mc2hessian

Common tools to estimate PDF uncertainties

Zahari Kassabov

in collaboration with S. Carrazza, S. Forte, J.I. Latorre and J. Rojo

First Annual Meeting of ITN HiggsTools, April 17, 2015, Freiburg

OUTLINE OF THE TALK

INTRODUCTION

PDF UNCERTAINTIES FROM DATA

PDFs have to be estimated from experimental data.

- Several groups perform *global fits.*

PDF UNCERTAINTIES FROM DATA

PDFs have to be estimated from experimental data.

- Several groups perform *global fits*.
  - Different statistical treatments, data sets, theory assumptions, ...

PDF UNCERTAINTIES FROM DATA

PDFs have to be estimated from experimental data.

- Several groups perform *global fits*.
  - Different statistical treatments, data sets, theory assumptions, ...
  - Need tools to combine, compare, benchmark.

PDFs have to be estimated from experimental data.

- Several groups perform *global fits*.
  - Different statistical treatments, data sets, theory assumptions, …
  - Need tools to combine, compare, benchmark.

PDF UNCERTAINTIES FROM DATA

PDFs have to be estimated from experimental data.

- Several groups perform *global fits*.
  - Different statistical treatments, data sets, theory assumptions, ...
  - Need tools to combine, compare, benchmark.
  - Need to distribute in a way useful for the community.

PDF PARAMETIZATIONS

There are two main ways of fitting PDFs from data:

There are two main ways of fitting PDFs from data:

Hessian approach Imagine the functional form is known, and guess parameters from data by maximum likelihood.

ORIGINAL IDEAS

There are two main ways of fitting PDFs from data:

Hessian approach Imagine the functional form is known, and guess parameters from data by maximum likelihood.

- Provide the mean and a *eigenvector error set*.

ORIGINAL IDEAS

There are two main ways of fitting PDFs from data:

Hessian approach Imagine the functional form is known, and guess parameters from data by maximum likelihood.

- Provide the mean and a *eigenvector error set*.

Monte Carlo approach Assume a very general functional form and fix parameters form data by maximum a posteriori.

ORIGINAL IDEAS

There are two main ways of fitting PDFs from data:

Hessian approach Imagine the functional form is known, and guess parameters from data by maximum likelihood.

- Provide the mean and a *eigenvector error set*.

Monte Carlo approach Assume a very general functional form and fix parameters form data by maximum a posteriori.

- Provide a set of functional forms, *“replicas”*.

There are two main ways of fitting PDFs from data:

Hessian approach Imagine the functional form is known, and guess parameters from data by maximum likelihood.

- Provide the mean and a *eigenvector error set*.

Monte Carlo approach Assume a very general functional form and fix parameters form data by maximum a posteriori.

- Provide a set of functional forms, “*replicas*”.

However

Both can be delivered as Hessian and MC representations.

- Can separate *fit strategy* from *representation*.

There are two main ways of fitting PDFs from data:

Hessian approach Imagine the functional form is known, and guess parameters from data by maximum likelihood.

- Provide the mean and a *eigenvector error set*.

Monte Carlo approach Assume a very general functional form and fix parameters form data by maximum a posteriori.

- Provide a set of functional forms, *"replicas"*.

However

Both can be delivered as Hessian and MC representations.

- Can separate *fit strategy* from *representation*.

We show how to transform Monte Carlo to Hessian.

Hessian approach Apply simple error propagation formula to the "eigenvalues" z_k, ie $f(x, Q^2) = f[z_k](x, Q^2)$.

Hessian approach Apply simple error propagation formula to the "eigenvalues" z_k, ie $f(x, Q^2) = f[z_k](x, Q^2)$.

- Assume small (linear Taylor expansion), and Gaussian errors.

$$(\Delta\mathcal{O}[f])^2 \propto \sum_k \left(\frac{\partial\mathcal{O}}{\partial z_k}\right)^2, \; z_k \sim \mathcal{N}(0,1)$$

Hessian approach Apply simple error propagation formula to the "eigenvalues" z_k, ie $f(x, Q^2) = f[z_k](x, Q^2)$.

