# Statistical Methods for Parton Distribution Functions

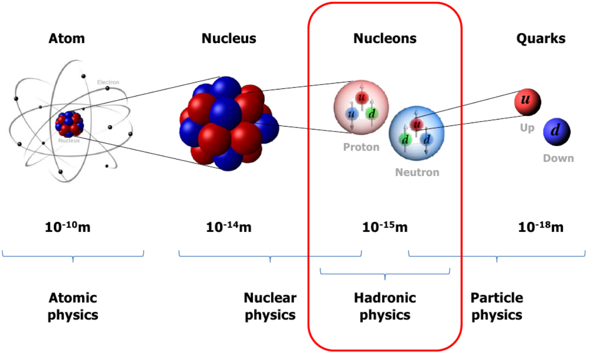
Uncertainty Quantification at the Extremes (ISNET-6)

Emanuele R. Nocera

Nikhef Theory Group, Amsterdam

TU Darmstadt – 10 October 2018

# Foreword



Nucleons make up all nuclei, and hence most of the visible matter in the Universe

Nucleons are bound states of quarks and gluons (partons)
whose dynamics is described by Quantum Chromodynamics (QCD)

Parton Distribution Functions (PDFs), $f(x, Q^2)$, can be (roughly) regarded
as the probability of finding a parton $f$ in a nucleon
$x$ is the fraction of the nucleon's momentum carried by the parton
$Q^2$ is the characteristic energy scale of the interaction

# Theoretical Input

1. Collinear leading twist factorisation of physical observables [Adv.Ser.Direct.HEP 5 (1988) 1]

$$\mathcal{O}_I = \sum_{f=q,\bar{q},g} C_{If}(x,\alpha_s(Q^2))\otimes f(x,Q^2) + \text{p.s. corrections} \qquad f\otimes g = \int_x^1 \frac{dy}{y} f\left(\frac{x}{y}\right) g(y)$$

2. Evolution of PDFs; DGLAP equations [NPB 126 (1977) 298]

$$\frac{\partial}{\partial \ln Q^2} f_i(x,Q^2) = \sum_j^{n_f} \int_x^1 \frac{dz}{z} P_{ji}\left(z,\alpha_s(Q^2)\right) f_j\left(\frac{x}{z},Q^2\right)$$

3. Perturbative expansion of coefficient and splitting functions

$$C_{If}(y,\alpha_s) = \sum_{k=0} a_s^k C_{If}^{(k)}(y), \qquad P_{ji}(z,\alpha_s) = \sum_{k=0} a_s^{k+1} P_{ji}^{(k)}(z), \qquad a_s = \alpha_s/(4\pi)$$

$C_{If}$ known up to NNLO for an increasing number of processes $P_{ji}$ known up to NNLO

4. Theoretical constraints
positivity of cross sections, sum rules, symmetries, integrability of some PDF combinations

5. The dependence of the PDF on $x$ cannot be computed from perturbative QCD
it must be determined from a (global) analysis of experimental data

# The problem: a global determination of PDFs

Determine the probability density (measure) $\mathcal{P}[f]$ in the space of PDFs $[f]$

For any observable $\mathcal{O}$ depending on a set of PDFs $[f]$

its expectation value and uncertainty are functional integrals over the space of PDFs

$$\langle\mathcal{O}[f]\rangle = \int \mathcal{D}f\,\mathcal{P}[f]\,\mathcal{O}[f] \qquad \text{expectation value}$$

$$\sigma_{\mathcal{O}}[f] = \left[\int \mathcal{D}f\,\mathcal{P}[f]\,\left(\mathcal{O}[f]-\langle\mathcal{O}[f]\rangle\right)^2\right]^{1/2} \qquad \text{uncertainty}$$

ILL-DEFINED PROBLEM

determine a set of infinite-dimensional objects (the PDFs of each parton)
from a finite piece of information (the experimental data points $N_{\rm dat}\approx\mathcal{O}(5000)$)

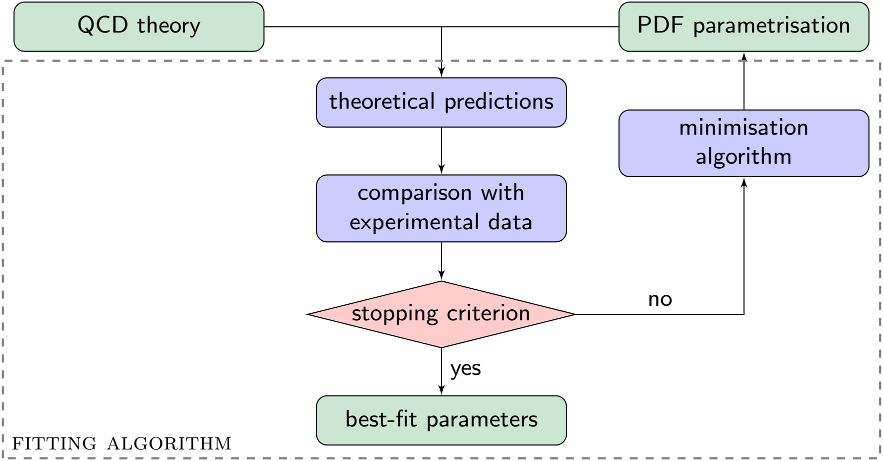
Choose a PDF parametrisation at an initial scale $Q_0^2$ for each independent parton $f$

$$xf(x,Q_0^2) = A_f\,x^{a_f}\,(1-x)^{b_f}\,\mathscr{F}(x,\{c_f\})$$

that should be sufficiently GENERAL, SMOOTH, FLEXIBLE (or MINIMALLY BIASED)

Problem reduced to determine the optimal set of parameters $\{\mathbf{a}\}=\{A_f,a_f,b_f,\{c_f\}\}$

