Mixture Models for Uncertainty Quantification in PDFs

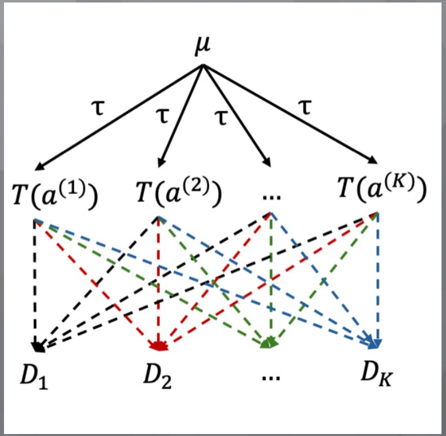
Kirtimaan Mohan – Michigan State University

with

Mengshi Yan, Tie-Jiun Hou, Zhao Li & C.-P. Yuan

arxiv: 2406.01664

@PDFLattice2024 - JLab



MICHIGAN STATE UNIVERSITY

Motivation

- Precision measurements need precise PDFs
- PDF fitting groups have to contend with tension in data
  - Many strategies to deal with this: For example, the use of tolerance $(\Delta\chi^2 = T^2)$
- This talk will describe the use of Gaussian Mixture Model (GMM) and how it can be used to
  - find inconsistencies or tension in data sets
  - Implement Bayesian Model Averaging (BMA) in order to determine uncertainties in a statistically robust way.

What is the **Gaussian Mixture Model**?

- Widely used an unsupervised machine learning technique
  - Could be used to classify PDF data
- Class of Finite Mixture Models
  - https://doi.org/10.1146/annurev-statistics-031017-100325
- Widely used in astronomy and astrophysics to distinguish between different sources in the sky
- First proposed by Karl Pearson (1894) – to study characteristics of a population of crabs
- **Focus of this talk: How can this machine learning technique be used to implement Bayesian Model Averaging for uncertainties in PDFs?**

Outline

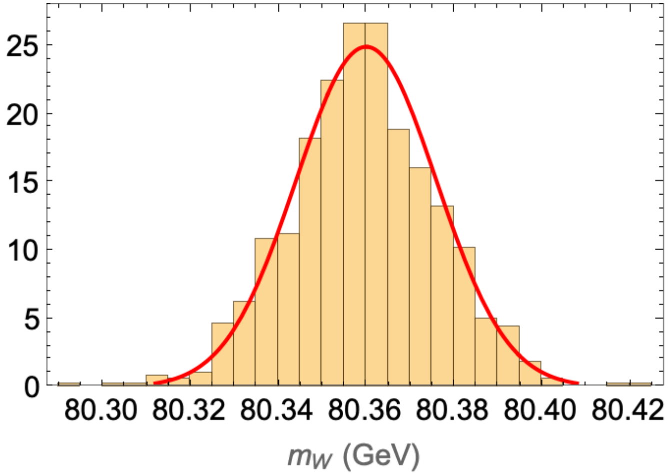
- Motivation for GMM use in PDFs ✓
- Description of Gaussian Mixture Model(GMM) in a simple 1-D example
- Formalism: Bayesian Model Averaging
- Demonstrate idea with a toy model of PDFs
- Summary

Simple 1-D example

Measuring Mass (Weight) PHY-101 Lab

- Measure mass of W-boson
- Repeat measurement several times
- Minimize log-likelihood or loss function
  - $\chi^2 = \sum_i \frac{(\mu - x_i)^2}{\sigma_i^2}$
  - $L = \prod_i \frac{e^{\left[\frac{(\mu - x_i)^2}{\sigma_i^2}\right]}}{\sqrt{2\pi}\sigma_i}$
- Determine best-fit value
  - $m_W = \mu = 80.36 \pm 0.016\ GeV$

[ATLAS-CONF-2023-004]



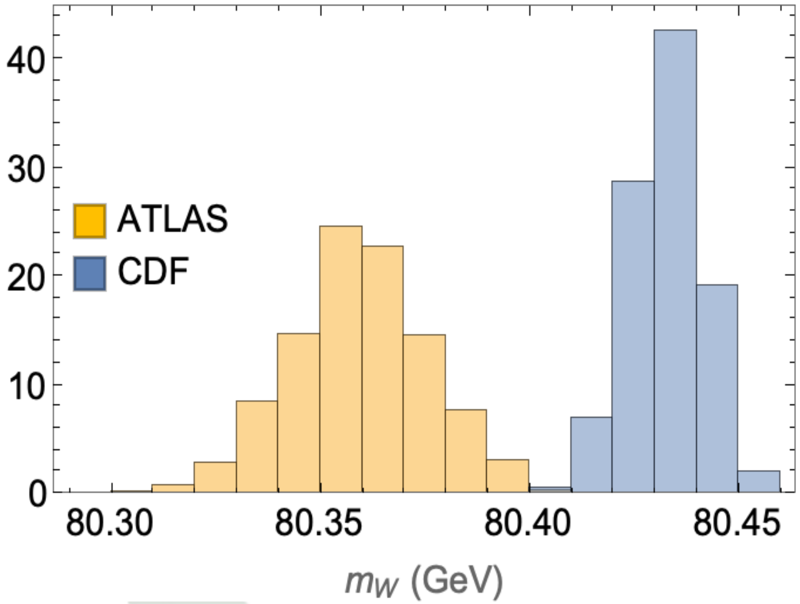
Manufactured by ATLAS

Measuring Mass (Weight) PHY-101 Lab

Repeat measurements with another balance

CDF *Science 376* (2022)

