

Extracting the xFitter Likelihood

Ali Al Kadhim & Harrison B. Prosper



Outline

- Introduction
- Procedure
- Results
- Available Code
- xFitter Wishlist
- Summary

Introduction

- Characterizing PDF uncertainties is important – they directly affect our inferences from data.
- This will become increasingly important as we move to Run 3 and the HL-LHC era.
- It is well-known that if a small number of datasets are used, one can safely apply standard statistical procedures to estimate confidence intervals ($\Delta\chi^2 = 1$).
- However, when the fit includes a large number of datasets, tolerance factors (T) are used to arrive at uncertainties that are deemed to be meaningful ($T = \sqrt{\Delta\chi^2}$).

Introduction

- Our goal is to extract the xFitter likelihood for increasing numbers of datasets and study the tolerance factors.
- Ideally, every experimental result would be published along with its statistical model $P(x|\theta)$, see K. Cranmer, S. Kraml, H.B. Prosper et. al. [1], and PDF fits would be performed using the sum of the associated negative log-likelihoods.
- However, all PDF fits are performed by minimizing a χ^2 function.
- In our studies, we assume that the xFitter likelihood function, $L(\theta) \equiv P(D|\theta)$, is given by

$$-2 \log L(\theta) = \chi^2$$

- [1] K. Cranmer, S. Kraml, H.B. Prosper et. al. "Publishing statistical models: Getting the most out of particle physics experiments" <https://scipost.org/SciPostPhys.12.1.037/pdf>

Procedure

- We sample the PDF parameters θ from a prior $\pi(\theta)$, whose support, ideally, roughly matches that of the likelihood $L(\theta)$.
- In order to approximate the likelihood, we weight each point, i , by

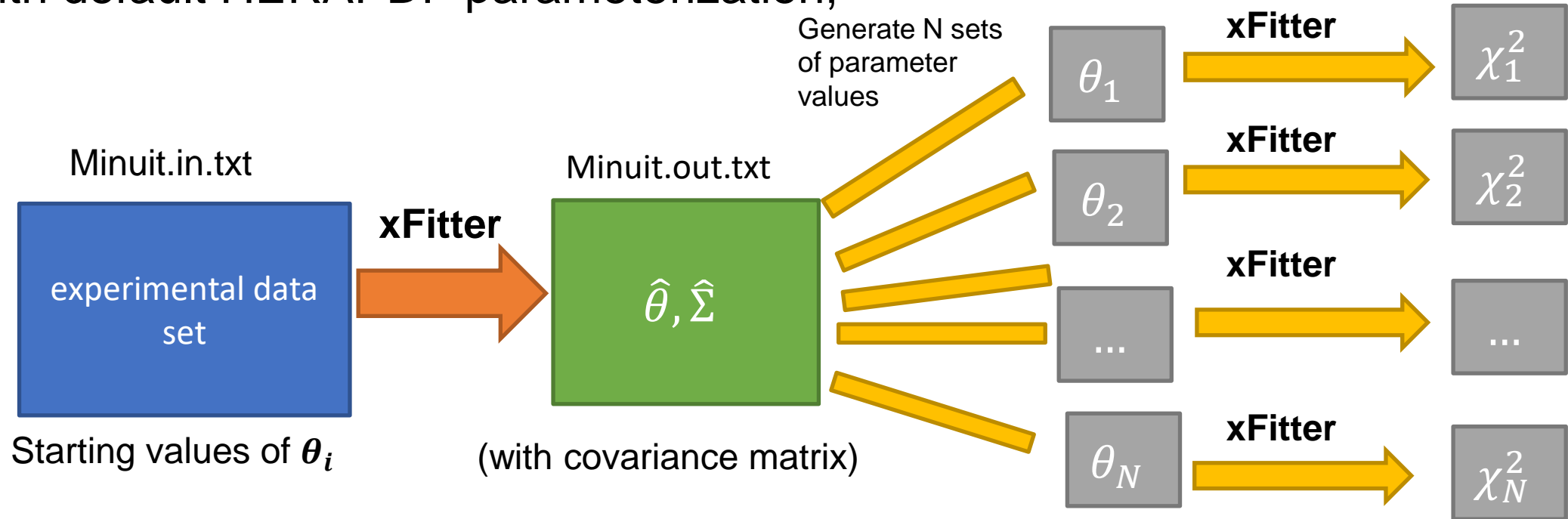
$$w_i = \frac{L(\theta_i)}{\pi(\theta_i)}$$