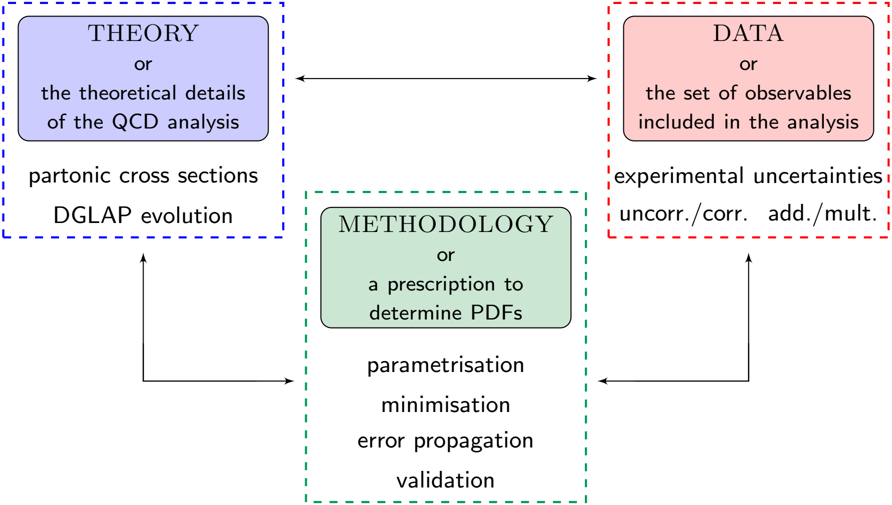
# A global PDF determination: the underlying strategy



$$\chi^2 = \frac{1}{N_{\rm dat}} \sum_{i,j}^{\rm N_{dat}} \left(\mathcal{T}_i[f] - D_i\right) \left({\rm cov}^{-1}\right)_{ij} \left(\mathcal{T}_j[f] - D_j\right)$$

$$({\rm cov})_{ij} = \delta_{ij} s_i^2 + \sum_{\alpha}^{N_c} \sigma_{i,\alpha}^{(c)} \sigma_{j,\alpha}^{(c)} D_i D_j + \sum_{\alpha}^{N_{\mathcal{L}}} \sigma_{i,\alpha}^{(\mathcal{L})} \sigma_{j,\alpha}^{(\mathcal{L})} \mathcal{T}_i^0 \mathcal{T}_j^0$$

# A global PDF determination: the ingredients we need



THEORY

or

the theoretical details of the QCD analysis

partonic cross sections

DGLAP evolution

DATA

or

the set of observables included in the analysis

experimental uncertainties

uncorr./corr. add./mult.

METHODOLOGY

or

a prescription to determine PDFs

parametrisation

minimisation

error propagation

validation

Each of these ingredients is a source of uncertainty in the PDF determination
PDFs with *faithful* uncertainties are crucial *e.g.* for SM/BSM physics at the LHC

# A. Uncertainty representation

- Propagating the data uncertainty: the Hessian and the Monte Carlo methods
- Characterising methodological uncertainties: closure tests
- Theoretical uncertainties in PDF fits: a proposal

# The Hessian method: general strategy

1. Expand the $\chi^2$ about its global minimum at first (nontrivial) order

$$\chi^2\{\mathbf{a}\} \approx \chi^2\{\mathbf{a_0}\} + \delta a^i H_{ij} \delta a^j , \qquad H_{ij} = \frac{1}{2}\left.\frac{\partial^2\chi^2\{\mathbf{a}\}}{\partial a_i \partial a_j}\right|_{\{\mathbf{a}\}=\{\mathbf{a_0}\}}$$

2. Assume linear error propagation for any observable $\mathcal{O}$ depending on $\{\mathbf{a}\}$

$$\langle\mathcal{O}\{\mathbf{a}\}\rangle \approx \mathcal{O}\{\mathbf{a_0}\} + a_i \left.\frac{\partial\mathcal{O}\{\mathbf{a}\}}{\partial a_i}\right|_{\{\mathbf{a}\}=\{\mathbf{a_0}\}} \qquad \sigma_{\mathcal{O}\{\mathbf{a}\}} \approx \sigma_{ij} \left.\frac{\partial\mathcal{O}\{\mathbf{a}\}}{\partial a_i} \; \frac{\partial\mathcal{O}\{\mathbf{a}\}}{\partial a_j}\right|_{\{\mathbf{a}\}=\{\mathbf{a_0}\}}$$

3. Determine $\sigma_{ij}$ from $H_{ij}$ from maximum likelihood (under Gaussian hypothesis)

$$\sigma_{ij}^{-1} = \left.\frac{\partial^2\chi^2\{\mathbf{a}\}}{\partial a_i \partial a_j}\right|_{\{\mathbf{a}\}=\{\mathbf{a_0}\}} = H_{ij}$$

4. A C.L. about the best fit is obtained as the volume (in parameter space) about $\chi^2\{\mathbf{a_0}\}$ that corresponds to a fixed increase of the $\chi^2$; for Gaussian uncertainties:

$$68\% \text{ C.L.} \Longleftrightarrow \Delta\chi^2 = \chi^2\{\mathbf{a}\} - \chi^2\{\mathbf{a_0}\} = 1$$

# The Hessian method: some remarks

1. Compact representation and computation of observables and their uncertainties

$$\langle\mathcal{O}[f(x,Q^2)]\rangle = \mathcal{O}[f_0(x,Q^2)]$$

$$\sigma_{\mathcal{O}}[f(x,Q^2)] = \frac{1}{2}\left[\sum_{i=1}^{N_{\rm par}}\left(\mathcal{O}[f_i(x,Q^2)] - \mathcal{O}[f_0(x,Q^2)]\right)^2\right]^{1/2}$$

2. Parameters can always be adjusted so that all eigenvalues of $H_{ij}$ are equal to one (diagonalise $H_{ij}$ and rescale the eigenvectors by their eigenvalues)

$$\delta a_i H_{ij}\delta a_j = \sum_{i=1}^{N_{\rm par}}\left[a_i'(a_i)\right]^2 \Longleftrightarrow \sigma_{\mathcal{O}\{\mathbf{a}'\}} = \left|\nabla'\mathcal{O}\{\mathbf{a}'\}\right|$$

   The total contribution to the uncertainty due to two different sources (possibly correlated) is obtained by simply adding them in quadrature

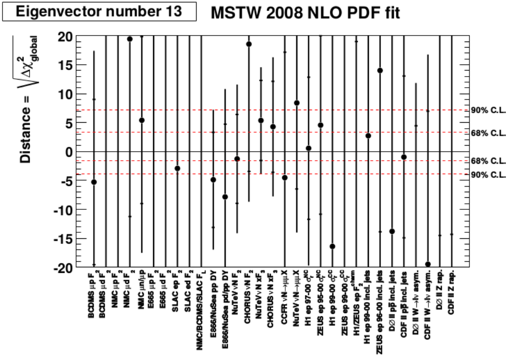
3. Any rotation in the space of parameters preserves the gradient (one can diagonalise a chosen observable without spoiling the result)