Manufactured by CDF

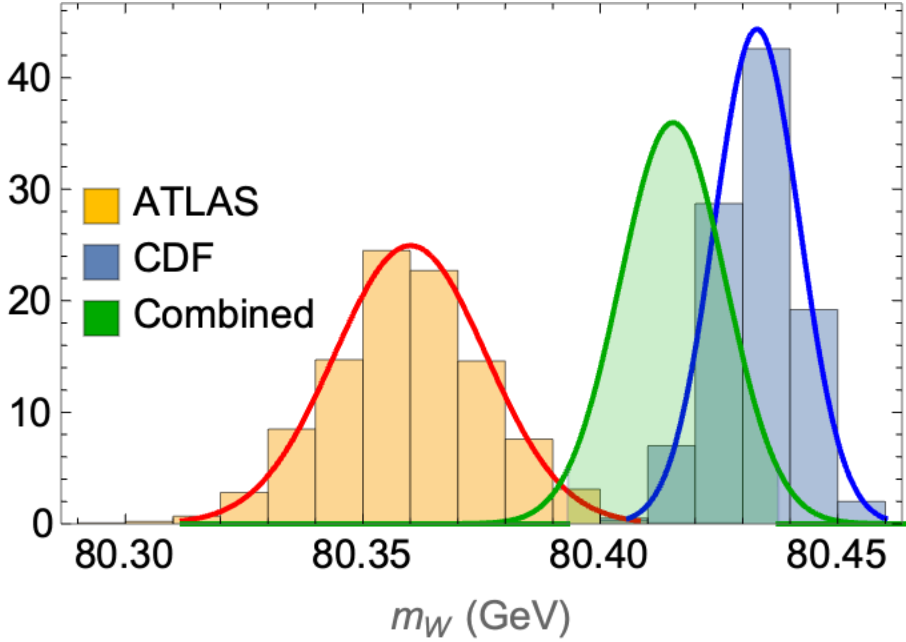


$$m_W^{CDF} = 80.433 \pm 0.009\ GeV$$
$$m_W^{ATLAS} = 80.36 \pm 0.016\ GeV$$

Manufactured by ATLAS

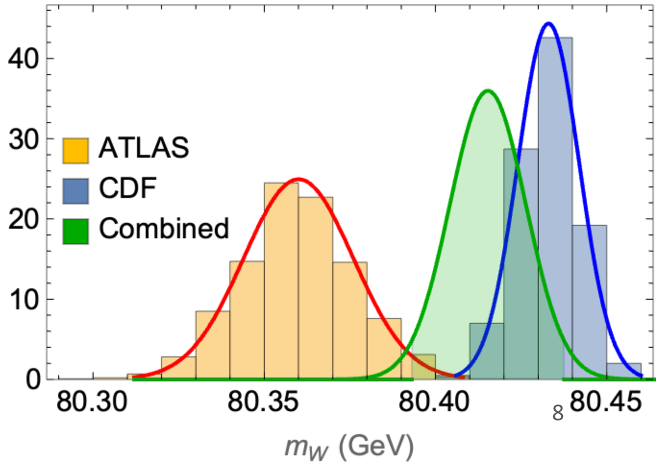
Measuring Mass (Weight) PHY-101 Lab

- How should we combine these two discrepant measurements to give one value of mass?
- **Attempt #1**: Let's repeat earlier exercise
  - Minimize loss function
    - $\chi^2 = \sum_i \frac{(\mu - x_i)^2}{\sigma_i^2}$
    - $m_W = 80.415 \pm 0.011\ GeV$
- 2σ band does not cover both means
  - How should we interpret this?
- One familiar proposal
  - Increase tolerance $\Delta\chi^2 = T^2; T > 1$
  - Does not provide a faithful representation of the probability distribution of m_W, drawn from our sample of experiments



Shortcomings of χ^2 fits

- Why didn't our usual approach reproduce the probability distribution function for m_W work?
- In this simple example
  - We ignored individual likelihoods from each experiment
  - We minimized the χ^2 which is
    - Just like taking the weighted mean
    - And adding errors in quadrature
    - Then defining a new gaussian likelihood (green)
    - Starting assumption is that m_W likelihood is a single gaussian
    - Good assumption **if** data is consistent
- **Attempt #2:** Combine likelihoods