- We anticipate that as more datasets are added and more discrepancies appear between them, the true width W_L of the 68% intervals computed using the posterior density, derived from the likelihood, will satisfy

$$\frac{W_{L(\theta)}}{W_{\Delta\chi^2=1}} \approx T$$

Procedure

- With default HERAPDF parameterization,



- We generate N sets of parameter values according to $\theta_i \sim \pi(\theta)$.

Procedure: Prior $\pi(\theta)$

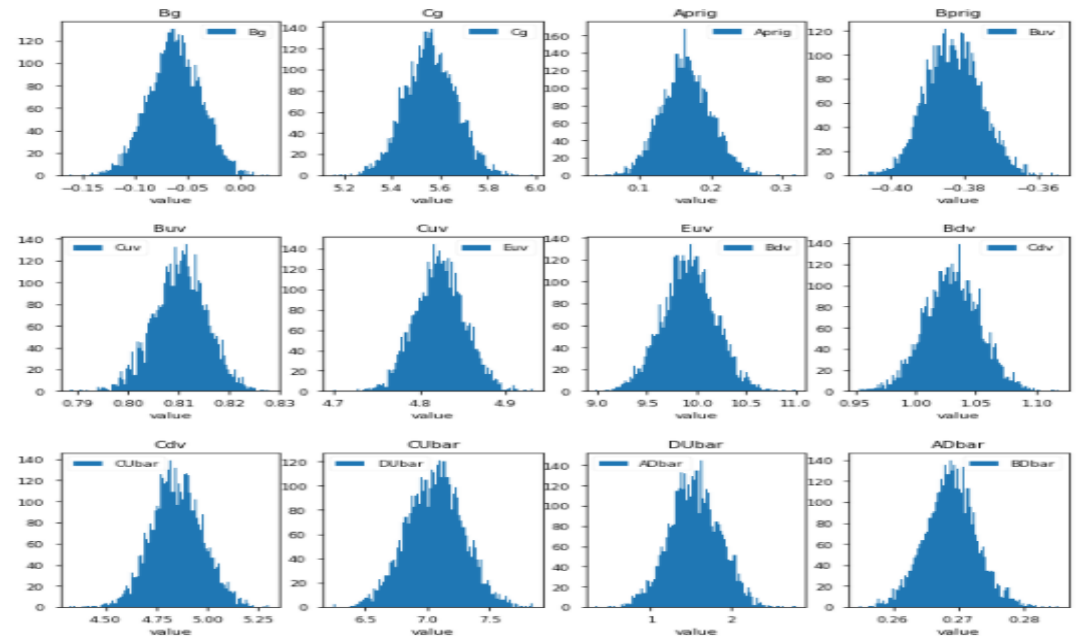
- In our current studies we take $\pi(\theta) = \mathcal{N}(\mu = \hat{\theta}, \Sigma = \hat{\Sigma})$.

- HERAPDF parameterization:

$$xf(x, \mu_0^2) = Ax^B(1-x)^C(1+Dx+Ex^2) - A'x^{B'}(1-x)^{C'}$$

- Data: HERA I & II + ZEUS combined

Parameter	xFitter Name	Starting Value	Step Size	Best-Fit Value	Approximate Error
B_g	Bg	-0.061953	0.027133	-00.61856	0.25134E-01
C_g	Cg	5.562367	0.318464	5.5593	0.10838
A'_g	Aprig	0.166118	0.028009	0.16618	0.34574E-01
B'_g	Bprig	-0.383100	0.009784	-0.38300	0.76253E-02
B_{uv}	Buv	0.810476	0.016017	0.81056	0.53604E-02
C_{uv}	Cuv	4.823512	0.063844	4.8239	0.29342E-01
E_{uv}	Euv	9.921366	0.835891	9.9226	0.27481
B_{dv}	Bdv	1.029995	0.061123	1.0301	0.23240E-01
C_{dv}	Cdv	4.846279	0.295439	4.8456	0.12584
$C_{\bar{U}}$	CUbar	7.059694	0.809144	7.0603	0.22306
$D_{\bar{U}}$	DUbar	1.548098	1.096540	1.5439	0.31340
$A_{\bar{D}}$	ADbar	0.268798	0.008020	0.26877	0.39536E-02
$B_{\bar{D}}$	BDbar	-0.127297	0.003628	-0.12732	0.17428E-02
$C_{\bar{D}}$	CDbar	9.586246	1.448861	9.5810	0.60834



Procedure: Reweighting

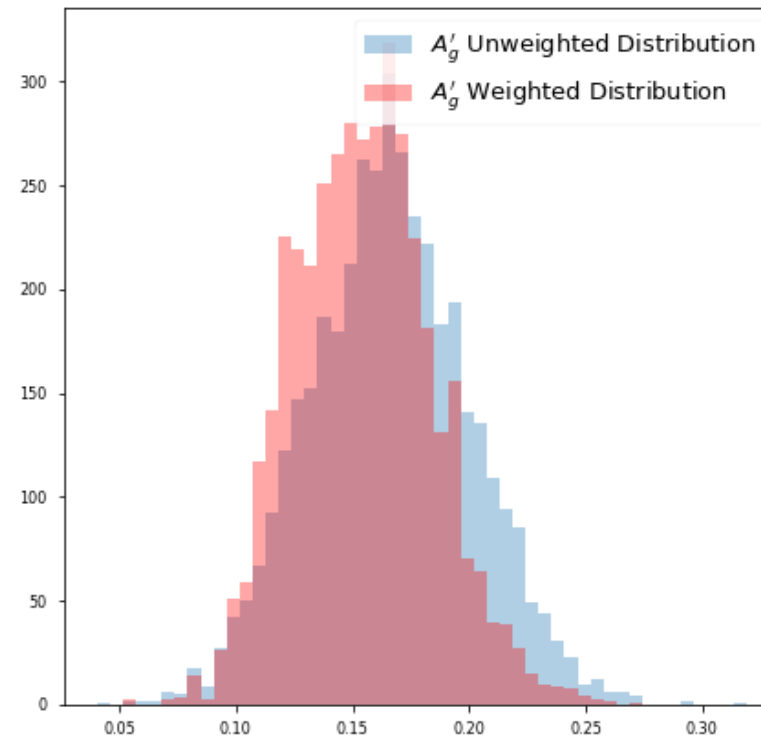
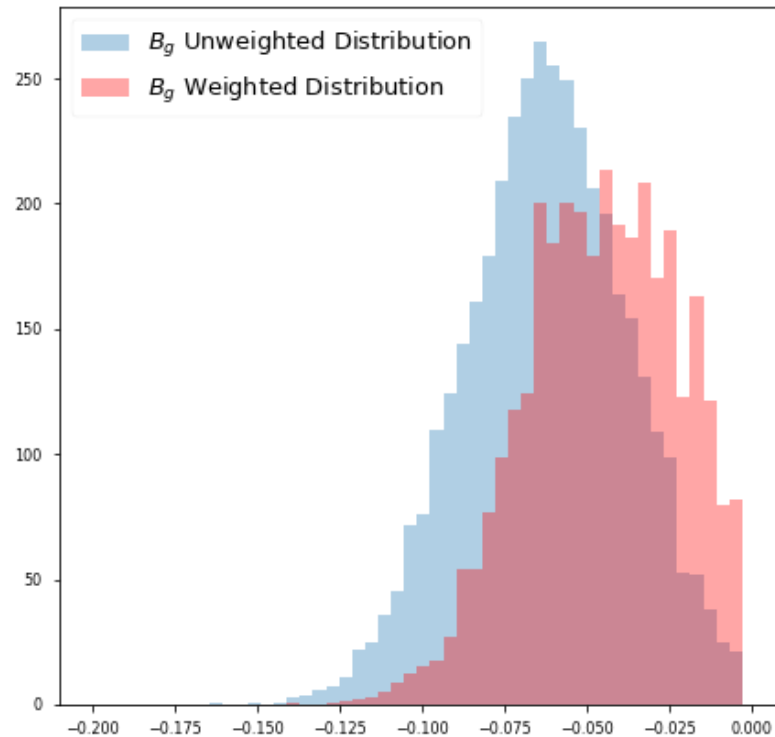
- As noted, to approximate the likelihood, we weight each point by