- Assume small (linear Taylor expansion), and Gaussian errors.

$$(\Delta\mathcal{O}[f])^2 \propto \sum_k \left(\frac{\partial\mathcal{O}}{\partial z_k}\right)^2, \quad z_k \sim \mathcal{N}(0,1)$$

Hessian approach Apply simple error propagation formula to the "eigenvalues" z_k, ie $f(x, Q^2) = f[z_k](x, Q^2)$.

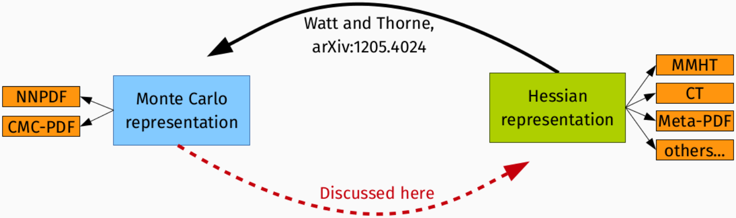
- Assume small (linear Taylor expansion), and Gaussian errors.

$$(\Delta\mathcal{O}[f])^2 \propto \sum_k \left(\frac{\partial\mathcal{O}}{\partial z_k}\right)^2,\ z_k \sim \mathcal{N}(0,1)$$

Monte Carlo approach Perform a Monte Carlo simulation sampling from the distribution of replicas.

$$\mathcal{O} \sim \mathcal{O}(f)$$

SUMMARY



Problem addressed here:

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

Hessian approach:

- Linear error propagation.

Monte Carlo approach:

ADVANTAGES AND DISADVANTAGES

Hessian approach:

- Linear error propagation.
- Errors Gaussian in the space of parameters.

Monte Carlo approach:

ADVANTAGES AND DISADVANTAGES

Hessian approach:

- Linear error propagation.
- Errors Gaussian in the space of parameters.
- Straightforward combination with other sources of uncertainty.

Monte Carlo approach:

Hessian approach:

- Linear error propagation.
- Errors Gaussian in the space of parameters.
- Straightforward combination with other sources of uncertainty.
- Efficient and easy implementation

Monte Carlo approach:

ADVANTAGES AND DISADVANTAGES

Hessian approach:

- Linear error propagation.
- Errors Gaussian in the space of parameters.
- Straightforward combination with other sources of uncertainty.
- Efficient and easy implementation

Monte Carlo approach:

- Arbitrary error propagation.

ADVANTAGES AND DISADVANTAGES

Hessian approach:

- Linear error propagation.
- Errors Gaussian in the space of parameters.
- Straightforward combination with other sources of uncertainty.
- Efficient and easy implementation

Monte Carlo approach:

- Arbitrary error propagation.
- Easy combination of multiple PDF sets.

ADVANTAGES AND DISADVANTAGES

Hessian approach:

- Linear error propagation.
- Errors Gaussian in the space of parameters.
- Straightforward combination with other sources of uncertainty.
- Efficient and easy implementation

Monte Carlo approach:

- Arbitrary error propagation.
- Easy combination of multiple PDF sets.
- Much less *functional bias.*

THE `mc2hessian` ALGORITHM

Meta-PDF [arXiv:1401.0013] is an alternative method to combine different PDF sets:

Meta-PDF [arXiv:1401.0013] is an alternative method to combine different PDF sets:

1. Convert Hessian sets to Monte Carlo.

Meta-PDF [arXiv:1401.0013] is an alternative method to combine different PDF sets:

1. Convert Hessian sets to Monte Carlo.
2. Combine all MC replicas.

Meta-PDF [arXiv:1401.0013] is an alternative method to combine different PDF sets:

1. Convert Hessian sets to Monte Carlo.
2. Combine all MC replicas.
3. Fit each replica to a “Hessian like” functional form.