4. Unmanageable Hessian matrix if the numer of parameters is huge

# The Hessian method: limitations

Uncertainties obtained with $\Delta\chi^2 = 1$ might be unrealistically small
(inadequacy of the linear approximation)

MSTW TOLERANCE PLOT FOR 13TH EIGENVEC.



uncertainties tuned to the distribution of deviations from best-fits for single experiments

for each eigenvector in parameter space

determine the CL for the distribution of best-fits of each experiment

rescale to the $\Delta\chi^2 = T$ interval such that correct confidence intervals are reproduced

no statistically rigorous interpretation of $T$ (tolerance)

# The Monte Carlo method: general strategy

1. Generate $(art)$ replicas of $(exp)$ data according to the distribution

$$\mathcal{O}_i^{(art)(k)} = \mathcal{O}_i^{(exp)} + r_i^{(k)}\sigma_{\mathcal{O}_i}\,, \qquad i=1,\dots N_{\rm dat}\,, \qquad k=1,\dots,N_{\rm rep}$$

where $r_i^{(k)}$ are (Gaussianly distributed) random numbers for each $k$-th replica ($r_i^{(k)}$ can be generated with any distribution, not neccesarily Gaussian)

2. Perform a fit for each replica $k=1,\dots,N_{\rm rep}$

3. Compact computation of observables and their uncertainties (PDF replicas are equally probable members of a statistical ensemble)

$$\langle\mathcal{O}[f(x,Q^2)]\rangle = \frac{1}{N_{\sf rep}}\sum_{k=1}^{N_{\sf rep}}\mathcal{O}[f^{(k)}(x,Q^2)]$$

$$\sigma_{\mathcal{O}}[f(x,Q^2)] = \left[\frac{1}{N_{\sf rep}-1}\sum_{k=1}^{N_{\sf rep}}\Big(\mathcal{O}[f^{(k)}(x,Q^2)] - \langle\mathcal{O}[f(x,Q^2)]\rangle\Big)^2\right]^{1/2}$$

⇒ **no need to rely on linear approximation**

⇒ **computational expensive: need to perform $N_{\rm rep}$ fits instead of one**

# The Monte Carlo method: determining the sample size

1. Generalise the way in which artificial replicas are generated

$$\mathcal{O}_i^{(\mathrm{art}),(k)}(x,Q^2) = \left[1 + \sum_c r_{c,i}^{(k)}\sigma_{c,i} + r_{s,i}^{(k)}\sigma_{s,i}\right]\mathcal{O}_i^{(\mathrm{exp})}(x,Q^2)$$

$\sigma_{c,p}$: correlated uncertainties

$\sigma_{s,p}$: uncorrelated uncertainties

$r_{c,p}^{(k)}$, $r_{s,p}^{(k)}$: Gaussian random numbers

2. Define proper estimators to determine the sample size

| | $\left\langle \mathrm{PE}\left[\langle\mathcal{O}^{(\mathrm{art})}\rangle\right]\right\rangle$ [%] | | | $r\left[\mathcal{O}^{(\mathrm{art})}\right]$ | | |
|---|---|---|---|---|---|---|
| $N_{\mathrm{rep}}$ | 10 | 100 | 1000 | 10 | 100 | 1000 |
| Exp.1 | 23.7 | 3.5 | 2.9 | .76037 | .99547 | .99712 |
| Exp.2 | 19.4 | 5.6 | 1.2 | .94789 | .99908 | .99993 |
| ... | ... | ... | ... | ... | ... | ... |

$$\left\langle PE\left[\langle\mathcal{O}^{(\mathrm{art})}\rangle_{\mathrm{rep}}\right]\right\rangle_{\mathrm{dat}} = \frac{1}{N_{\mathrm{dat}}}\sum_{i=1}^{N_{\mathrm{dat}}}\left|\frac{\langle\mathcal{O}_{\mathrm{i}}^{(\mathrm{art})}\rangle_{\mathrm{rep}} - \mathcal{O}_{\mathrm{i}}^{(\mathrm{exp})}}{\mathcal{O}_{\mathrm{i}}^{(\mathrm{exp})}}\right|$$

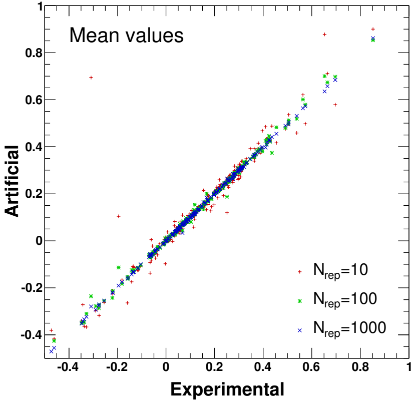
Percentage Error

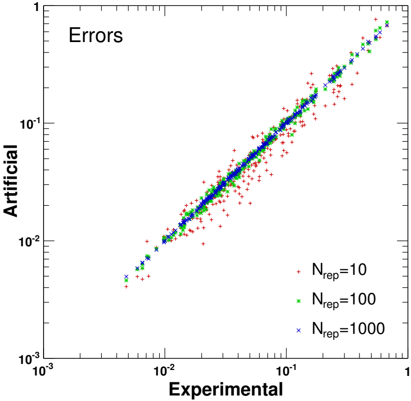
$$r\left[\mathcal{O}^{(\mathrm{art})}\right] = \frac{\left\langle\mathcal{O}^{(\mathrm{exp})}\langle\mathcal{O}^{(\mathrm{art})}\rangle_{\mathrm{rep}}\right\rangle_{\mathrm{dat}} - \left\langle\mathcal{O}^{(\mathrm{exp})}\right\rangle_{\mathrm{dat}}\left\langle\langle\mathcal{O}^{(\mathrm{art})}\rangle_{\mathrm{rep}}\right\rangle_{\mathrm{dat}}}{\sigma_{\mathrm{s}}^{(\mathrm{exp})}\sigma_{\mathrm{s}}^{(\mathrm{art})}}$$

Scatter Correlation

# The Monte Carlo method: determining the sample size

Require that the average over the replicas reproduces the central value of the original experimental data to a desired accuracy (the standard deviation reproduces the error and so on)

