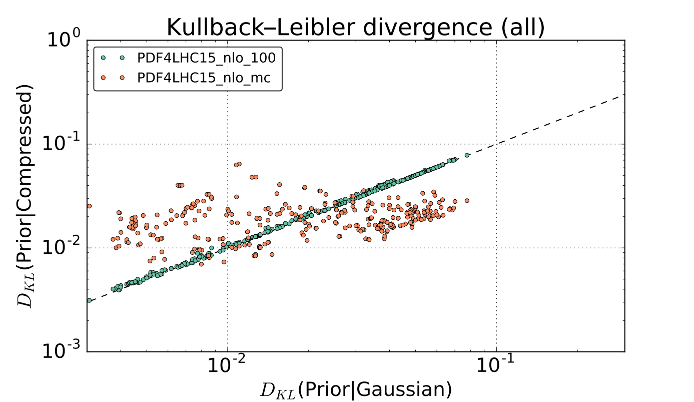
MC2H works best for Gaussian bins and when using the results as Gaussian.



CMC works best for non-Gaussian bins, when treating the results as MC.



Gaussianity of the PDF4LHC15 combinations

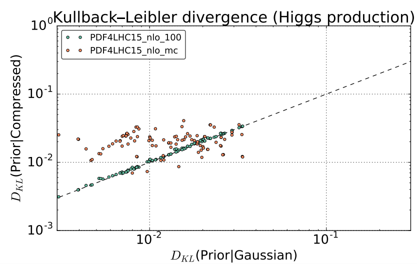


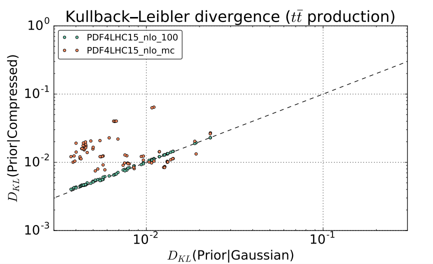
Points in the diagonal ⇒ agrees exactly with the purely Gaussian approx.

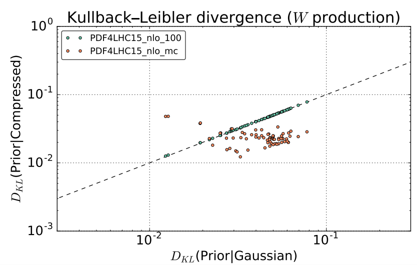
Orange points below diagonal ⇒ CMC better than MC2H

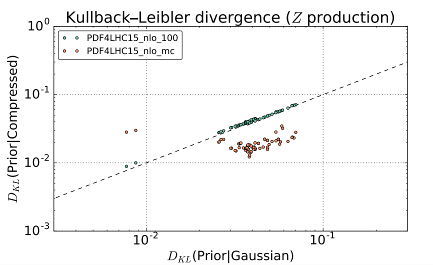
Gaussianity of the PDF4LHC15 combinations

KL divergence process by process:









SMPDF

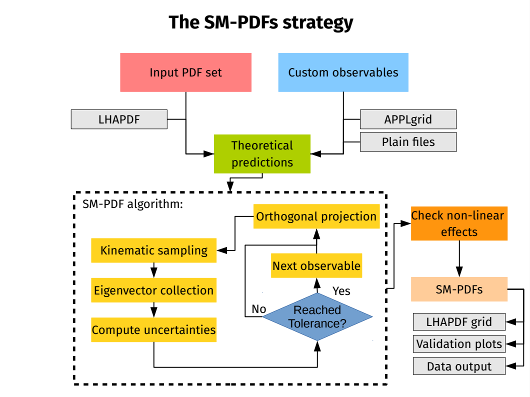
(S.C., Forte, Kassabov, Rojo, arXiv:1601.00005)

Specialized Minimal PDFs

Idea overview

Efficient and accurate PDF **process-specific** PCA Hessian reduction algorithm.