Meta-PDF [arXiv:1401.0013] is an alternative method to combine different PDF sets:

1. Convert Hessian sets to Monte Carlo.
2. Combine all MC replicas.
3. Fit each replica to a “Hessian like” functional form.
4. Produce final Hessian set from fitted parameters.

Meta-PDF [arXiv:1401.0013] is an alternative method to combine different PDF sets:

1. Convert Hessian sets to Monte Carlo.
2. Combine all MC replicas.
3. Fit each replica to a “Hessian like” functional form.
4. Produce final Hessian set from fitted parameters.

- Introduces functional bias.

Given a Monte Carlo prior set of PDFs

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

use a subset of replicas as parameters of linear expansion:

$$f_\alpha^{(k)} \approx f_{H,\alpha}^{(k)} \equiv f_\alpha^{(0)} + \sum_{i=1}^{N_{\rm eig}} a_i^{(k)} (\eta_\alpha^{(i)} - f_\alpha^{(0)})\,, \quad k = 1, \ldots, N_{\rm rep}$$

Given a Monte Carlo prior set of PDFs

$$\{f_\alpha^{(k)}\}_{k=1,\ldots,N_{\rm rep}}\,,\qquad \alpha=\{g,u,d,s,\ldots\}\,,$$

use a subset of replicas as parameters of linear expansion:

$$f_\alpha^{(k)} \approx f_{H,\alpha}^{(k)} \equiv f_\alpha^{(0)} + \sum_{i=1}^{N_{\rm eig}} a_i^{(k)}(\eta_\alpha^{(i)} - f_\alpha^{(0)})\,,\qquad k=1,\ldots,N_{\rm rep}$$

- approximate each replica of the original MC ensemble $f_\alpha^{(k)}$

Given a Monte Carlo prior set of PDFs

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

use a subset of replicas as parameters of linear expansion:

$$f_\alpha^{(k)} \approx f_{H,\alpha}^{(k)} \equiv f_\alpha^{(0)} + \sum_{i=1}^{N_{\mathrm{eig}}} a_i^{(k)} (\eta_\alpha^{(i)} - f_\alpha^{(0)})\,, \quad k = 1, \ldots, N_{\mathrm{rep}}$$

- approximate each replica of the original MC ensemble $f_\alpha^{(k)}$
- by the linear combination $f_{H,\alpha}^{(k)}$

Given a Monte Carlo prior set of PDFs

$$\{f_\alpha^{(k)}\}_{k=1,\dots,N_{\rm rep}}\,, \quad \alpha = \{g, u, d, s, \dots\}\,,$$

use a subset of replicas as parameters of linear expansion:

$$f_\alpha^{(k)} \approx f_{H,\alpha}^{(k)} \equiv f_\alpha^{(0)} + \sum_{i=1}^{N_{\rm eig}} a_i^{(k)} (\eta_\alpha^{(i)} - f_\alpha^{(0)})\,, \quad k = 1, \dots, N_{\rm rep}$$

- approximate each replica of the original MC ensemble $f_\alpha^{(k)}$
- by the linear combination $f_{H,\alpha}^{(k)}$
- with coefficients $a_i^{(k)}$

Given a Monte Carlo prior set of PDFs

$$\{f_\alpha^{(k)}\}_{k=1,\dots,N_{\rm rep}}\,, \quad \alpha = \{g, u, d, s, \dots\}\,,$$

use a subset of replicas as parameters of linear expansion:

$$f_\alpha^{(k)} \approx f_{H,\alpha}^{(k)} \equiv f_\alpha^{(0)} + \sum_{i=1}^{N_{\rm eig}} a_i^{(k)}(\eta_\alpha^{(i)} - f_\alpha^{(0)})\,, \quad k = 1, \dots, N_{\rm rep}$$

- approximate each replica of the original MC ensemble $f_\alpha^{(k)}$
- by the linear combination $f_{H,\alpha}^{(k)}$
- with coefficients $a_i^{(k)}$
- of deviations from the central value $f_\alpha^{(0)}$