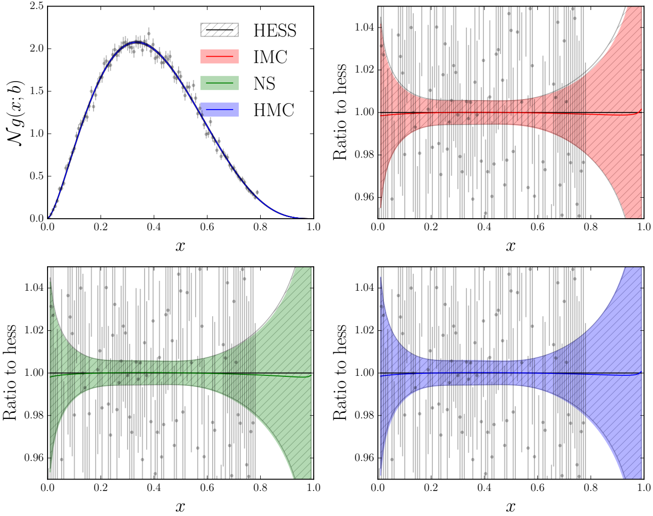




Accuracy of few % requires $\sim 100$ replicas

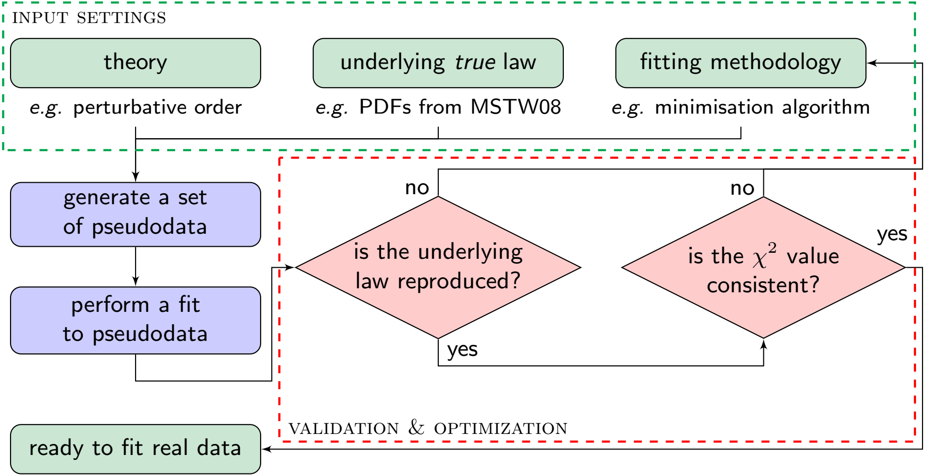
# Alternative approaches to represent the data uncertainty

1. Nested Sampling (NS) [PRL,120 (2018) 152502, arXiv:1804.01965]
2. (Hybrid) Monte Carlo Markov Chain (MCMC) [EPJ C75 (2015) 304]



# Closure tests: general idea [JHEP 1504 (2015) 040]

Validation and optimisation of the fitting strategy with known underlying physical law


INPUT SETTINGS
theory
*e.g.* perturbative order
underlying *true* law
*e.g.* PDFs from MSTW08
fitting methodology
*e.g.* minimisation algorithm
generate a set of pseudodata
perform a fit to pseudodata
is the underlying law reproduced?
no
yes
is the $\chi^2$ value consistent?
no
yes
VALIDATION & OPTIMIZATION
ready to fit real data

Full control/characterisation of methodological uncertainties

# Closure tests: levels

1. Level 0: generate pseudodata $D_i^0$ with zero uncertainty
   (but $(\mathrm{cov})_{ij}$ in the $\chi^2$ is the data covariance matrix)
   → fit quality can be arbitrarily good, if the fitting methodology is efficient: $\chi^2/N_{\rm dat} \sim 0$
   → validate fitting methodology (parametrisation, minimisation)
   → interpolation and extrapolation uncertainty
2. Level 1: generate pseudodata $D_i^1$ with stochastic fluctuations (no replicas)

$$D_i^1 = (1 + r_i^{\rm nor}\sigma_i^{\rm nor})\left(D_i^0 + \sum_p^{N_{\rm sys}} r_{i,p}^{\rm sys}\sigma_{i,p}^{\rm sys} + r_i^{\rm stat}\sigma_i^{\rm stat}\right)$$

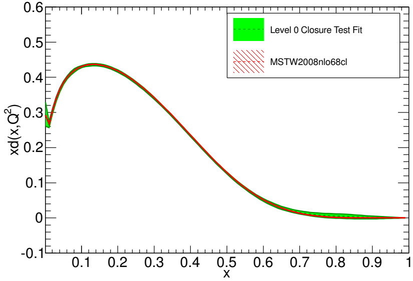
   → experimental uncertainties are not propagated into FFs: $\chi^2/N_{\rm dat} \sim 1$
   → functional uncertainty (a large number of functional forms with equally good $\chi^2$)
3. Level 2: generate $N_{\rm rep}$ Monte Carlo pseudodata replicas $D_i^{2,k}$ on top of Level 2

$$D_i^{2,k} = (1 + r_i^{{\rm nor},k}\sigma_i^{\rm nor})\left(D_i^1 + \sum_p^{N_{\rm sys}} r_{i,p}^{{\rm sys},k}\sigma_{i,p}^{\rm sys} + r_i^{{\rm stat},k}\sigma_i^{\rm stat}\right)$$

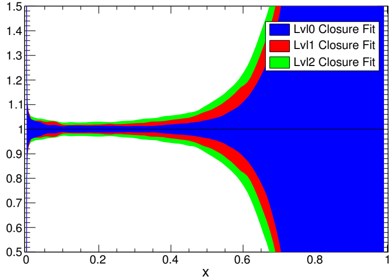
   → propagate the fluctuations due to experimental uncertainties into PDFs: $\chi^2/N_{\rm dat} \sim 1$
   → input PDFs within the one-sigma band of the fitted PDFs with a probability of ∼ 68%
   → data uncertainty

# Closure tests: examples

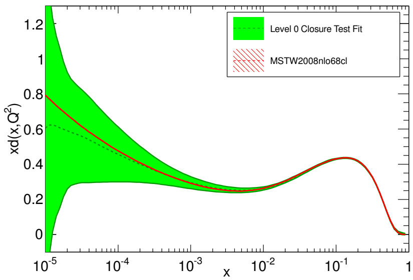
Level 0 closure test vs. MSTW



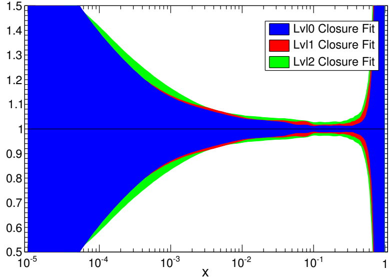
Ratios of d at different closure test levels