MICHIGAN STATE UNIVERSITY

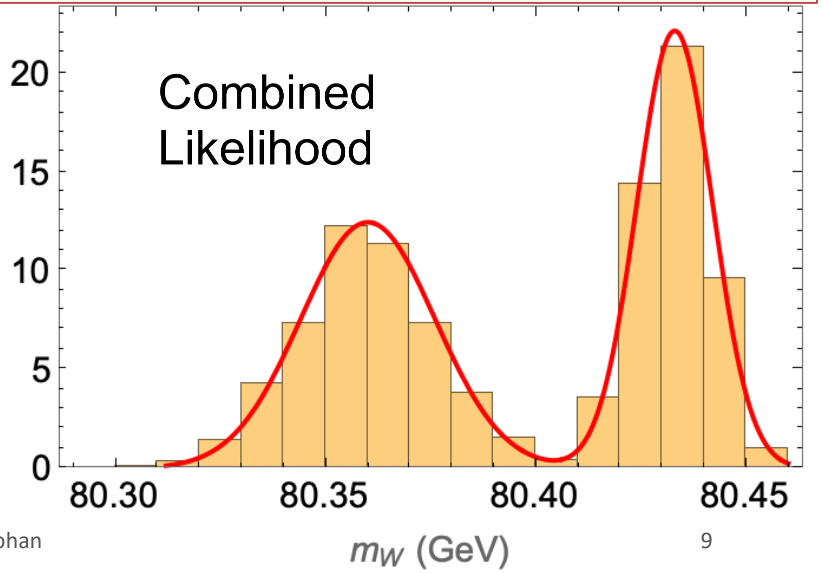
Combining Likelihoods – Gaussian Mixture Model

$$\mathcal{N} = \frac{e^{\left[\frac{(\mu - x_i)^2}{\sigma_i^2}\right]}}{\sqrt{2\pi}\sigma_i}$$

- Start by parameterizing the likelihood as a sum of Gaussians
- In this simple example we know there are two Gaussians, i.e. K= 2
- In general, the value of K needs to be determined – discussed later
- Introduced a new parameter ω_k - weights
- Constraints on ω_k; ensures proper normalization and interpretation as a probability distribution function
- For simplicity we'll use equal weights here
- In reality – it is an additional fit parameter
- See Interpretation in Bayesian formalism later.

$$\pi(Y|\vec{\theta}) = \prod_{j=1}^{N_{\mathrm{pt}}} \pi(y_j, \Delta y_j|\vec{\theta}) = \prod_{j=1}^{N_{\mathrm{pt}}} \sum_{i=1}^{K} \omega_i \mathcal{N}(y_j, \Delta y_j|\theta_i),$$

$$0 \leq \omega_k \leq 1 \quad \text{and} \quad \sum_k \omega_k = 1,$$



Determine mean and variance for GMM

$$\pi(Y|\vec{\theta}) = \prod_{j=1}^{N_{\rm pt}} \pi(y_j, \Delta y_j|\vec{\theta}) = \prod_{j=1}^{N_{\rm pt}} \sum_{i=1}^{K} \omega_i \mathcal{N}(y_j, \Delta y_j|\theta_i),$$

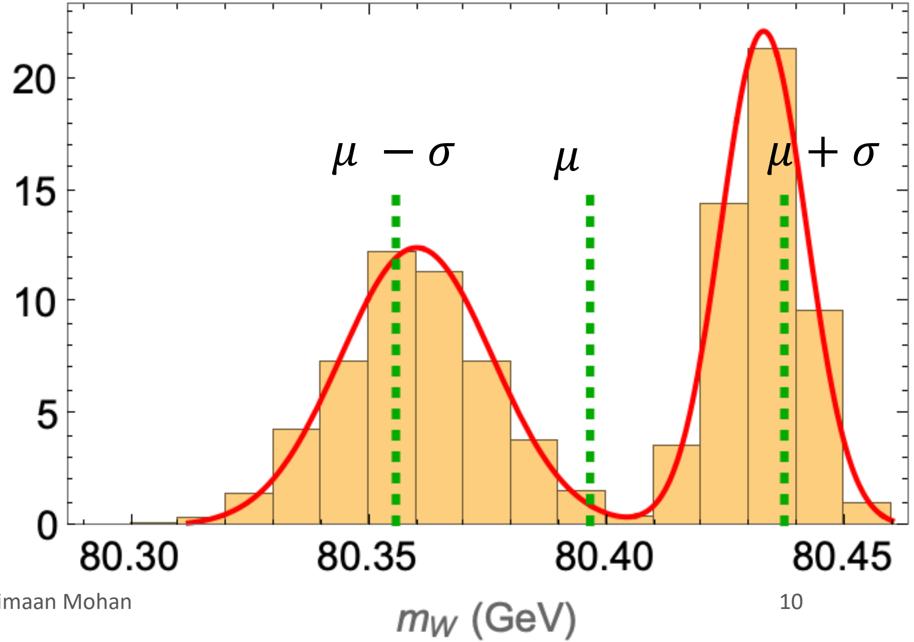
$$0 \leq \omega_k \leq 1 \quad \text{and} \quad \sum_k \omega_k = 1,$$