Given a Monte Carlo prior set of PDFs

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

use a subset of replicas as parameters of linear expansion:

$$f_\alpha^{(k)} \approx f_{H,\alpha}^{(k)} \equiv f_\alpha^{(0)} + \sum_{i=1}^{N_{\rm eig}} a_i^{(k)}(\eta_\alpha^{(i)} - f_\alpha^{(0)})\,, \quad k=1,\ldots,N_{\rm rep}$$

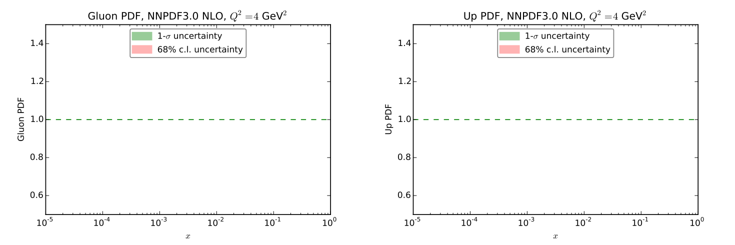
- approximate each replica of the original MC ensemble $f_\alpha^{(k)}$
- by the linear combination $f_{H,\alpha}^{(k)}$
- with coefficients $a_i^{(k)}$
- of deviations from the central value $f_\alpha^{(0)}$
- expanded in the basis of a subset of replicas $\{\eta_\alpha^{(i)}\}_{i=1,\ldots,N_{\rm eig}} \subset \{f_\alpha^{(k)}\}$

DESCRIPTION OF THE METHOD

- We want to go from $N_{rep} = 1000$ MC replicas to N_{eig} eigenvectors.

- We want to go from $N_{rep} = 1000$ MC replicas to N_{eig} eigenvectors.
- We are interested in reproducing Gaussian regions of the PDF:

$$\epsilon = \left| \frac{\sigma - (68\%\,\text{c.l})}{\sigma} \right|$$



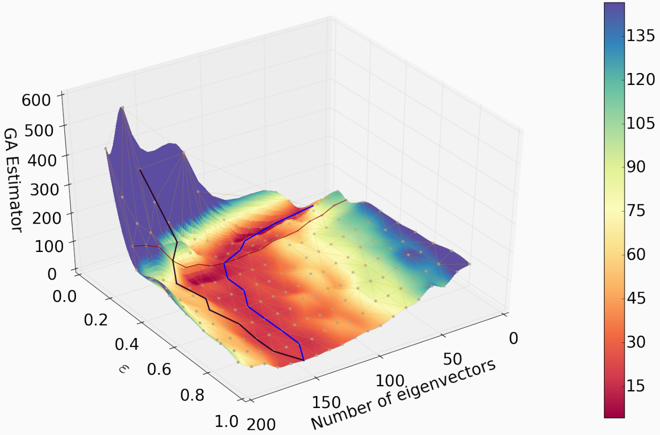
- We want to go from $N_{rep} = 1000$ MC replicas to N_{eig} eigenvectors.
- We are interested in reproducing Gaussian regions of the PDF:

$$\epsilon = \left| \frac{\sigma - (68\%\,\text{c.l})}{\sigma} \right|$$

- We construct a *figure of merit* and optimize with a *genetic algorithm*:

$$\mathrm{ERF}_{\sigma} = \sum_{i=1}^{N_x} \sum_{\alpha=1}^{N_f} \left| \frac{\sigma^{\mathrm{PDF}}_{H,\alpha}(x_i, Q_0^2) - \sigma^{\mathrm{PDF}}_{\alpha}(x_i, Q_0^2)}{\sigma^{\mathrm{PDF}}_{\alpha}(x_i, Q_0^2)} \right|$$