$$w_i = \frac{L(\theta_i)}{\pi(\theta_i)}$$

- In the limit of infinite number of sampled points, the weighted distribution will recover the true likelihood.
- The weighted distributions are what we expect to arrive at if we could do a Markov chain sampling of $L(\theta)$.

Results - 1

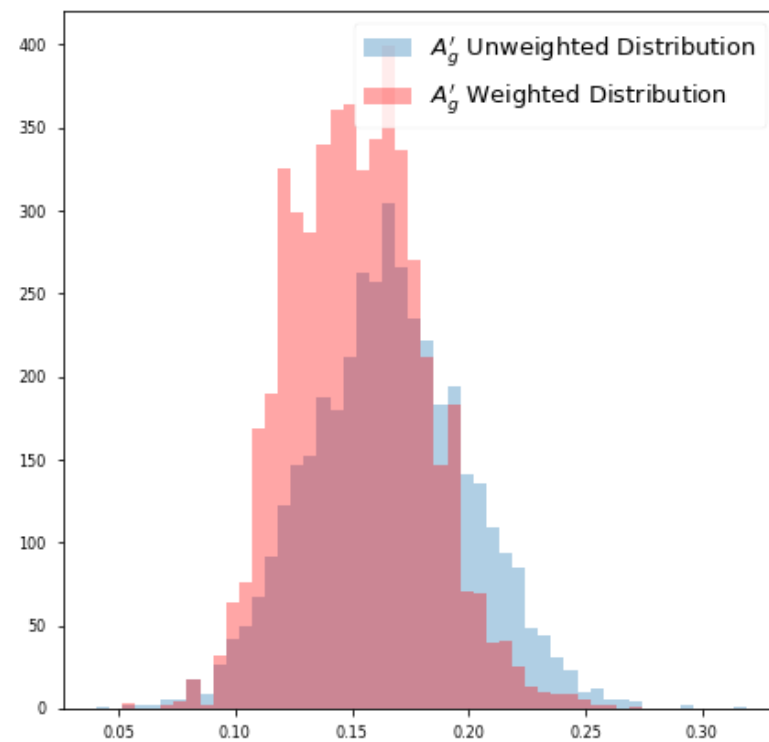
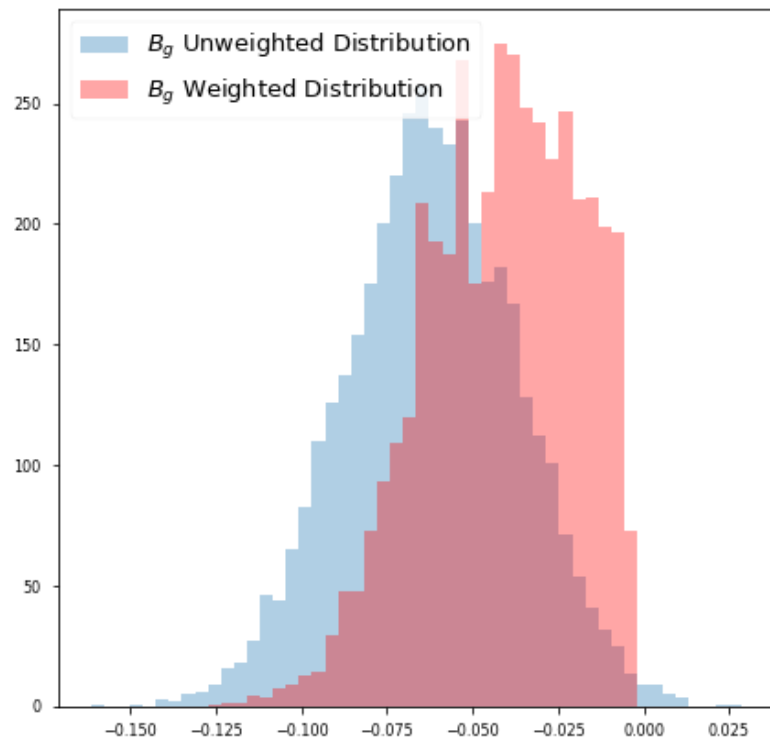
- Data: HERA I, II & ZEUS