- Prior PDF, list of observables $\Longrightarrow$ Reduced representation (**SMPDF**)



Example cases

We have generated SMPDFs for the most important **Higgs prod. processes**:

Process	PDF4LHC15_prior		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

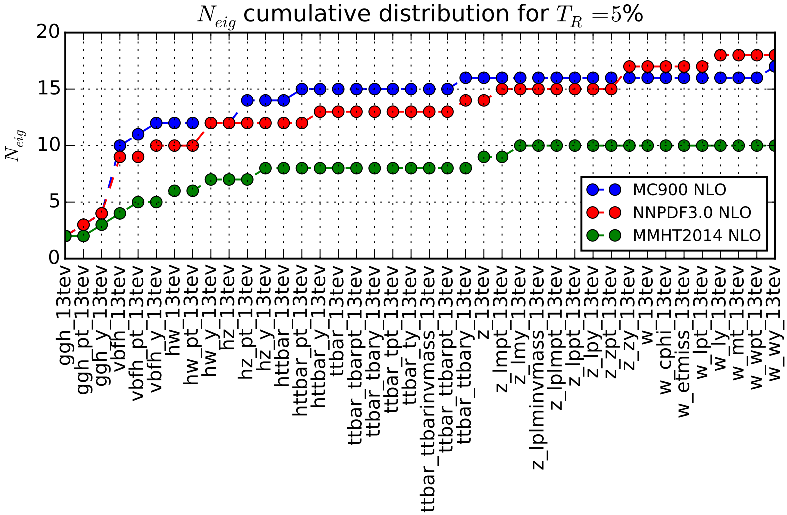
and the **main backgrounds**:

Process	PDF4LHC15_prior		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.
$\Rightarrow$ Typical difference is **much smaller**.

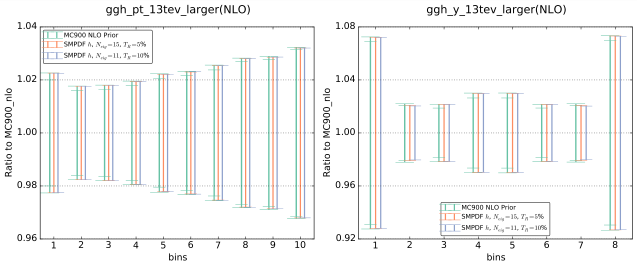
Ladder SMPDF

Multiple processes can be efficiently stacked together.

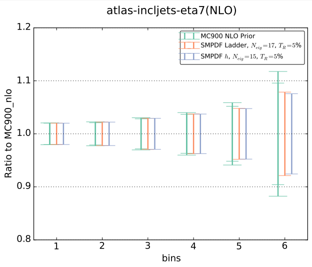


SMPDF stability

Kinematical ranges that double those used as input (p_T^h and y^h)



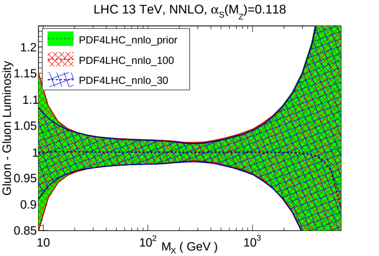
Breakdown only when going in extreme regions (large $|\eta|$):



Summary

Summary

- We presented the **PDF4LHC15 recommendation strategy**.
  - New ways of combining PDFs.
  - New **public tools** for the reduction of PDF sets are available.
- ***Public tools:***
  - CMC-PDFs: `https://github.com/scarrazza/compressor`
  - MC2H: `https://github.com/scarrazza/mc2hessian`
  - SM-PDFs: `https://github.com/scarrazza/smpdf`
- ***Future developments:***
  - Development of **compression algorithms** with better **performance**.



Thanks for your attention!

The PDF4LHC15 deliverables

LHAPDF6 grid	Pto.	ErrorType	N_{mem}	$\alpha_S(m_Z^2)$
PDF4LHC15_(n)nlo_mc	(N)NLO	replicas	100	0.118
PDF4LHC15_(n)nlo_100	(N)NLO	symmhessian	100	0.118
PDF4LHC15_(n)nlo_30	(N)NLO	symmhessian	30	0.118
PDF4LHC15_(n)nlo_mc_pdfas	(N)NLO	replicas+as	102	mem 0:100→0.118
				mem 101→0.1165
				mem 102→0.1195
PDF4LHC15_(n)nlo_100_pdfas	(N)NLO	symmhessian+as	102	mem 0:100→0.118
				mem 101→0.1165
				mem 102→0.1195
PDF4LHC15_(n)nlo_30_pdfas	(N)NLO	symmhessian+as	32	mem 0:30→0.118
				mem 31→0.1165
				mem 32→0.1195
PDF4LHC15_(n)nlo_asvar	(N)NLO	-	1	mem 0→0.1165
				mem 1→0.1195

Table 1: Summary of the combined PDF4LHC15 sets with $n_f^{\max} = 5$.