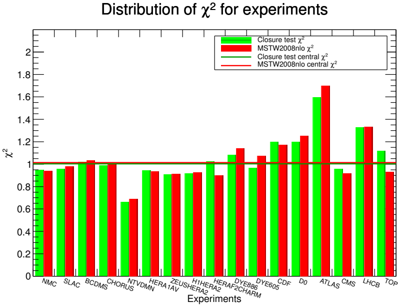
Level 0 closure test vs. MSTW

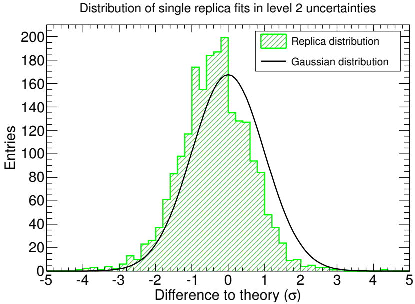


Ratios of d at different closure test levels



# Closure tests: statistical estimators





$$\chi^2 = \frac{1}{N_{\rm dat}} \sum_{i,j}^{\rm N_{dat}} \left(\mathcal{T}_i[f] - D_i\right) \left(\mathrm{cov}^{-1}\right)_{ij} \left(\mathcal{T}_j[f] - D_j\right)$$

$$\varphi_{\chi^2} \equiv \sqrt{\langle \chi^2[\mathcal{T}[f_{\rm fit}], \mathcal{D}_0] \rangle - \chi^2[\langle \mathcal{T}[f_{\rm fit}] \rangle, \mathcal{D}_0]}$$

$$\Delta_{\chi^2} = \frac{\chi^2[\langle \mathcal{T}[f] \rangle, \mathcal{D}_1] - \chi^2[\mathcal{T}[f_{\rm in}], \mathcal{D}_1]}{\chi^2[\mathcal{T}[f_{\rm in}], \mathcal{D}_1]}$$

$$\xi_{n\sigma} = \frac{1}{N_{\rm PDF}} \frac{1}{N_x} \frac{1}{N_{\rm fits}} \sum_{i=1}^{N_{\rm PDF}} \sum_{j=1}^{N_x} \sum_{l=1}^{N_{\rm fits}} I_{\left[-n\sigma_{\rm fit}^{i(l)}(x_j), n\sigma_{\rm fit}^{i(l)}(x_j)\right]} \left(\langle f_{\rm fit}^{i(l)}(x_j) \rangle - f_{\rm in}^{i}(x_j)\right)$$

$\chi^2 = 1.15$ $\varphi_{\chi^2} = 0.254$ $\Delta_{\chi^2} = -0.011$ $\xi_{1\sigma} = 0.699$ $\xi_{2\sigma} = 0.948$ (lvl2)

$\chi^2 = 1.12$ $\varphi_{\chi^2} = 0.173$ $\Delta_{\chi^2} = -0.015$ $\xi_{1\sigma} = 0.512$ $\xi_{2\sigma} = 0.836$ (lvl1)

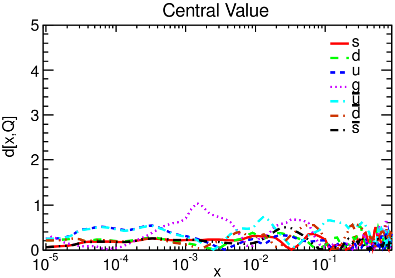
# Closure tests: applications

$$d_\sigma[x,Q] \equiv \sqrt{\frac{\left(\bar{f}_{i,\mathrm{fit}}(x,Q)-f_{i,\mathrm{in}}(x,Q)\right)^2}{\sigma^2\left[f_{i,\mathrm{fit}}\right](x,Q)}}$$

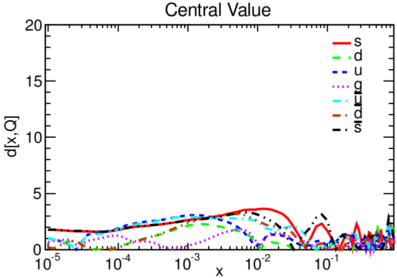
$$\bar{f}(x,Q^2) \equiv \left\langle f(x,Q^2)\right\rangle_{\mathrm{rep}} = \frac{1}{N_{\mathrm{rep}}}\sum_k^{N_{\mathrm{rep}}} f^{(k)}(x,Q^2)$$

$$\sigma^2\left[f\right](x,Q^2) = \frac{1}{N_{\mathrm{rep}}-1}\sum_k^{N_{\mathrm{rep}}}\left(f^{(k)}(x,Q^2)-\left\langle f(x,Q^2)\right\rangle_{\mathrm{rep}}\right)^2$$

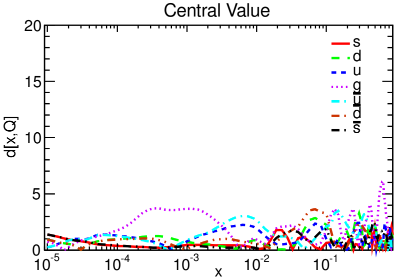
30k vs 80k generations



flavour basis vs evolution basis



300 vs 37 parameters



# Theoretical uncertainties in PDF fits [arXiv:1801.04842]

1. Denote the $n$ data in a particular process by a vector $y$, and the corresponding theoretical predictions by $\mathcal{T}[f]=\sum_{m=0}^{p}c_m[f]$. By Bayes' theorem

$$P(\mathcal{T}|y)=\frac{P(y|\mathcal{T})P(\mathcal{T})}{P(y|f)}$$

2. We assume Gaussian uncertainties for the data, *i.e.*,

$$P(y|\mathcal{T})\sim\exp\left\{-\frac{1}{2}(y-\mathcal{T})^T\sigma^{-1}(y-\mathcal{T})\right\}$$

3. We assume Gaussian theoretical uncertainties, *i.e.*,

$$P(\mathcal{T})=\prod_{m=0}^{p}P(c_m)\qquad P(c_m)\propto\exp\left(-\frac{1}{2}c_m^Ts^{-1}c_m\right)$$

4. Therefore, $P(\mathcal{T}|y)\sim\exp\left(-\frac{1}{2}\chi^{2\prime}\right)$ with

$$\chi^2\to\chi^{2'}=(y-\mathcal{T}[f])^T(\sigma+s)^{-1}(y-\mathcal{T}[f])$$

Problem reduced to determine the theoretical covariance matrix $s$

# B. Monte Carlo optimisation and delivery

- Fitting without refitting: Bayesian reweighting
- Compression of Monte Carlo sets and specialised minimal PDF sets
- Hessian to Monte Carlo and Monte Carlo to Hessian

