SPECIALIZED MINIMAL PDFS

PDF4LHC Meeting

Zahari Kassabov

In collaboration with S. Carrazza, S. Forte, J.Rojo

October 27th 2015

University of Turin, University of Milan

SMPDF

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

OVERVIEW

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

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

OVERVIEW

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

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

- Suitable for use in experimental analysis:

OVERVIEW

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

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.

OVERVIEW

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.

OVERVIEW

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

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

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

- Motivation:

APPROACH

- Motivation:
    - General purpose sets require 30 (lower accuracy) to 100 (higher accuracy) error sets (to reproduce the combined PDF4LHC Monte Carlo prior, *MC900*).

- Motivation:
    - General purpose sets require 30 (lower accuracy) to 100 (higher accuracy) error sets.
    - A better balance can be achieved if we only need to reproduce some processes.

APPROACH

- Motivation:
  - General purpose sets require 30 (lower accuracy) to 100 (higher accuracy) error sets.
  - A better balance can be achieved if we only need to reproduce some processes.
- Other methods:

  Reduced META ensemble P. Nadolsky at PDF4LHC November 2014

  (based on Dataset Diagonalization method, J Pumplin arxiv:0904.2425).

- Motivation:
  - General purpose sets require 30 (lower accuracy) to 100 (higher accuracy) error sets.
  - A better balance can be achieved if we only need to reproduce some processes.
- Other methods:

  Reduced META ensemble P. Nadolsky at PDF4LHC November 2014.
- Our approach is based on:

- Motivation:
  - General purpose sets require 30 (lower accuracy) to 100 (higher accuracy) error sets.
  - A better balance can be achieved if we only need to reproduce some processes.
- Other methods:

  Reduced META ensemble P. Nadolsky at PDF4LHC November 2014.
- Our approach is based on:
  - Correlations on x space.

APPROACH

- Motivation:
  - General purpose sets require 30 (lower accuracy) to 100 (higher accuracy) error sets.
  - A better balance can be achieved if we only need to reproduce some processes.
- Other methods:

  Reduced META ensemble P. Nadolsky at PDF4LHC November 2014.
- Our approach is based on:
  - Correlations on x space.
    - Provides stability, e.g. can reproduce cross sections for the same process in a wide kinematical range.

APPROACH

- Motivation:
  - General purpose sets require 30 (lower accuracy) to 100 (higher accuracy) error sets.
  - A better balance can be achieved if we only need to reproduce some processes.
- Other methods:

  Reduced META ensemble P. Nadolsky at PDF4LHC November 2014.
- Our approach is based on:
  - Correlations on x space.
  - Iterative orthogonal projections

- Motivation:
    - General purpose sets require 30 (lower accuracy) to 100 (higher accuracy) error sets.
    - A better balance can be achieved if we only need to reproduce some processes.
- Other methods:

  Reduced META ensemble P. Nadolsky at PDF4LHC November 2014.
- Our approach is based on:
    - Correlations on x space.
    - Iterative orthogonal projections
        - Allows to reproduce smaller uncertainty contributions required to reach the tolerance.

APPROACH

- Motivation:
  - General purpose sets require 30 (lower accuracy) to 100 (higher accuracy) error sets.
  - A better balance can be achieved if we only need to reproduce some processes.
- Other methods:

  Reduced META ensemble P. Nadolsky at PDF4LHC November 2014.
- Our approach is based on:
  - Correlations on x space.
  - Iterative orthogonal projections
- We can combine:

APPROACH

- Motivation:
  - General purpose sets require 30 (lower accuracy) to 100 (higher accuracy) error sets.
  - A better balance can be achieved if we only need to reproduce some processes.
- Other methods:

  Reduced META ensemble P. Nadolsky at PDF4LHC November 2014.
- Our approach is based on:
  - Correlations on x space.
  - Iterative orthogonal projections
- We can combine:
  - Different sets of observables, by generating a common SM-PDF.

- Motivation:
  - General purpose sets require 30 (lower accuracy) to 100 (higher accuracy) error sets.
  - A better balance can be achieved if we only need to reproduce some processes.
- Other methods:

  Reduced META ensemble P. Nadolsky at PDF4LHC November 2014.
- Our approach is based on:
  - Correlations on *x* space.
  - Iterative orthogonal projections
- We can combine:
  - Results from independent SM-PDFs a posteriori.