The PDF4LHC15 deliverables

LHAPDF6 grid	Pto.	ErrorType	N_{mem}	$\alpha_S^{(n_f=5)}(m_Z^2)$
PDF4LHC15_nlo_nf4_100	NLO	symmhessian	100	0.118
PDF4LHC15_nlo_nf4_30	NLO	symmhessian	30	0.118
PDF4LHC15_nlo_nf4_100_pdfas	NLO	symmhessian+as	102	mem 0:100→0.118
				mem 101→0.1165
				mem 102→0.1195
PDF4LHC15_nlo_nf4_30_pdfas	NLO	symmhessian+as	32	mem 0:30→0.118
				mem 31→0.1165
				mem 32→0.1195
PDF4LHC15_nlo_nf4_asvar	NLO	-	1	mem 0→0.1165
				mem 1→0.1195

Table 2: Summary of the combined PDF4LHC15 in the $n_f = 4$.

Monte Carlo combination

The combination strategy

1. We select the PDF sets that enter the combination
   $\Rightarrow$ must be **reasonably consistent** among them.
2. **Transform** the Hessian PDF sets into their **Monte Carlo representation** (Watt and Thorne '12):

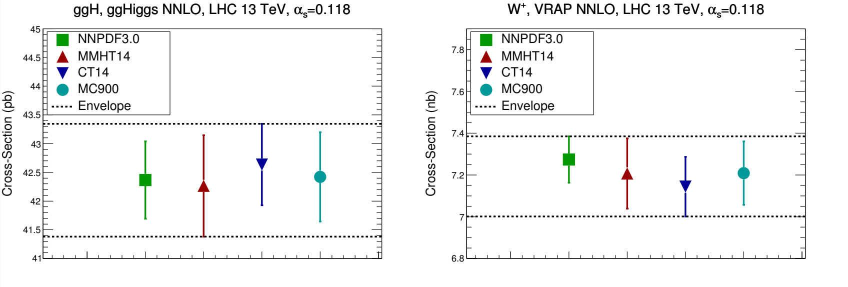
$$F^k = F(q_0) + \frac{1}{2}\sum_{j=1}^{N_{eig}} \left[F(q_j^+) - F(q_j^-)\right] R_j^k, \quad k = 1, \ldots, N_{\mathrm{rep}}$$

3. **Combine** the same number of replicas from each of the prior sets, assuming **equal weight** in the combination (i.e. an unweighted set).

PDF4LHC15: we combine $N_{rep} = 300$ replicas from NNPDF3.0, CT14 and MMHT2014, however any **other choice is possible**.

Monte Carlo combination

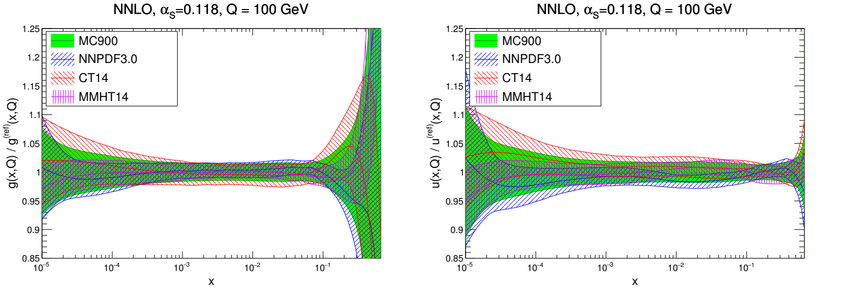
The resulting combined Monte Carlo set has **statistical properties** which lead to **smaller uncertainties** than the PDF4LHC11 envelope.



Proper treatment of **outliers** $\Rightarrow$ the envelope gives more weight to **outliers**.

Monte Carlo combination

PDF4LHC15_prior 900 MC replicas required to **stabilize** the combination.