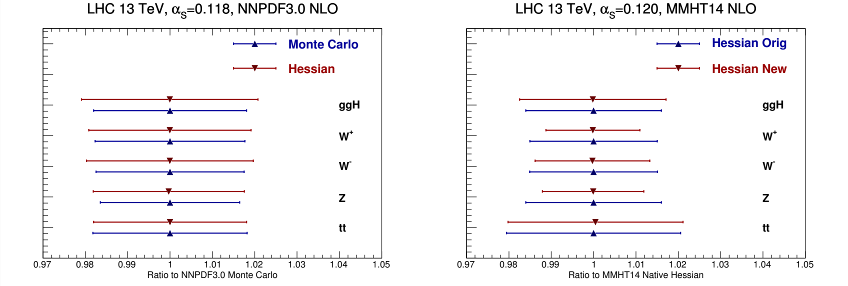
SELECTING THE OPTIMAL BASIS



- **Surface:** GA minimum for estimator in function of ϵ and N_{eig}.
- **Blue curve:** surface minimum; **black curve:** estimator with large ϵ.

PHENOMENOLOGY

LHC inclusive cross-sections @ 13 TeV



- **Good agreement** for LHC inclusive cross-sections, below 10%.
- **Also** for a large number of differential distributions at the LHC 7 TeV.

ANOTHER IDEA

- Use the whole replica set to form the linear combinations (not a subset).

ANOTHER IDEA

- Use the whole replica set to form the linear combinations (not a subset).
- Maximize *"agreement"* of Hessian and MC covariance matrices.

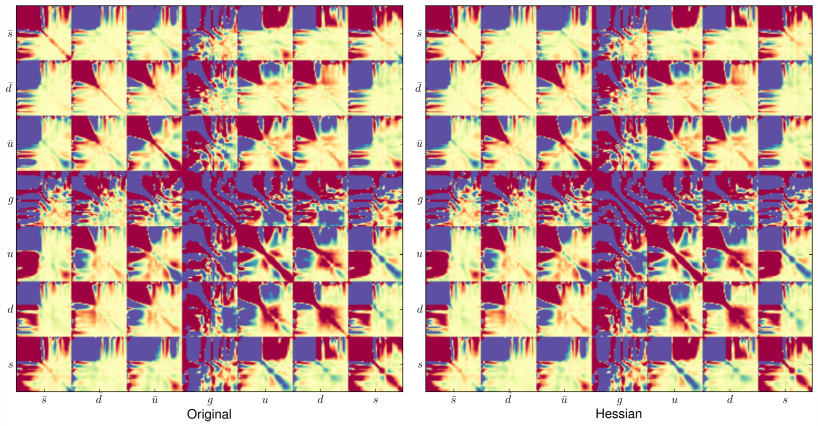
ANOTHER IDEA

- Use the whole replica set to form the linear combinations (not a subset).
- Maximize *"agreement"* of Hessian and MC covariance matrices.
- It can be reduced to a lineal algebra problem!

ANOTHER IDEA

- Use the whole replica set to form the linear combinations (not a subset).
- Maximize *"agreement"* of Hessian and MC covariance matrices.
- It can be reduced to a lineal algebra problem!
  - (pick linear combinations of replicas corresponding to the dominant eigenvalues, using singular value decomposition).

Results for 100-eigenvector Hessian.



DELIVERY

- The `mc2hessian` program is public available at

  `github.com/scarrazza/mc2hessian`

- Further **optimizations** in progress before final release.
- NNPDF3.0 Hessian version available in `LHAPDF6` soon:
  - `NNPDF30_nlo_as_0118_hessian`
  - `NNPDF30_nnlo_as_0118_hessian`
- Any other MC set can be converted using directly the public code.

Current PDF4LHC prescription is to use combined PDF sets.

- Combine using a Monte Carlo sample of each set.

OUTLOOK

Current PDF4LHC prescription is to use combined PDF sets.

- Combine using a Monte Carlo sample of each set.
- Deliver final Hessian set (as experiments prefer).

Current PDF4LHC prescription is to use combined PDF sets.

- Combine using a Monte Carlo sample of each set.
- Deliver final Hessian set (as experiments prefer).

Hopefully `mc2hessian` will be used to deliver the Standard PDFs for tasks like Higgs cross section measurements.

Questions?