Mean $\mathbb{E}[\theta] = \sum_{i=1}^{K} \omega_i \hat{\theta}_i.$

$$\begin{aligned} \text{cov}_{\rm GMM} &= \sum_{i=1}^{K} \omega_i \, \text{cov}_{{\rm GMM},i} + \sum_{i=1}^{K} \omega_i (\mathbb{E}[\theta] - \hat{\theta}_i)^2 \\ &= \sum_{i=1}^{K} \omega_i \left(\sum_{j=1}^{N_{\rm pt}} \frac{1}{\Delta y_j^2} \left(\frac{\partial y_j(\theta_i)}{\partial \theta_i} \right)^2 \frac{\mathcal{N}(y_j, \Delta y_j|\theta_i)}{\pi(y_j, \Delta y_j|\vec{\theta})} \right)^{-1} + \sum_{i=1}^{K} \omega_i (\mathbb{E}[\theta] - \hat{\theta}_i)^2. \end{aligned}$$

Weighted sum of covariances of each Gaussian

Difference between Gaussians



Here we use the variance as an estimator for the standard error.
Alternatively, we could use the Observed Fisher Information Matrix

Determine mean and variance for GMM

$$\pi(Y|\vec{\theta}) = \prod_{j=1}^{N_{\rm pt}} \pi(y_j, \Delta y_j|\vec{\theta}) = \prod_{j=1}^{N_{\rm pt}} \sum_{i=1}^{K} \omega_i \mathcal{N}(y_j, \Delta y_j|\theta_i),$$

$$0 \le \omega_k \le 1 \quad \text{and} \quad \sum_k \omega_k = 1,$$

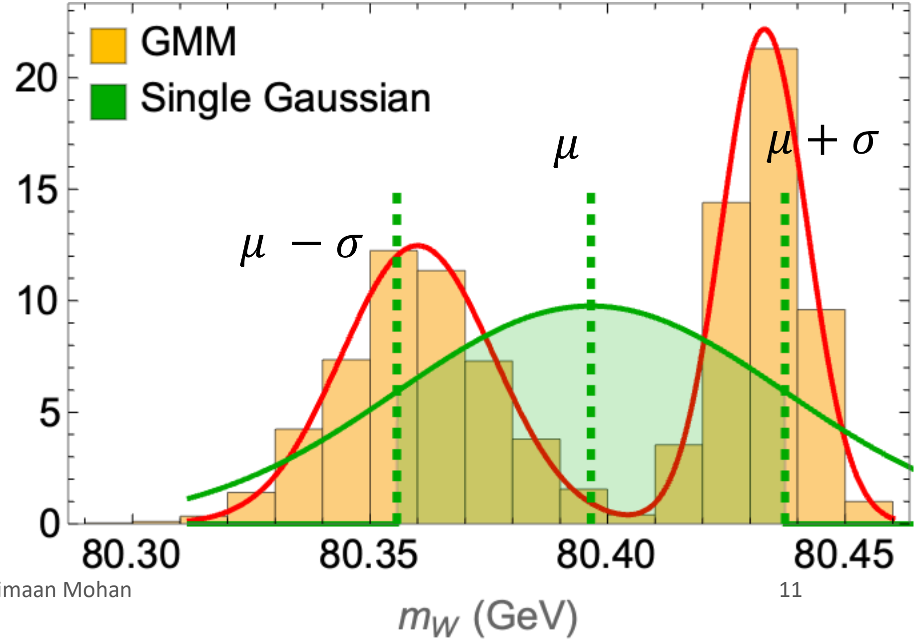
Mean $$\mathbb{E}[\theta] = \sum_{i=1}^{K} \omega_i \hat{\theta}_i.$$

$$\text{cov}_{\rm GMM} = \sum_{i=1}^{K} \omega_i \, \text{cov}_{{\rm GMM},i} + \sum_{i=1}^{K} \omega_i (\mathbb{E}[\theta] - \hat{\theta}_i)^2$$

$$= \sum_{i=1}^{K} \omega_i \left(\sum_{j=1}^{N_{\rm pt}} \frac{1}{\Delta y_j^2} \left(\frac{\partial y_j(\theta_i)}{\partial \theta_i} \right)^2 \frac{\mathcal{N}(y_j, \Delta y_j|\theta_i)}{\pi(y_j, \Delta y_j|\vec{\theta})} \right)^{-1} + \sum_{i=1}^{K} \omega_i (\mathbb{E}[\theta] - \hat{\theta}_i)^2.$$

Weighted sum of covariances of each Gaussian

Difference between Gaussians



Caveat about green curve: because we are used to it, it is possible to model this as a single Gaussian (green) – but we must be careful - it is **not** a faithful representation of the likelihood.