APPROACH

- Motivation:
  - General purpose sets require 30 (lower accuracy) to 100 (higher accuracy) error sets.
  - A better balance can be achieved if we only need to reproduce some processes.
- Other methods:

  Reduced META ensemble P. Nadolsky at PDF4LHC November 2014.
- Our approach is based on:
  - Correlations on x space.
  - Iterative orthogonal projections
- We can combine:
- Combined SM-PDFs including HERAPDF?

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 (see backup slides 22-24).

RESULTS

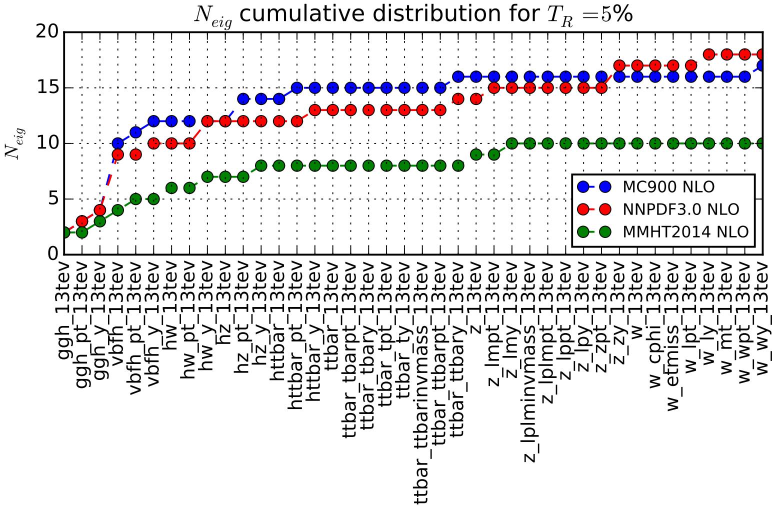
Process	MC900		NNPDF3.0		MMHT14	
	$T_R = 5\%$	$T_R = 10\%$	$T_R = 5\%$	$T_R = 10\%$	$T_R = 5\%$	$T_R = 10\%$
h	15	11	13	8	8	7
$t\bar{t}$	4	4	5	4	3	3
W, Z	14	11	13	8	10	9
Ladder	17	14	18	11	10	10

- 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

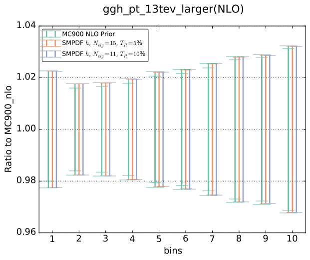
Process	MC900		NNPDF3.0		MMHT14	
	$T_R = 5\%$	$T_R = 10\%$	$T_R = 5\%$	$T_R = 10\%$	$T_R = 5\%$	$T_R = 10\%$
$gg \to h$	4	5	4	4	3	3
VBF hjj	7	5	10	5	4	3
hW	6	5	6	4	6	3
hZ	11	7	6	4	8	5
$ht\bar{t}$	3	2	4	4	3	2
Total h	15	11	13	8	8	7

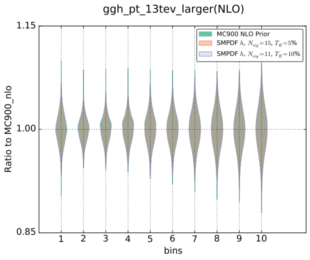
Multiple processes can be efficiently stacked together:

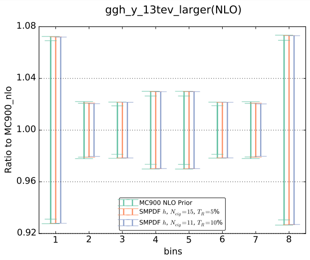


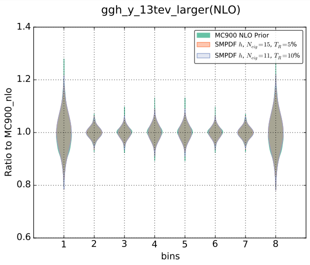
Kinematical ranges that double those used as input:

($p_T^h \in (0, 400)$ GeV, $y^h \in (-5, 5)$).