Issues before the development of reduction strategies:

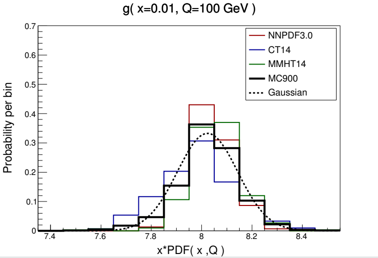
- too many replicas for practical applications ($N_{rep} = 900$)
- no possible Hessian representation
- no reduced way to preserve non-Gaussian features

Monte Carlo combination

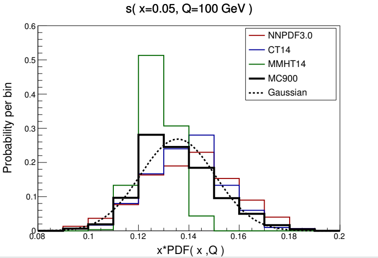
The MC combination is usually **Gaussian** but in many cases **non-Gaussian features** are observed.

Particular important for **BSM searches**, which rely on PDFs in regions where PDF errors are large.

Gaussian



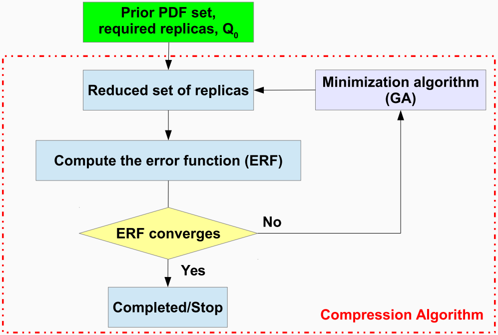
Non-Gaussian



The compression strategy

The algorithm **selects replicas** from the prior that minimize the **error function**. The minimization is driven by a **genetic algorithm**.

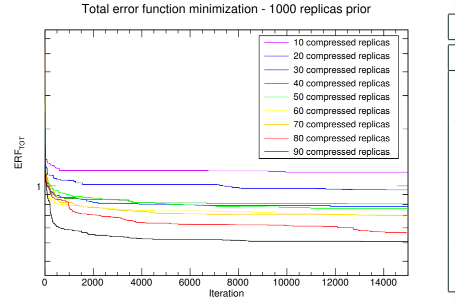
Validation: estimators, PDF plots, theoretical predictions, distances, χ^2 to experimental data, etc.



The compression strategy

Test case:

Example of ERF minimization for $N_{rep} = 1000$ from NNPDF3.0 NLO.

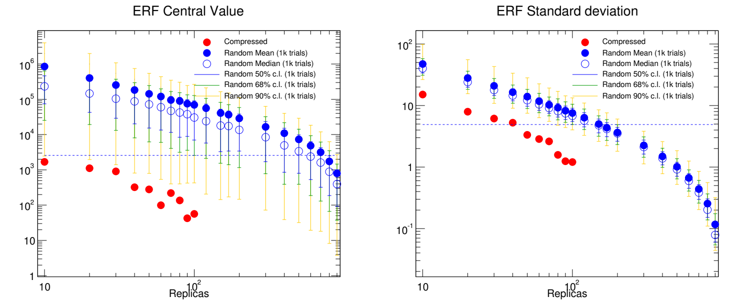


compressor v1.0.0	
GA Parameters	
$N_{\mathrm{gen}}^{\mathrm{max}}$	15000
N_{mut}	5
N_{x}	70
x_{min}	10^{-5}
x_{max}	0.9
n_f	7
Q_0	user-defined
$N_{\mathrm{x}}^{\mathrm{corr}}$	5
N_{rand}	1000

- The algorithm reaches the stability plateau after 2k iterations.
- Large prior of MC replicas $\Rightarrow$ increases possible combinations.

The compression strategy

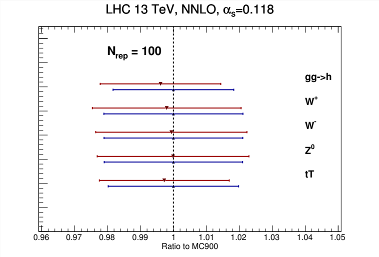
Moment estimators for the **compression** and **random selections**.
- horizontal lines ⇒ lower 68% c.l. for random selection with $N_{rep} = 100$

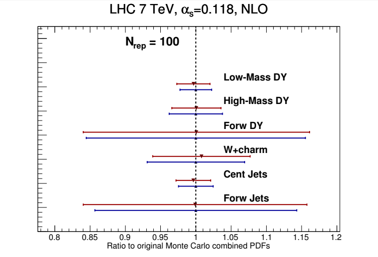


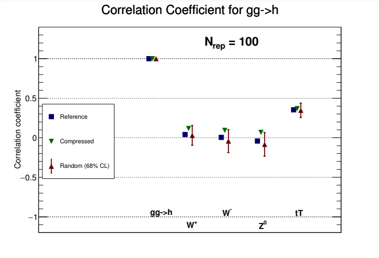
- Substantial **improvements** as compared to random selections.
- Compression is able to successfully reproduce **higher moments** and **correlations**.
- In this test case $N_{rep} = 50$ are equivalent to MC fits with 100 replicas.