Formalism

Bayesian Model Averaging

Review of Bayesian Formalism for χ^2

MICHIGAN STATE UNIVERSITY

Data $$D_i = \langle D_i \rangle + \sigma_i \Delta_i \ .$$

$$\langle f \rangle = (2\pi)^{N_D/2} \int f(\Delta) \prod_{i=1}^{N_D} d\Delta_i \exp\left(-\frac{1}{2}\Delta_i^2\right)$$

$$\langle g \rangle = \frac{1}{\sqrt{(2\pi)^{N_D} \det C}} \int g(D) \prod_{i,j=1}^{N_D} dD_i \exp\left(-\frac{1}{2}(D_i - \langle D_i \rangle)(D_j - \langle D_j \rangle) C_{ij}^{-1}\right)$$

$$P(D|T(a)) = \frac{1}{\sqrt{(2\pi)^{N_D} \det C}} dD \exp\left(-\frac{1}{2}\sum_{i,j=1}^{N_D} (D_i - T_i(a))(D_j - T_j(a)) C_{ij}^{-1}\right)$$

$$P(T(a)|D) = \frac{P(D|T(a))P(T(a))}{P(D)}$$

See Kovarık, Nadolsky & Soper arXiv:1905.06957

Bayesian Model Averaging and GMMs

Data from K different experiments $D_i^{(k)} = \langle D_i^{(k)} \rangle + \sigma_i^{(k)} \Delta_i^{(k)} = T_i(a^{(k)}) + \sigma_i^{(k)} \Delta_i^{(k)}$

$$P(T(a^{(k)})) = \int d\mu d\tau P(T(a^{(k)})|\mu,\tau)p(\mu,\tau) \equiv w_k \qquad \sum_{k=1}^{K} w_k = 1.$$

Bayes' Theorem

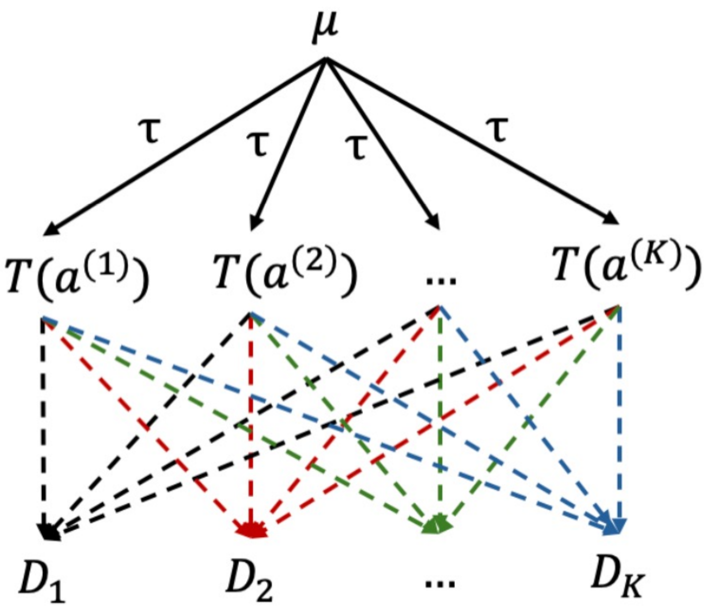
$$P(D_i|T(a^{(k)}))P(T(a^{(k)})) = w_k P(D_i|T(a^{(k)})) = P(T(a^{(k)})|D_i)P(D_i)$$

$$\prod_{i=1}^{N_D}\left(\sum_{k=1}^{K} P(T(a^{(k)})|D_i)\right) \propto \prod_{i=1}^{N_D}\left(\sum_{k=1}^{K} w_k \mathcal{N}(D_i|T(a^{(k)}),\sigma_i)\right)$$

Likelihood of GMM

MICHIGAN STATE UNIVERSITY

Bayesian Model Averaging (BMA)



See also talk by Ethan Neil
and arxiv:2008.01069

$$\prod_{i=1}^{N_D}\left(\sum_{k=1}^{K} P(T(a^{(k)})|D_i)\right) \propto \prod_{i=1}^{N_D}\left(\sum_{k=1}^{K} w_k \mathcal{N}(D_i|T(a^{(k)}), \sigma_i)\right)$$