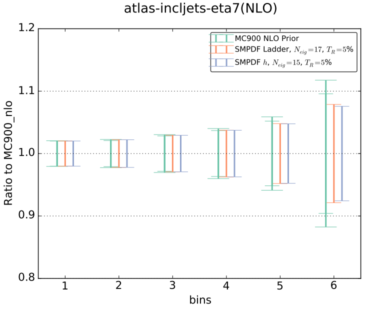




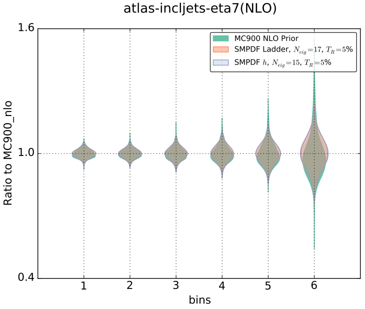




“BREAKDOWN”



atlas-incljets-eta7(NLO)
MC900 NLO Prior
SMPDF Ladder, N_{eig}=17, T_R=5%
SMPDF h, N_{eig}=15, T_R=5%
Ratio to MC900_nlo
bins
1.2
1.1
1.0
0.9
0.8
1
2
3
4
5
6



atlas-incljets-eta7(NLO)
MC900 NLO Prior
SMPDF Ladder, N_{eig}=17, T_R=5%
SMPDF h, N_{eig}=15, T_R=5%
Ratio to MC900_nlo
bins
1.6
1.0
0.4
1
2
3
4
5
6

- The code is public and can be used to generate custom `SM-PDFs` from:
  - Observables in `APPLgrid` or text format (e.g. NNLO codes).
  - `LHAPDF6` Prior PDF (MC or `symhessian`).
- The aforementioned `SM-PDFs` will be made public.
  - Including the `APPLgrids` used to generate them.
- Custom `SM-PDFs` can be generated upon request.

HESSIAN REPRESENTATION: GENERAL STRATEGY

1. Start with prior PDF (Monte Carlo or Symmetric Hessian)

$$f^{(k)}_{\alpha(Prior)}(x, Q),\ k = 1 \ldots N_{\text{rep}}$$

1. Start with prior PDF (Monte Carlo or Symmetric Hessian)

$$f^{(k)}_{\alpha(Prior)}(x,Q),\ k=1\ldots N_{\text{rep}}$$

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

$$f^{(0)}_{\alpha(\text{Hessian})}(x,Q) \equiv f^{(0)}_{\alpha(Prior)}(x,Q)$$

HESSIAN REPRESENTATION: GENERAL STRATEGY

1. Start with prior PDF (Monte Carlo or Symmetric Hessian)

$$f^{(k)}_{\alpha(Prior)}(x,Q),\ k=1\dots N_{\rm rep}$$

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

$$f^{(0)}_{\alpha(\text{Hessian})}(x,Q)\equiv f^{(0)}_{\alpha(Prior)}(x,Q)$$

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

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

$$\mathrm{cov}(Q)=\frac{1}{N_{\rm rep}-1}XX^t$$

$$\boxed{f^{(k)}_{\alpha(\mathrm{Hesian})}(x_i,Q)=f^{(0)}_{\alpha}(x_i,Q)+X_{lk}(Q)\,,\quad k=1,\dots,N_{\rm eig}\,.}$$

1. Start with prior PDF (Monte Carlo or Symmetric Hessian)

$$f^{(k)}_{\alpha(Prior)}(x, Q),\ k = 1 \ldots N_{\text{rep}}$$

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

$$f^{(0)}_{\alpha(\text{Hessian})}(x, Q) \equiv f^{(0)}_{\alpha(Prior)}(x, Q)$$

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

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

$$\mathrm{cov}(Q) = \frac{1}{N_{\mathrm{rep}} - 1} X X^t$$

$$\boxed{f^{(k)}_{\alpha(\mathrm{Hesian})}(x_i, Q) = f^{(0)}_{\alpha}(x_i, Q) + X_{lk}(Q)\,, \quad k = 1, \ldots, N_{\mathrm{eig}}\,.}$$

l can be arbitrarily fine grained. Completely unbiased!