# Bayesian reweighting [PRD 58 (1998) 094023, NPB 849 (2011) 112, NPB 855 (2012) 608]

Assess the impact of including a new data set $\{y\} = \{y_1, \dots, y_n\}$ in an old PDF set

1. Evaluate the agreement between new data and each replica $f_k$ in a prior ensemble

$$\chi^2_k(\{y\},\{f_k\}) = \sum_{i,j}^{n} \{y_i - \mathcal{T}_i[f_k]\}\, \sigma^{-1}_{ij}\, \{y_j - \mathcal{T}_j[f_k]\}$$

2. Apply Bayes' theorem to determine the conditional probability of PDF upon the inclusion of the new data and update the probability density in the space of PDFs

$$\mathcal{P}_{\rm new} = \mathcal{N}_\chi \mathcal{P}(\chi^2_k|\{f_k\})\mathcal{P}_{\rm old}(\{f_k\}) \qquad \mathcal{P}(\chi^2_k|\{f_k\}) = [\chi^2_k(\{y\},\{f_k\}]^{\frac{1}{2}(n-1)} e^{-\frac{1}{2}\chi^2_k(\{y\},\{f_k\})}$$

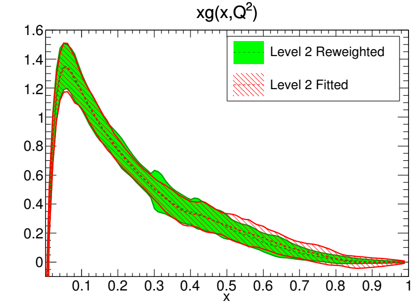
3. Replicas are no longer equally probable. Expectation values are given by

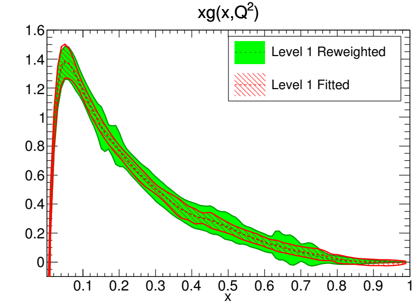
$$\langle \mathcal{O}[f_i(x,Q^2]\rangle_{\rm new} = \sum_{k=1}^{N_{\rm rep}} w_k \mathcal{O}[f_i^{(k)}(x,Q^2)]$$

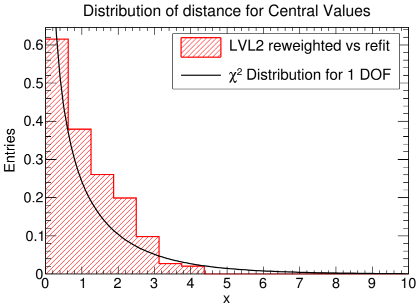
$$w_k \propto [\chi^2_k(\{y\},\{f_k\})]^{\frac{1}{2}(n-1)} e^{-\frac{1}{2}\chi^2_k(\{y\},\{f_k\})} \quad \text{with} \quad N_{\rm rep} = \sum_{k=1}^{N_{\rm rep}} w_k$$

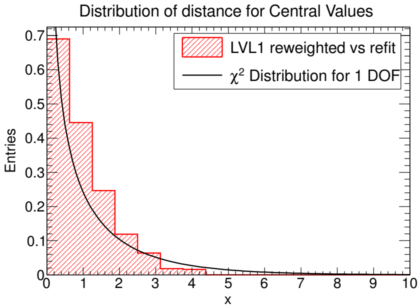
4. Loss of efficiency: $N_{\rm eff} \equiv \exp\left[-\sum_{k=1}^{N_{\rm rep}} p_k \log p_k\right]$ with $p_k = w_k/N_{\rm rep}$
   $0 < N_{\rm eff} < N_{\rm rep}$; $N_{\rm eff}$ must not be too low ⇒ increase the number of replicas in prior

# Bayesian reweighting in action



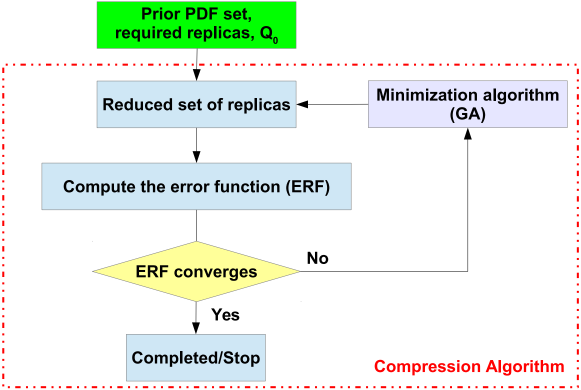






# Compressed Monte Carlo PDF sets [EPJ C75 (2015) 474]

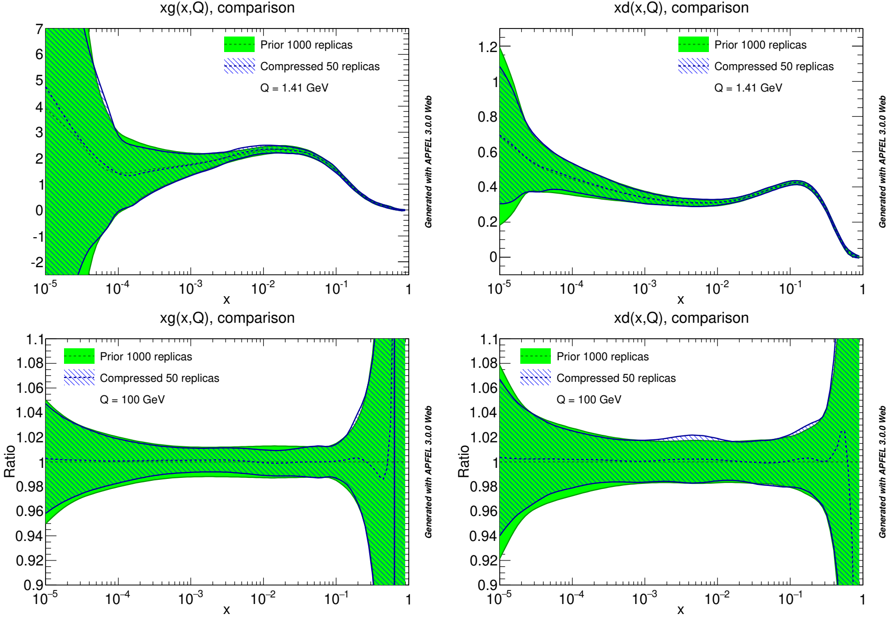
Reduce $N_{\rm rep}$ PDFs to $N'_{\rm rep} < N_{\rm rep}$ statistically equivalent PDFs in a MC set



$$\mathrm{ERF} = \sum_k \frac{1}{N_k} \sum_i \left( \frac{C_i^{(k)} - O_i^{(k)}}{O_i^{(k)}} \right)^2 \qquad k = \text{CV, STD, SKE, KUR} \quad i = \text{sampling grid}$$