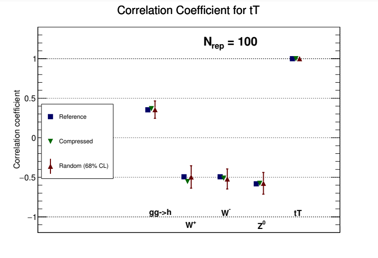
LHC Phenomenology

CMC-PDFs also validated for LHC inclusive cross-sections and differential distributions, including correlations.



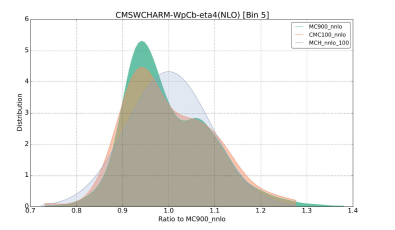


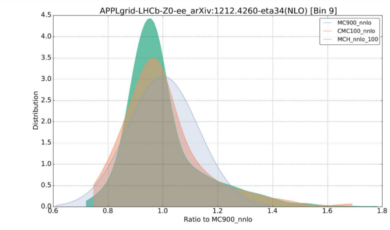




Non-Gaussian features in LHC cross-sections

Non-Gaussian features are also clearly observed at the level of LHC processes, e.g. most forward bin of the CMS W+charm differential cross-section measurement and DY measurement from LHCb.





Hessian reduction fails by construction when reproducing such features.

However, in regions where the **Gaussian approximation** is reasonable, one should use a Hessian representation.

General strategy

1. Given a Monte Carlo prior set of PDFs

$$\{f^{(k)}_{\alpha,\mathrm{mc}}\}_{k=1,\ldots,N_{\mathrm{rep}}}\,,\quad \alpha=\{g,u,d,s,\ldots\}\,,$$

2. Fix the central value to be the same as the prior:

$$f^{(0)}_{\alpha,\mathrm{hessian}}=f^{(0)}_{\alpha,\mathrm{mc}}$$

3. We define the matrix for the deviations wrt central value:

$$X_{lk}(Q)\equiv f^{(k)}_{\alpha,\mathrm{mc}}(x_i,Q)-f^{(0)}_{\alpha}(x_i,Q),\quad l\equiv N_x(\alpha-1)+i$$

4. The covariance matrix is given in terms of X:

$$\mathrm{cov}^{\mathrm{pdf}}_{\mathrm{ij},\alpha\beta}(Q)\equiv\frac{1}{N_{\mathrm{rep}}-1}XX^t$$

General strategy

SVD

A diagonal representation of the covariance matrix in terms of replicas is found by SVD of the matrix X:

$$X = USV^t,$$

V is an orthogonal $N_{\mathrm{rep}} \times N_{\mathrm{rep}}$ matrix of coefficients, and

$$XV,$$

provides a representation of the multigaussian covariance matrix in terms of the original replicas.

PCA Reduction

Many eigenvectors lead to a very small contribution to the covariance matrix $\Rightarrow$ we can select a smaller set of N_{eig}, with largest eigenvalues, which still provides a good approximation to the covariance matrix.

General strategy

The PCA optimization retains the principal components, i.e. the largest singular values.

- U, S are replaced by their submatrices u, s respectively.
  - $\dim u = N_x N_f \times N_{\rm eig}$ and $\dim s = N_{\rm eig} \times N_{\rm rep}$
- Only the $N_{\rm rep} \times N_{\rm eig}$ orthogonal upper left submatrix of V contributes
- This is the principal submatrix P of V:

$$P_{ki} = V_{ki}, \quad k = 1, \ldots N_{\rm rep}; i = 1, \ldots, N_{\rm eig}$$

General strategy

The PCA optimization retains the principal components, i.e. the largest singular values.