1. Start with prior PDF (Monte Carlo or Symmetric Hessian)

$$f^{(k)}_{\alpha(Prior)}(x,Q),\ k = 1 \dots N_{\text{rep}}$$

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

$$f^{(0)}_{\alpha(\text{Hessian})}(x,Q) \equiv f^{(0)}_{\alpha(Prior)}(x,Q)$$

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

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

$$\mathrm{cov}(Q) = \frac{1}{N_{\mathrm{rep}} - 1} X X^t$$

$$\boxed{f^{(k)}_{\alpha(\mathrm{Hesian})}(x_i,Q) = f^{(0)}_{\alpha}(x_i,Q) + X_{lk}(Q)\,, \quad k = 1,\dots,N_{\mathrm{eig}}\,.}$$

l can be arbitrarily fine grained. Completely unbiased!

HESSIAN REPRESENTATION: GENERAL STRATEGY

1. Start with prior PDF (Monte Carlo or Symmetric Hessian)

$$f^{(k)}_{\alpha(Prior)}(x,Q),\ k=1\dots N_{\mathrm{rep}}$$

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

$$f^{(0)}_{\alpha(\mathrm{Hessian})}(x,Q)\equiv f^{(0)}_{\alpha(Prior)}(x,Q)$$

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

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

$$\mathrm{cov}(Q)=\frac{1}{N_{\mathrm{rep}}-1}XX^t$$

$$\boxed{f^{(k)}_{\alpha(\mathrm{Hesian})}(x_i,Q)=f^{(0)}_{\alpha}(x_i,Q)+X_{lk}(Q)\,,\quad k=1,\dots,N_{\mathrm{eig}}(=N_{\mathrm{rep}})\,.}$$

l can be arbitrarily fine grained. Completely unbiased!

mc2hessian: GLOBAL OPTIMIZATION

- In mc2hessian PCA (arxiv:1505.06736, appendix) we optimize for the absolute value of the covariance matrix, at some scale Q_0.
    - This is used for the PDFLHC15_100 sets.

- In `mc2hessian PCA` (arxiv:1505.06736, appendix) we optimize for the absolute value of the covariance matrix, at some scale Q_0.
    - This is used for the PDFLHC15_100 sets.
- We compute the *Singular Value Decomposition* and keep the N_{eig} largest singular values:

$$X(Q_0) = USV^t$$

`mc2hessian`: GLOBAL OPTIMIZATION

- In `mc2hessian PCA` (arxiv:1505.06736, appendix) we optimize for the absolute value of the covariance matrix, at some scale Q_0.
  - This is used for the PDFLHC15_100 sets.
- We compute the *Singular Value Decomposition* and keep the $N_{\rm eig}$ largest singular values:

$$X(Q_0) = USV^t$$

$$V \in \mathbb{R}^{N_{\rm rep}} \times \mathbb{R}^{N_{\rm rep}}$$

mc2hessian: GLOBAL OPTIMIZATION

- In `mc2hessian PCA` (arxiv:1505.06736, appendix) we optimize for the absolute value of the covariance matrix, at some scale Q_0.
  - This is used for the PDFLHC15_100 sets.
- We compute the *Singular Value Decomposition* and keep the $N_{\rm eig}$ largest singular values:

$$X(Q_0) = USV^t$$

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

- In `mc2hessian PCA` (arxiv:1505.06736, appendix) we optimize for the absolute value of the covariance matrix, at some scale Q_0.
  - This is used for the PDFLHC15_100 sets.
- We compute the *Singular Value Decomposition* and keep the N_{eig} largest singular values:

$$X(Q_0) = USV^t$$

$$\widetilde{X} = XP$$

mc2hessian: GLOBAL OPTIMIZATION

- In `mc2hessian PCA` (arxiv:1505.06736, appendix) we optimize for the absolute value of the covariance matrix, at some scale Q_0.
  - This is used for the PDFLHC15_100 sets.
- We compute the *Singular Value Decomposition* and keep the N_{eig} largest singular values:

$$X(Q_0) = USV^t$$

$$\widetilde{X} = XP$$

- `mc2hessian` needs ~ 100 eigenvectors to reproduce the covariance matrix of the PDF4LHC prior and most phenomenology at percent level.

We have:

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

which gives N_{rep} Hessian parameters, from $k = 1$, to $k = N_{\mathrm{rep}}$.