4,000 parameter
sets
HERAPDF
parameterization

Results - 2

- Data: HERA I+II & ZEUS + CDF W asymmetry + D0 Run II cone jets



4,000 data points
(parameter sets)
HERAPDF
parameterization

Available Code

- We have done many studies with xFitter (focusing on master version). All our code is available at:
https://github.com/AliAlkadhim/PDF_Uncertainty/
- We also made a docker image with full xfitter-master and its dependencies and all datasets and our code installed

```
docker run -it alialkadhim/pdf_uncertainty:v0
```
- Most recent studies with xFitter-master can be found in
https://github.com/AliAlkadhim/PDF_Uncertainty/master_version/local

xFitter Wishlist

- Initialize all the datasets (and packages) only once.
- Once the initialization is done, be able to feed xFitter a sequence of parameter points.
- For each parameter point, xFitter computes and returns a χ^2 value.
- In the future, xFitter would return $-2 \log L(\theta)$.

Summary

- The weighting procedure for recovering the true likelihood seems to work, although it would be good to try different priors.
- However, in this study, we rerun xFitter for every parameter set, which is inefficient.
- We found xFitter to be an excellent tool for performing such studies, and with the changes we propose, it would make studies easier.
- We would like to thank the members of the xFitter team and Simone for inviting us to make this presentation.

Backup

Technical Aspects & Suggestions

The more datasets used, the more programs that have to be initialized \Rightarrow More overhead time! e.g. a parameters.yml could look like:

- **Decompositions: Proton**
 - **DefaultEvolution: proton-QCDNUM**
- **Evolutions:**
 - **proton-APFELff**; include evolutions/APFEL.yaml
 - **proton-QCDNUM**; include evolutions/QCDNUM.yaml
 - ...
- **Decomposition:**
 - **Proton**
 - **Antiproton**; class: FlipCharge
 - **Neutron**; class: FlipUD
 - **proton-LHAPDF**; class: LHAPDF ; set: "NNPDF30_nlo_as_0118"

Ali Al Kadhim - xFitter Workshop 2022

xFitter-master Datasets used
(all available but 1)