Application of GMM and BMA to a toy model of PDFs

MICHIGAN STATE UNIVERSITY

A toy model of PDFs with inconsistent data

"truth" $$g(x) = a_0\ x^{a_1}(1-x)^{a_2} e^{x a_3}(1+x e^{a_4})^{a_5}$$

Parameters of model: $\{a_0, a_1, a_2, a_3, a_4, a_5\}$

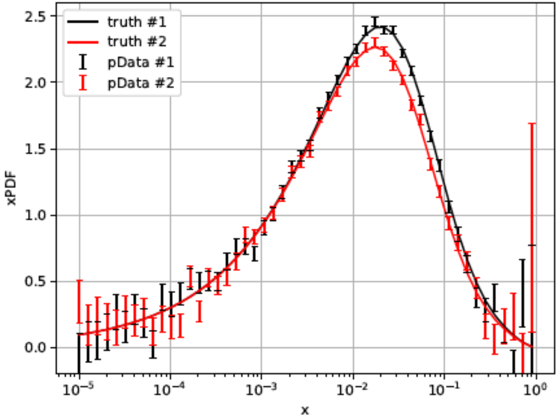
Pseudo-data generation

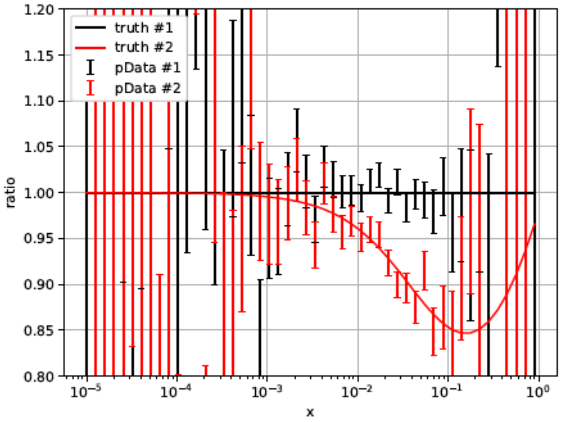
Central value $$g_D(x) = \Big(1 + r \times \Delta g(x)\Big) g(x)$$

Uncertainty $$\Delta g(x) = \frac{\alpha}{\sqrt{g(x)}}$$

	$N_{\rm pt}$	a_0	a_1	a_2	a_3	a_4	a_5
pseudo-data #1	50	30	0.5	2.4	4.3	2.4	-3.0
pseudo-data #2	50	30	0.5	2.4	4.3	2.6	-2.8

Inconsistent Pseudo-data generated by starting with different values of a_4 & a_5



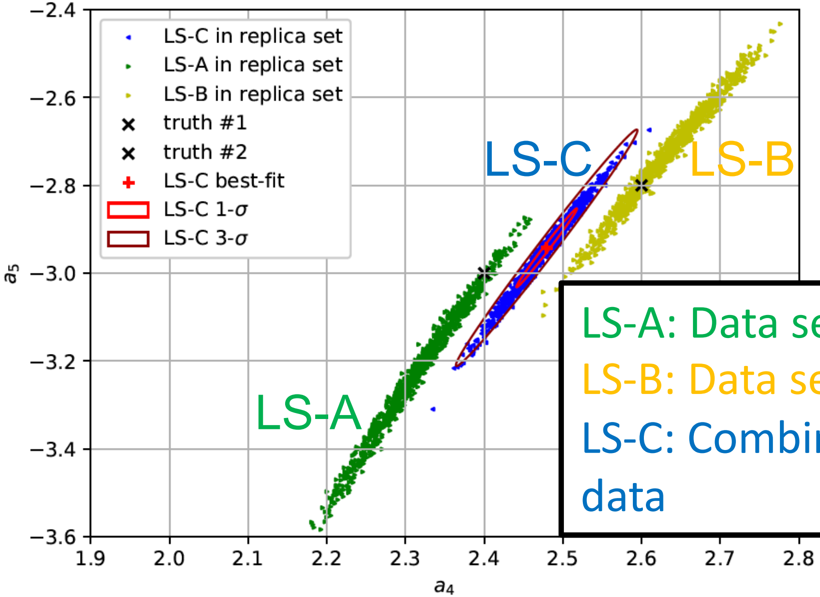


Fits to pseudo-data

$$\chi^2 = \sum_{j=1}^{N_{\mathrm{pt}}} \left(\frac{D_i - T_i(\theta)}{\Delta D_i} \right)^2$$