MORE COMPRESSED HESSIAN

We have:

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

which gives N_{rep} Hessian parameters, from $k = 1$, to $k = N_{\mathrm{rep}}$.

- Any rotation of the space of linear combinations of replicas gives the same results for linear error propagation of PDF dependent quantities.

We have:

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

which gives N_{rep} Hessian parameters, from $k = 1$, to $k = N_{\mathrm{rep}}$.

- Any rotation of the space of linear combinations of replicas gives the same results for linear error propagation of PDF dependent quantities.
- We can find a subspace with a smaller number of parameters, which optimizes the agreement for some quantities:

MORE COMPRESSED HESSIAN

We have:

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

which gives N_{rep} Hessian parameters, from $k = 1$, to $k = N_{\mathrm{rep}}$.

- Any rotation of the space of linear combinations of replicas gives the same results for linear error propagation of PDF dependent quantities.
- We can find a subspace with a smaller number of parameters, which optimizes the agreement for some quantities:

$$X \in \mathbb{R}^{N_x N_{\mathrm{pdf}}} \times \mathbb{R}^{N_{\mathrm{rep}}}$$

MORE COMPRESSED HESSIAN

We have:

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

which gives N_{rep} Hessian parameters, from $k = 1$, to $k = N_{\mathrm{rep}}$.

- Any rotation of the space of linear combinations of replicas gives the same results for linear error propagation of PDF dependent quantities.
- We can find a subspace with a smaller number of parameters, which optimizes the agreement for some quantities:

$$X \in \mathbb{R}^{N_x N_{\mathrm{pdf}}} \times (\mathcal{P} \oplus \mathcal{R})$$

MORE COMPRESSED HESSIAN

We have:

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

which gives N_{rep} Hessian parameters, from $k = 1$, to $k = N_{\rm rep}$.

- Any rotation of the space of linear combinations of replicas gives the same results for linear error propagation of PDF dependent quantities.
- We can find a subspace with a smaller number of parameters, which optimizes the agreement for some quantities:

$$XP \in \mathbb{R}^{N_x N_{\rm pdf}} \times (\mathcal{P} \cancel{\oplus \mathcal{R}})$$

We have:

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

which gives N_{rep} Hessian parameters, from $k = 1$, to $k = N_{\mathrm{rep}}$.

- Any rotation of the space of linear combinations of replicas gives the same results for linear error propagation of PDF dependent quantities.
- We can find a subspace with a smaller number of parameters, which optimizes the agreement for some quantities:

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

We have:

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

which gives N_{rep} Hessian parameters, from $k = 1$, to $k = N_{\mathrm{rep}}$.

- Any rotation of the space of linear combinations of replicas gives the same results for linear error propagation of PDF dependent quantities.
- We can find a subspace with a smaller number of parameters, which optimizes the agreement for some quantities:

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

- The reduced PDF representation is:

$$\boxed{\widetilde{f}_{\alpha}^{(k)}(x_i, Q) = f_{\alpha}^{(0)}(x_i, Q) + \widetilde{X}_{lk}(Q)\,, \quad k = 1, \ldots, N_{\mathrm{eig}}\,.}$$

- We can greatly improve the reduction by targeting specific processes:

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

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

- We can greatly improve the reduction by targeting specific processes:

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

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

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

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

- We can greatly improve the reduction by targeting specific processes:

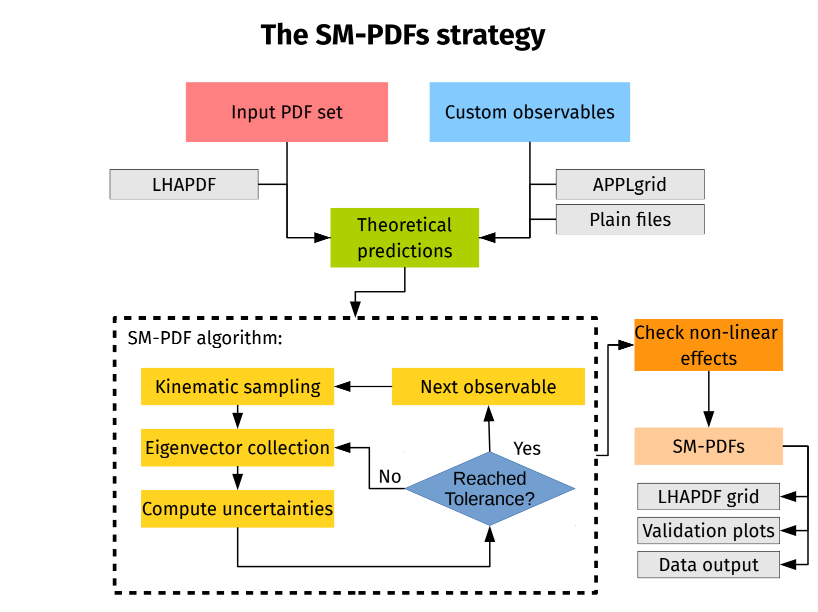
$$\{\sigma_i\}\,, \quad i = 1, \dots N_\sigma$$