+ Kolmogorov-Smirnov test + PDF correlations

# The compression algorithm in action


xg(x,Q), comparison

xd(x,Q), comparison

xg(x,Q), comparison

xd(x,Q), comparison

# Monte Carlo to Hessian [EPJC 75 (2015) 369]

1. Find a subset of Monte Carlo replicas $\{\eta_\alpha^{(i)}\}_{i=1,\dots,N_{\rm eig}} \subset \left\{f_\alpha^{(k)}\right\}$ such that

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

2. Sample the replicas at a discrete set of points and construct the covariance matrix

$$({\rm cov})^f_{ij,\alpha\beta} \equiv \frac{N_{\rm rep}}{N_{\rm rep}-1}\left(\left\langle f_\alpha^{(k)}(x_i)\cdot f_\beta^{(k)}(x_j)\right\rangle_{\rm rep} - \left\langle f_\alpha^{(k)}(x_i)\right\rangle_{\rm rep}\left\langle f_\beta^{(k)}(x_i)\right\rangle_{\rm rep}\right)$$

3. Determine the set of coefficients $\{a^{(k)}\}$ by singular value decomposition of

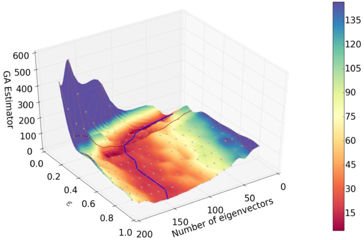
$$\chi_f^{2(k)} = \sum_{i,j}^{N_x}\sum_{\alpha,\beta}^{N_f}\left\{\left[f_{H,\alpha}^{(k)}(x_i) - f_\alpha^{(k)}(x_i)\right]\left({\rm cov}^f\right)^{-1}_{ij,\alpha\beta}\left[f_{H,\beta}^{(k)}(x_j) - f_\beta^{(k)}(x_j)\right]\right\}$$

4. Determine the Hessian matrix by inverting the covariance matrix

$${\rm cov}_{ij}^{(a)} = \frac{N_{\rm rep}}{N_{\rm rep}-1}\left(\left\langle a_i^{(k)} a_j^{(k)}\right\rangle_{\rm rep} - \left\langle a_i^{(k)}\right\rangle_{\rm rep}\left\langle a_j^{(k)}\right\rangle_{\rm rep}\right)$$

5. Optimise the number of grid points $N_x$ and of eigenvectors $N_{\rm eig}$

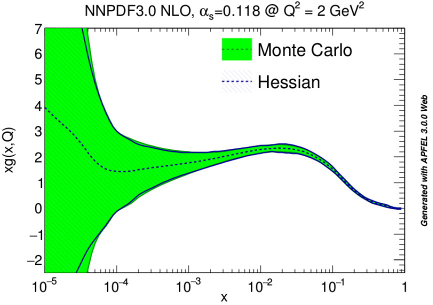
# Monte Carlo to Hessian: optimisation and validation

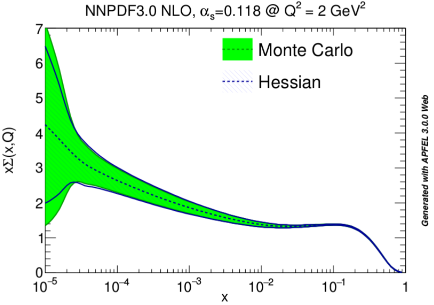


$$\epsilon_\alpha(x_i) = \frac{|\sigma_\alpha(x_i) - \sigma_\alpha^{68}(x_i)|}{\sigma_\alpha^{68}(x_i)} < \epsilon = 0.25$$

$$\mathrm{ERF}_\sigma = \sum_i^{N_x} \sum_\alpha^{N_f} \left| \frac{\sigma_{H,\alpha}^f(x_i) - \sigma_\alpha^f(x_i)}{\sigma_\alpha^f(x_i)} \right|$$

$$N_{\mathrm{rep}} = 1000 \qquad N_{\mathrm{eig}} = 120$$



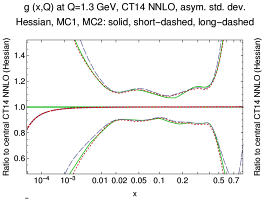


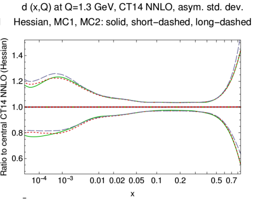
Similarly produce minimal PDF sets for specific processes [EPJ C76 (2016) 205]

# Hessian to Monte Carlo [JHEP 1208 (2012) 052; JHEP 1703 (2017) 099]

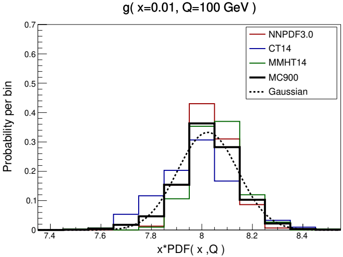
1. Generate multi-Gaussian replicas in the space of parameters

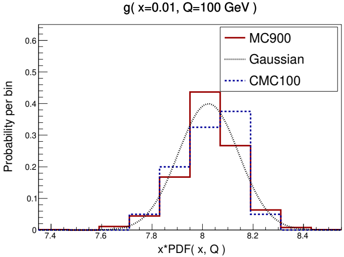
$$f^{(k)} = f_0 + d^{(k)} - \Delta \qquad d^{(k)} = \sum_{i=1}^{n_{\rm par}} \frac{1}{2}(f_{+i} - f_{-i})R_i^{(k)} \qquad \Delta = \frac{1}{N_{\rm rep}} \sum_{k=1}^{N_{\rm rep}} d^{(k)}$$


g (x,Q) at Q=1.3 GeV, CT14 NNLO, asym. std. dev.
Hessian, MC1, MC2: solid, short–dashed, long–dashed


d (x,Q) at Q=1.3 GeV, CT14 NNLO, asym. std. dev.
Hessian, MC1, MC2: solid, short–dashed, long–dashed