MICHIGAN STATE UNIVERSITY

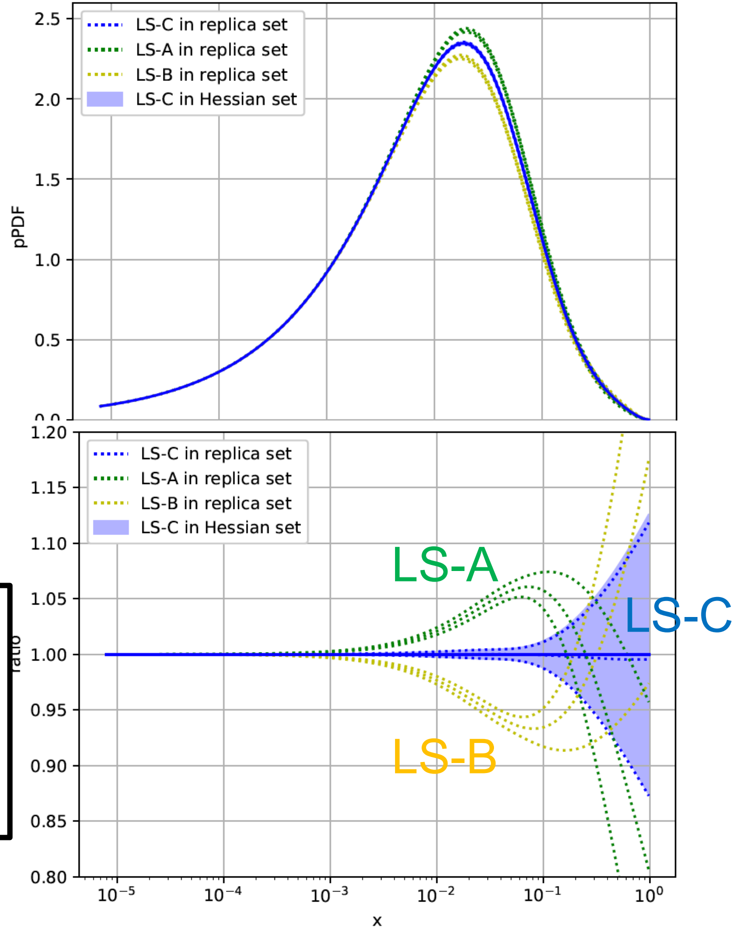
fits	pseudo-data	best-fit a_4	best-fit a_5	$\chi^2_{\#1}/N_{\mathrm{pt}}$	$\chi^2_{\#2}/N_{\mathrm{pt}}$
LS-A	# 1	2.32	-3.22	0.88	6.55
LS-B	# 2	2.63	-2.73	7.00	1.02
LS-C	# 1 and # 2	2.48	-2.94	2.27	2.56
truth	# 1	2.4	-3.0	-	-
truth	# 2	2.6	-2.8	-	-



LS-A: Data set 1 only

LS-B: Data set 2 only

LS-C: Combines all data

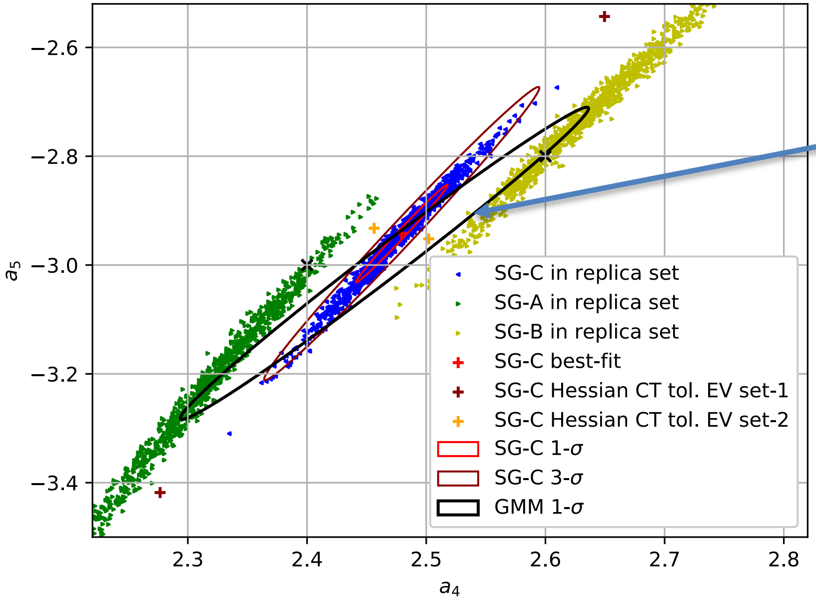


Fits to pseudo-data using the GMM

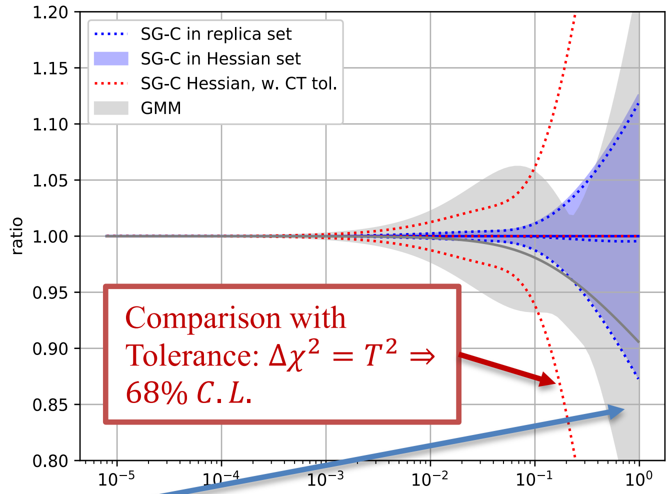
GMM uncertainty ellipse spans both replica sets. Unlike usual χ^2 method