HERA I+II combined inclusive DIS

CDF Jets, W, Z production

D0 Jets, W, Z production

CMS W, Z production

CMS Jets

ATLAS W, Z production

ATLAS Drell-Yann

ATLAS Jets

ATLAS Dec 2016 W,Z

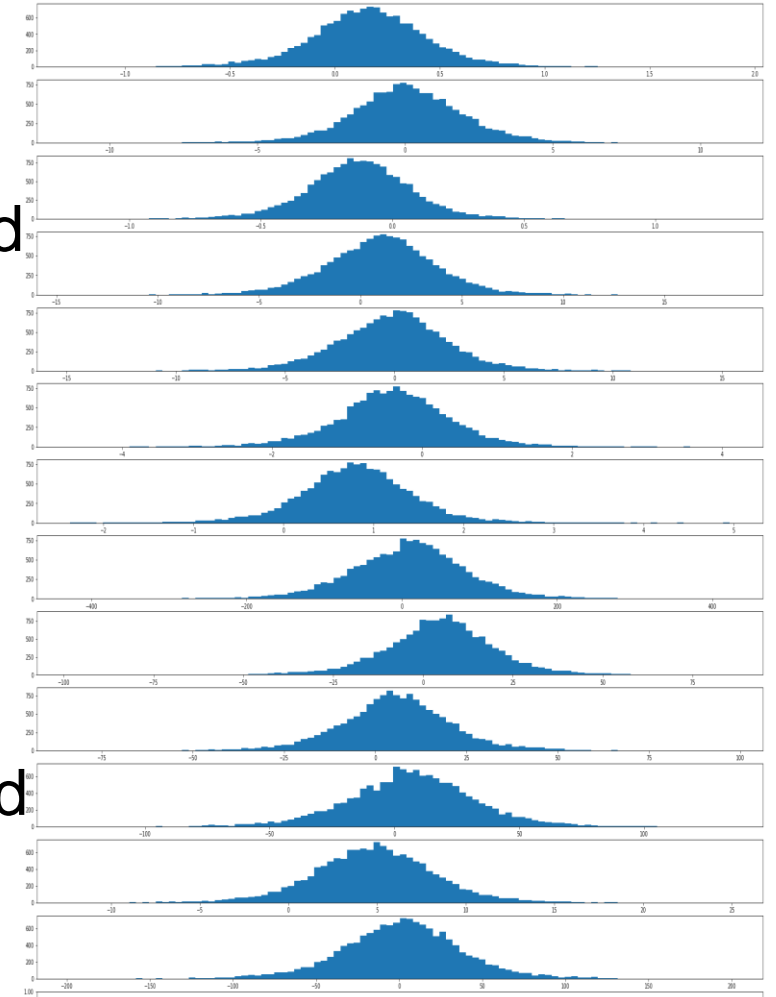
LHCb charm and beauty

CMS W+c

CMS 8 TeV jets

MCMC using all sets, Different $L'(\theta)$

- MCMC is not efficient since it cannot be parallelized and computationally expensive.
- However, it should return the true shape of the likelihoods.
- Using 14336 points (many weeks of running), MCMC (Metropolis-Hastings) still yields Gaussian parameter likelihoods using the xFitter χ^2 values.
- Reason: Not enough differing (discrepant) datasets (and not enough statistics, i.e. data points).



More Backup

- If we approximate $\pi(\boldsymbol{\theta}) = \mathcal{N}(\mathbf{D}|\hat{\boldsymbol{\theta}}_i, \hat{\boldsymbol{\Sigma}}_i)$, then

$$w_k = \frac{L(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})} = \frac{N_{samples}}{\sum_{k=1}^{N_{samples}} w_k} \frac{e^{-\frac{1}{2}\chi_k^2}}{\mathcal{N}(\mathbf{D}|\hat{\boldsymbol{\theta}}_k, \hat{\boldsymbol{\Sigma}}_k)} = \begin{cases} \mathbf{1}, & \text{Gauss. Approx. holds for } L(\boldsymbol{\theta}) \\ \text{else,} & L(\boldsymbol{\theta}) \text{ is non - Gauss.} \end{cases}$$

- If the likelihood for θ is multivariate normal, the likelihood of a single observation is of the form

$$\bullet L(\boldsymbol{\theta}|\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^D|\boldsymbol{\Sigma}|}} \exp\left\{-\frac{1}{2}[\mathbf{x} - g(\boldsymbol{\theta})]^T \boldsymbol{\Sigma}^{-1}[\mathbf{x} - g(\boldsymbol{\theta})]\right\} = \frac{1}{\sqrt{(2\pi)^D|\boldsymbol{\Sigma}|}} \exp\left\{-\frac{1}{2}\chi^2\right\} \longrightarrow \log L(\boldsymbol{\theta}) = -\frac{1}{2}\chi^2$$