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

- 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 iterative procedure.


The SM-PDFs strategy
Input PDF set
Custom observables
LHAPDF
APPLgrid
Plain files
Theoretical predictions
SM-PDF algorithm:
Kinematic sampling
Next observable
Eigenvector collection
Compute uncertainties
Yes
No
Reached Tolerance?
Check non-linear effects
SM-PDFs
LHAPDF grid
Validation plots
Data output

CORRELATION BASED SELECTION

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

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

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

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

CORRELATION BASED SELECTION

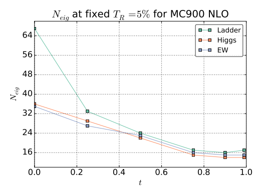
- For each iteration, select points in (x, α, Q) correlated with variations in σ:

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

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

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

- The correlation threshold t is the only free parameter of the algorithm: Set to $t = 0.9$ on phenomenological grounds.



CORRELATION BASED SELECTION

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

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

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

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

- The correlation threshold t is the only free parameter of the algorithm: Set to $t = 0.9$ on phenomenological grounds.
- The correlation-based approach allows to efficiently generalize to processes with similar PDF dependence, making the algorithm stable.
    - Same SMPDF can be used with different cuts.

- We compute the SVD of X_{Ξ} and select one eigenvector:

$$X_{\Xi}(Q_\sigma) = USV^t$$

$$\begin{pmatrix} P & R \end{pmatrix} = V \in \mathbb{R}^{N_{\mathrm{rep}}} \times \begin{pmatrix} \mathbb{R}^1 & \mathbb{R}^{N_{\mathrm{rep}}-1} \end{pmatrix}$$

- We compute the SVD of X_{Ξ} and select one eigenvector:

$$X_{\Xi}(Q_\sigma) = USV^t$$

$$\begin{pmatrix} P & R \end{pmatrix} = V \in \mathbb{R}^{N_{\mathrm{rep}}} \times \begin{pmatrix} \mathbb{R}^1 & \mathbb{R}^{N_{\mathrm{rep}}-1} \end{pmatrix}$$

- We project out the selected eigenvector for the next iteration:

$$X \longrightarrow XR$$

- We compute the SVD of X_{Ξ} and select one eigenvector:

$$X_{\Xi}(Q_\sigma) = USV^t$$

$$\begin{pmatrix} P & R \end{pmatrix} = V \in \mathbb{R}^{N_{\mathrm{rep}}} \times \begin{pmatrix} \mathbb{R}^1 & \mathbb{R}^{N_{\mathrm{rep}}-1} \end{pmatrix}$$

- We project out the selected eigenvector for the next iteration:

$$X \longrightarrow XR$$

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

- We compute the SVD of X_{Ξ} and select one eigenvector:

$$X_{\Xi}(Q_\sigma) = USV^t$$

$$\begin{pmatrix} P & R \end{pmatrix} = V \in \mathbb{R}^{N_{\mathrm{rep}}} \times \begin{pmatrix} \mathbb{R}^1 & \mathbb{R}^{N_{\mathrm{rep}}-1} \end{pmatrix}$$

- We project out the selected eigenvector for the next iteration:

$$X \longrightarrow XR$$

- We iterate (select more eigenvectors) until we met the tolerance criteria for the current observable, and move to the next observable, until we reproduce all.
- See upcoming publication or code for more details.