- U, S are replaced by their submatrices u, s respectively.
  - $\dim u = N_x N_f \times N_{\rm eig}$ and $\dim s = N_{\rm eig} \times N_{\rm rep}$
- Only the $N_{\rm rep} \times N_{\rm eig}$ orthogonal upper left submatrix of V contributes
- This is the principal submatrix P of V:

$$P_{ki} = V_{ki}, \quad k = 1, \ldots N_{\rm rep}; i = 1, \ldots, N_{\rm eig}$$

Thus we write the Hessian eigenvectors as a linear combination of replicas:

$$\begin{aligned} f^{(i)}_{\alpha,\rm hessian}(x_j, Q) &= f^{(0)}_{\alpha}(x_j, Q) + X_{lk} P_{ki}, \qquad l \equiv N_x(\alpha - 1) + j \\ &= f^{(0)}_{\alpha}(x_j, Q) + \sum_{k=1}^{N_{\rm rep}} a^{(i)}_k \left(f^{(k)}_{\alpha,mc}(x_j, Q) - f^{(0)}_{\alpha}(x_j, Q) \right) \end{aligned}$$

Note the $a^{(i)}_k$ independence in (x, Q). It takes care of evolution automatically.

Gaussianity of the PDF4LHC15 combinations

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 $Q(\sigma)$.

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

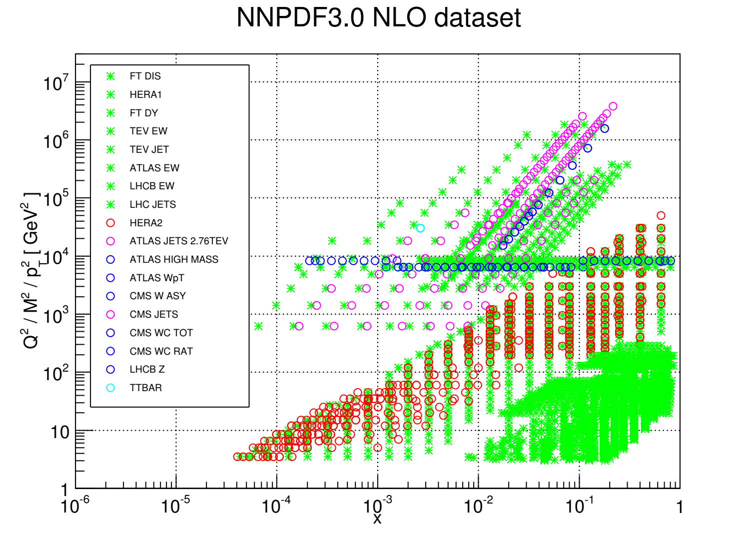
We compare the KDE of the prior with

- A Gaussian given by $\mu = \langle \sigma_i \rangle_i$, $\sigma = \frac{1}{N-1}\sqrt{\sum(\sigma_i - \mu)^2}$.
- The MC2H Gaussian.
- The CMC KDE.

Here we have used the SMPDF dataset.

Robustness of the PDF4LHC15 combinations

We have computed predictions with **PDF4LHC15_prior** and the three reduced sets, for all data in the NNPDF3.0 dataset.



SMPDF backup - algorithm strategy

Following the MC2H PCA methodology we can find a subspace with a smaller number of parameters which optimizes the agreement for some quantities.

$$\tilde{X} = XP \in \mathbb{R}^{N_x N_{\rm pdf}} \times \mathbb{R}^{N_{\rm eig} \ll N_{\rm rep}}$$

We can greatly improve the reduction by targeting specific processes:

$$\{\sigma_i\}, \quad i = 1, \ldots, N_\sigma$$

$$s_{\sigma_i} = \left(\frac{1}{N_{\rm rep} - 1} \sum_{k=1}^{N_{\rm rep}} \left(\sigma_i^{(k)} - \sigma_i^{(0)} \right)^2 \right)^{\frac{1}{2}}$$

The worst-case accuracy target can be tuned by user:

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

This is implemented in an interactive procedure.