2. Combine different MC PDF sets (unweighted average)


g( x=0.01, Q=100 GeV )


g( x=0.01, Q=100 GeV )

# C. Conclusions

# Summary

1. Parton Distribution Functions with *faithful* uncertainties
   - tested procedure to assess data and methodological uncertainties in PDF fits
   - representation of data uncertainties: are NS or MCMC more convenient than MC?
   - characterisation of procedural uncertainties: closure tests, sometimes impractical
   - is there an alternative/more efficient way of characterising uncertainties?
   - representation of theoretical uncertainties: work in progress, input useful
2. Monte Carlo optimisation and delivery
   - performing a fit is computational expensive
   - Bayesian reweighting to update the probability density in the space of PDFs
   - Bayesian reweighting only useful when little data is added
   - are there alternative approaches to Bayesian reweighting?
   - a posteriori compression of MC replicas to obtain lighter ensembles
   - possibility to convert MC to Hessian and vv. to perform statistical combinations
   - are there other statistical methods to reduce the number of replicas?

# Summary

1. Parton Distribution Functions with *faithful* uncertainties
   - tested procedure to assess data and methodological uncertainties in PDF fits
   - representation of data uncertainties: are NS or MCMC more convenient than MC?
   - characterisation of procedural uncertainties: closure tests, sometimes impractical
   - is there an alternative/more efficient way of characterising uncertainties?
   - representation of theoretical uncertainties: work in progress, input useful

2. Monte Carlo optimisation and delivery
   - performing a fit is computational expensive
   - Bayesian reweighting to update the probability density in the space of PDFs
   - Bayesian reweighting only useful when little data is added
   - are there alternative approaches to Bayesian reweighting?
   - a posteriori compression of MC replicas to obtain lighter ensembles
   - possibility to convert MC to Hessian and vv. to perform statistical combinations
   - are there other statistical methods to reduce the number of replicas?

**Thank you**

# Normalisation uncertainties: D'Agostini bias [JHEP 1005 (2010) 075]

1. Consider one experiment with $N_{\rm dat}$ data $d_i$ of one theoretical quantity $t$

$$\chi^2(t) = \sum_{i,j}^{N_{\rm dat}} (t - d_i) \left(\mathrm{cov}^{-1}\right)_{ij} (t - d_j)$$

2. The best-fit theoretical quantity $t_0$ and its variance $v_t$ are given by

$$\left.\frac{d\chi^2}{dt}\right|_{t=t_0} = 0 \Longleftrightarrow t_0 = \frac{\sum_{i,j}^{N_{\rm dat}} \left(\mathrm{cov}^{-1}\right)_{ij} d_j}{\sum_{i,j}^{N_{\rm dat}} \left(\mathrm{cov}^{-1}\right)_{ij}} \qquad v_t = \left(\frac{1}{2}\frac{d^2\chi^2}{dt^2}\right)^{-1} = \frac{1}{\sum_{i,j}^{N_{\rm dat}} \left(\mathrm{cov}^{-1}\right)_{ij}}$$

3. Consider completely uncorrelated additive errors: $(\mathrm{cov})_{ij} = s_i^2 \delta_{ij}$

$$t_0 = w = \Sigma^2 \sum_i^{N_{\rm dat}} \frac{d_i}{s_i^2} \qquad v_t = \Sigma^2 \qquad \text{with } \frac{1}{\Sigma^2} = \sum_i^{N_{\rm dat}} \frac{1}{s_i^2}$$

4. Consider an additional common normalisation error: $(\mathrm{cov})_{ij} = (s_i^2 + \sigma^2 d_i^2)\delta_{ij}$

$$t_0 = \frac{w}{1 + r^2\sigma^2 w^2/\Sigma^2} \qquad v_t = \frac{\Sigma^2 + \sigma^2 w^2 (1 + r^2)}{1 + r^2\sigma^2 w^2/\Sigma^2} \qquad \text{with } r^2 = \frac{\Sigma^2}{w^2} \sum_i^{N_{\rm dat}} \frac{(d_i - w)^2}{s_i^2}$$

5. Both $t_0$ and $v_t$ are affected by a downward bias
   smaller values of $d_i$ have a smaller normalization uncertainties $\sigma d_i$ and are thus preferred

# Normalisation uncertainties: D'Agostini bias [JHEP 1005 (2010) 075]

1. The penalty trick: redefine the fit quality

$$\chi^2(t) \to \chi^2(t,\mathcal{N}) = \sum_i^{N_{\rm dat}} \frac{(t/\mathcal{N}-d_i)^2}{s_i^2} + \frac{(\mathcal{N}-1)^2}{\sigma^2}$$

$$\left.\frac{\partial\chi^2}{\partial t}\right|_{t=t_0} = \frac{\partial\chi^2}{\partial\mathcal{N}} = 0 \Longleftrightarrow t_0 = w \qquad v_t = \left(\frac{1}{2}\frac{d^2\chi^2}{dt^2}\right)^{-1} = \Sigma^2 + \sigma^2 w^2$$

$\longrightarrow$ recover the unbiased estimators for $t_0$ and $v_t$

2. The $t_0$ method: redefine the covariance matrix

$$({\rm cov})_{ij} \to ({\rm cov}_{t_0})_{ij} \Longleftrightarrow (s_i^2+\sigma^2 d_i^2)\delta_{ij} \to s_i^2\delta_{ij} + \sigma^2 t_0^2$$

$$\left({\rm cov}_{t_0}^{-1}\right)_{ij} = \frac{\delta_{ij}}{s_i^2} - \frac{\sigma^2 t_0^2}{s_i^2 s_j^2}\frac{\Sigma^2}{\Sigma^2+\sigma^2 t_0^2} \Longleftrightarrow t_0 = w \qquad v_t = \Sigma^2 + \sigma^2 w^2$$

$\longrightarrow$ recover the unbiased estimators for $t_0$ and $v_t$, provided that $w$ is tuned to $t_0$
$\longrightarrow$ $w$ can be tuned to $t_0$ via an iterative procedure

The d'Agostini bias and its solution can be generalised to more than one experiment with different normalisation errors (per experiment/per data point)