- Results computed with different SMPDFs can be easily combined by expressing them in terms of the prior (within the tolerance):

$$d_k(\sigma)/\sqrt{N_{\mathrm{rep}}-1} = \sum_{j=1}^{N_{\mathrm{eig}}} P^t_{kj}\widetilde{d}_j(\sigma)\,, \qquad k=1,\ldots,N_{\mathrm{rep}}$$

$d_k(\sigma) = \sigma_k - \sigma_0,$ for σ_k expressed in terms of the N_{rep} prior replicas.

$d_j(\sigma) = \sigma_j - \sigma_0,$ for σ_j computed in terms of the N_{eig} SMPDF eigenvectors.

- Direct and inverse transformation matrices provided in the `SM-PDF` code.
  - With appropriate normalization constants and hash-based parameter names.

OTHER FEATURES OF THE CODE

- Python interface for `APPlgrid` and `LHAPDF`.
- Phenomenology comparisons for PDFs.
    - PDF4LHC Recommendation benchmarking plots.
    - Yellow report PDF correlations.
- Correlation plots
    - PDF with observable
    - Observable PDF uncertainties
- `mc2hessian` algorithm.
- Observable values as a function of α_s or N_f.
- Convolution result as a rich `HTML` and `CSV` table.

CODE INPUT

Execute with: `smpdf higgs.yaml –use-db`

```
#higgs.yaml
#Global parameters that are used until overwritten by parameters
#inside the actiongroups
observables: #Provide the Paths to the APPLGrids, and specify that
             #they are calculated at NLO
   # Higgs
   - {name: 'data/higgs/ggh_13tev.root', order: NLO}
   - {name: 'data/higgs/ggH_y_13tev.root', order: NLO}
   - {name: 'data/higgs/hw_13tev.root', order: NLO}
   - {name: 'data/higgs/hz_13tev.root', order: NLO}
   - {name: 'data/higgs/ggH_pt_13tev.root', order: NLO}
   - {name: 'data/higgs/httbar_13tev.root', order: NLO}
pdfsets:
   - MC900_nnlo #LHAPDF set to be used as prior
actions:
   - smpdf #Generate the SMPDFs from the prior and the observables
   - installgrids #Install the generated sets in the LHAPDF path

#The specification of the actions to actually be performed
#using the avove as default
actiongroups:
   - prefix: H05_ #Begin all exported filenames with this prefix
     smpdf_tolerance: 0.05 #Set T to 5% and execute the default
                           #actions above

   - prefix: H10_
     smpdf_tolerance: 0.10 #Set T to 10% and execute the default
                           #actions.

   - prefix: compall
     pdfsets: #Change the PDFsets for this actiongroup
        - MCH_nnlo_100
        - H05_smpdf* #Wildcard expansion is supported.
        - H10_smpdf*
        - MC900_nnlo
     actions: #Perform plots and save the data of the convolution.
        - violinplots
        - obscorrplots
        - ciplots
        - savedata
     base_pdf: MC900_nnlo #Plot values relative to this PDF.
```

Questions

BACKUP: SMPDF-HIGGS PROCESS DETAILS

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, $\lvert\eta^l\rvert \leq 2.5$
	$d\sigma/dp_t^h$	`hw_pt_13tev`	10	[0,200] GeV	$p_T(l) \geq$ 10 GeV, $\lvert\eta^l\rvert \leq 2.5$
	$d\sigma/dy^h$	`hw_y_13tev`	10	[-2.5,2.5]	$p_T(l) \geq$ 10 GeV, $\lvert\eta^l\rvert \leq 2.5$
hZ	incl xsec	`hz_13tev`	1	-	$p_T(l) \geq$ 10 GeV, $\lvert\eta^l\rvert \leq 2.5$
	$d\sigma/dp_t^h$	`hz_pt_13tev`	10	[0,200] GeV	$p_T(l) \geq$ 10 GeV, $\lvert\eta^l\rvert \leq 2.5$
	$d\sigma/dy^h$	`hz_y_13tev`	10	[-2.5,2.5]	$p_T(l) \geq$ 10 GeV, $\lvert\eta^l\rvert \leq 2.5$
$ht\bar{t}$	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]	-

BACKUP: SMPDF-$t\bar{t}$ PROCESS DETAILS

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]	-

BACKUP: SMPDF-EW PROCESS DETAILS

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

BACKUP: SELECTED CORRELATIONS FOR hz PRODUCTION