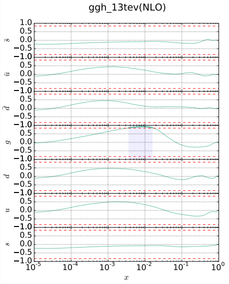
SMPDF backup - selection algorithm

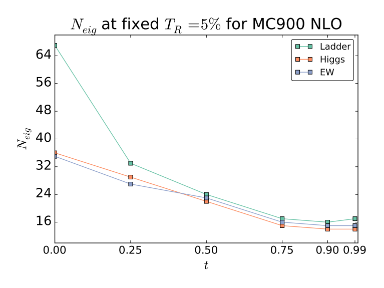
For each iteration, select points in (x, α, Q) correlated with variations in σ

$$\rho(x_i, Q, \alpha, \sigma) = \frac{N_{\rm rep}}{N_{\rm rep}-1} \frac{\langle X(Q_\alpha)_{lk} \cdot (\sigma^{(k)} - \sigma^{(0)})\rangle_{\rm rep} - \langle X(Q_\sigma)_{lk}\rangle_{\rm rep} \cdot \langle \sigma^{(k)} - \sigma^{(0)}\rangle_{\rm rep}}{s_\alpha^{\rm PDF} \cdot s_\sigma}$$

$$\Xi = \{(X_i, \alpha) : \rho(X_i, Q_\alpha, \alpha, \sigma) \geq t \cdot \rho_{\max}\}, \quad X \to X_\Xi(Q_\sigma)$$

The correlation threshold t is the only free parameter of the algorithm $\Rightarrow$ $t = 0.9$ optimal choice.





SMPDF backup - orthogonal projection algorithm

This approach allows to efficiently generalize to processes with similar PDF dependence, making the algorithm stable.

We compute the SVD of X_Ξ and select **one** eigenvector:

$$X_\Xi(Q_\alpha) = USV^t$$

$$(P \cdot R) = V \in \mathbb{R}^{N_{\rm rep}} \times \left(\mathbb{R}^1 \mathbb{R}^{N_{\rm rep}-1}\right)$$

We project out the selected eigenvector for the next iteration

$$X \to XR$$

We iterate (select more eigenvectors) until we meet the tolerance criteria for the current observable, and move to the next observable, until we reproduce all.

SMPDF backup - APPLgrids

Input cross-sections for SM-PDFs for Higgs physics

process	distribution	grid name	N_{bins}	range	kin. cuts
$gg \to h$	incl xsec	ggh_13tev	1	-	-
	$d\sigma/dp_t^h$	ggh_pt_13tev	10	[0,200] GeV	-
	$d\sigma/dy^h$	ggh_y_13tev	10	[-2.5,2.5]	-
VBF hjj	incl xsec	vbfh_13tev	1	-	-
	$d\sigma/dp_t^h$	vbfh_pt_13tev	5	[0,200] GeV	-
	$d\sigma/dy^h$	vbfh_y_13tev	5	[-2.5,2.5]	-
hW	incl xsec	hw_13tev	1	-	$p_T(l) \geq 10$ GeV, $\|\eta^l\| \leq 2.5$
	$d\sigma/dp_t^h$	hw_pt_13tev	10	[0,200] GeV	$p_T(l) \geq 10$ GeV, $\|\eta^l\| \leq 2.5$
	$d\sigma/dy^h$	hw_y_13tev	10	[-2.5,2.5]	$p_T(l) \geq 10$ GeV, $\|\eta^l\| \leq 2.5$
hZ	incl xsec	hz_13tev	1	-	$p_T(l) \geq 10$ GeV, $\|\eta^l\| \leq 2.5$
	$d\sigma/dp_t^h$	hz_pt_13tev	10	[0,200] GeV	$p_T(l) \geq 10$ GeV, $\|\eta^l\| \leq 2.5$
	$d\sigma/dy^h$	hz_y_13tev	10	[-2.5,2.5]	$p_T(l) \geq 10$ GeV, $\|\eta^l\| \leq 2.5$
$ht\bar{t}$	incl xsec	httbar_13tev	1	-	-
	$d\sigma/dp_t^h$	httbar_pt_13tev	10	[0,200] GeV	-
	$d\sigma/dy^h$	httbar_y_13tev	10	[-2.5,2.5]	-