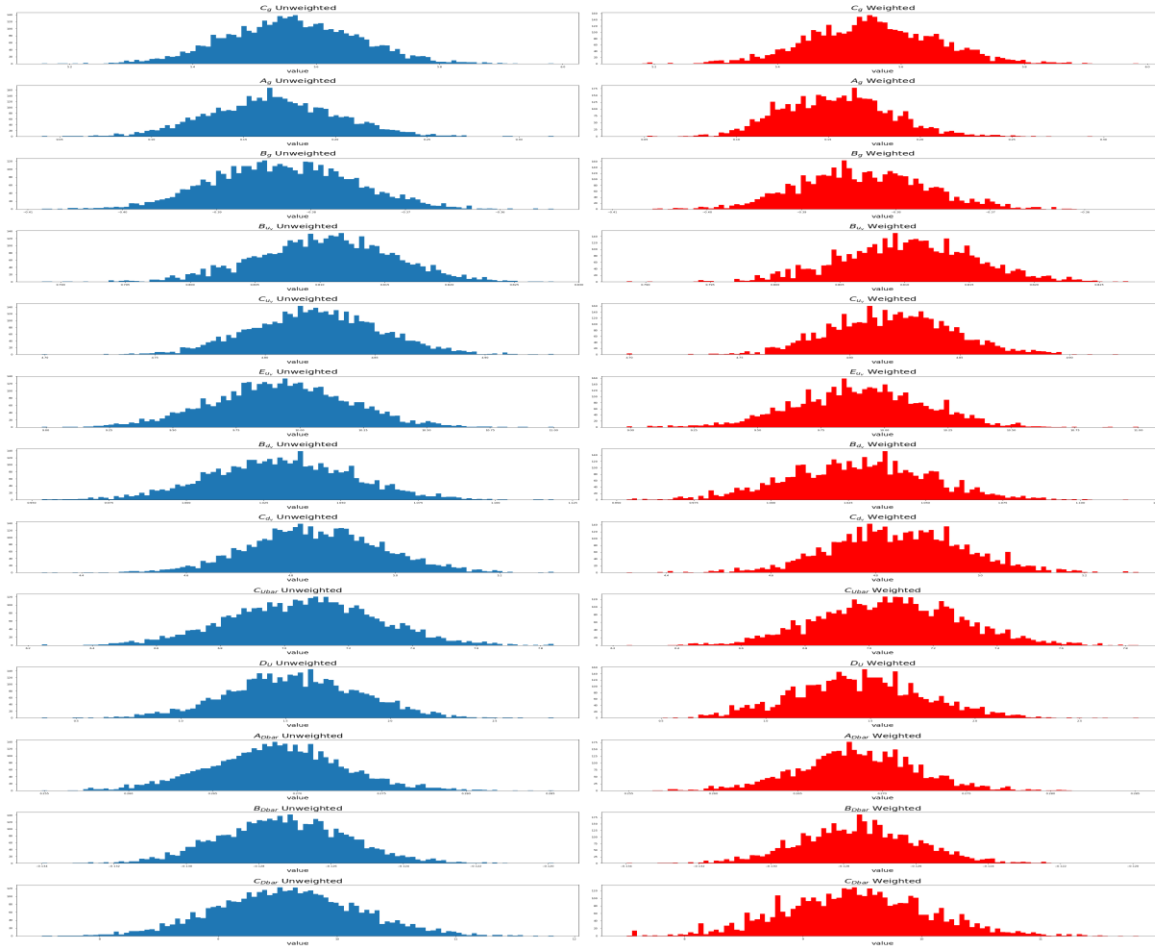
- 68% confidence intervals are obtained by finding points where $\Delta\chi^2 = 1$, i.e.
- $-2\Delta \log[L] = -2[\log[L(\boldsymbol{\theta}_{\pm}|\mathbf{x})] - \log[L(\hat{\boldsymbol{\theta}}|\mathbf{x})]] = 1 \longrightarrow (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_-, \hat{\boldsymbol{\theta}} + \boldsymbol{\theta}_+)$ but this assumes normal sampling of data.

- The tolerance $T = \sqrt{\Delta\chi_{global}^2}$, ideally $T = 1$, but this assumes ideal gaussian errors & well-defined theory.

- In global fits, $T > 1$ to account for discrepant data sets (e.g. see arxiv: 1410.8849).

All parameter Distributions

One Dataset



Multiple Datasets