SMPDF backup - APPLgrids

Input cross-sections for SM-PDFs for $t\bar{t}$ physics					
process	distribution	grid name	N_{bins}	range	kin. cuts
$t\bar{t}$	incl xsec	`ttbar_13tev`	1	-	-
	$d\sigma/dp_t^{\bar{t}}$	`ttbar_tbarpt_13tev`	10	[40,400] GeV	-
	$d\sigma/dy^{\bar{t}}$	`ttbar_tbary_13tev`	10	[-2.5,2.5]	-
	$d\sigma/dp_t^{t}$	`ttbar_tpt_13tev`	10	[40,400] GeV	-
	$d\sigma/dy^{t}$	`ttbar_ty_13tev`	10	[-2.5,2.5]	-
	$d\sigma/dm^{t\bar{t}}$	`ttbar_ttbarinvmass_13tev`	10	[300,1000]	-
	$d\sigma/dp_t^{t\bar{t}}$	`ttbar_ttbarpt_13tev`	10	[20,200]	-
	$d\sigma/dy^{t\bar{t}}$	`ttbar_ttbary_13tev`	12	[-3,3]	-

SMPDF backup - APPLgrids

Input cross-sections for SM-PDFs for electroweak boson production physics					
process	distribution	grid name	N_{bins}	range	kin. cuts
Z	incl xsec	z_13tev	1	-	$p_T(l) \geq 10$ GeV, $\lvert\eta^l\rvert \leq 2.5$
	$d\sigma / dp_t^{l^-}$	z_lmpt_13tev	10	[0,200] GeV	$p_T(l) \geq 10$ GeV, $\lvert\eta^l\rvert \leq 2.5$
	$d\sigma / dy^{l^-}$	z_lmy_13tev	10	[-2.5,2.5]	$p_T(l) \geq 10$ GeV, $\lvert\eta^l\rvert \leq 2.5$
	$d\sigma / dp_t^{l^+}$	z_lppt_13tev	10	[0,200] GeV	$p_T(l) \geq 10$ GeV, $\lvert\eta^l\rvert \leq 2.5$
	$d\sigma / dy^{l^-}$	z_lpy_13tev	10	[-2.5,2.5]	$p_T(l) \geq 10$ GeV, $\lvert\eta^l\rvert \leq 2.5$
	$d\sigma / dp_t^Z$	z_zpt_13tev	10	[0,200] GeV	$p_T(l) \geq 10$ GeV, $\lvert\eta^l\rvert \leq 2.5$
	$d\sigma / dy^Z$	z_zy_13tev	5	[-4,4]	$p_T(l) \geq 10$ GeV, $\lvert\eta^l\rvert \leq 2.5$
	$d\sigma / dm^{ll}$	z_lplminvmass_13tev	10	[50,130] GeV	$p_T(l) \geq 10$ GeV, $\lvert\eta^l\rvert \leq 2.5$
	$d\sigma / dp_t^{ll}$	z_lplmpt_13tev	10	[0,200] GeV	$p_T(l) \geq 10$ GeV, $\lvert\eta^l\rvert \leq 2.5$
W	incl xsec	w_13tev	1	-	$p_T(l) \geq 10$ GeV, $\lvert\eta^l\rvert \leq 2.5$
	$d\sigma / d\phi$	w_cphi_13tev	10	[-1,1]	$p_T(l) \geq 10$ GeV, $\lvert\eta^l\rvert \leq 2.5$
	$d\sigma / dE_t^{\mathrm{miss}}$	w_etmiss_13tev	10	[0,200] GeV	$p_T(l) \geq 10$ GeV, $\lvert\eta^l\rvert \leq 2.5$
	$d\sigma / dp_t^l$	w_lpt_13tev	10	[0,200] GeV	$p_T(l) \geq 10$ GeV, $\lvert\eta^l\rvert \leq 2.5$
	$d\sigma / dy^l$	w_ly_13tev	10	[-2.5,2.5]	$p_T(l) \geq 10$ GeV, $\lvert\eta^l\rvert \leq 2.5$
	$d\sigma / dm_t$	w_mt_13tev	10	[0,200] GeV	$p_T(l) \geq 10$ GeV, $\lvert\eta^l\rvert \leq 2.5$
	$d\sigma / dp_t^W$	w_wpt_13tev	10	[0,200] GeV	$p_T(l) \geq 10$ GeV, $\lvert\eta^l\rvert \leq 2.5$
	$d\sigma / dy^W$	w_wy_13tev	10	[-4,4]	$p_T(l) \geq 10$ GeV, $\lvert\eta^l\rvert \leq 2.5$

SMPDF results