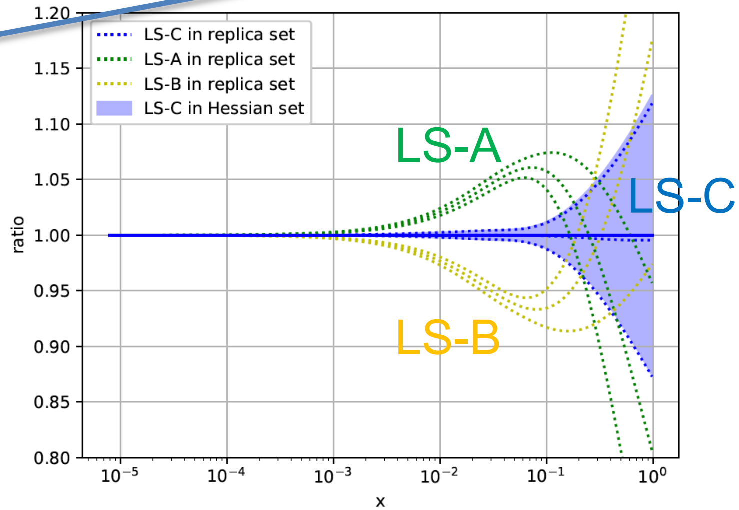
Axis of ellipse is different – covers uncertainties from individual data sets

Tolerance criteria both over and underestimates uncertainties in different regions

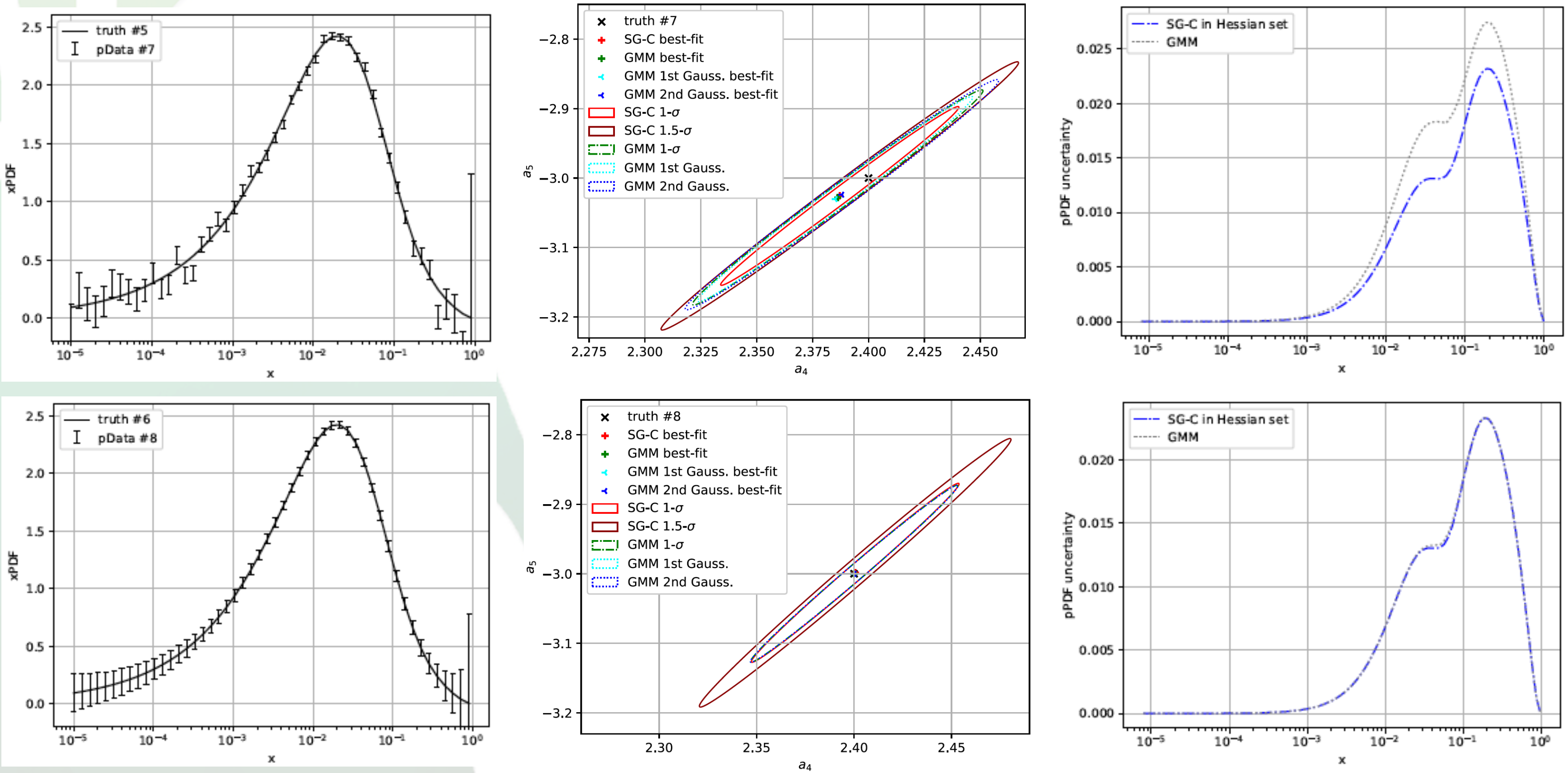


GMM "1σ"





GMM reduces to the χ^2 likelihood (K= 1), when data is consistent



How many Gaussians? How do we determine K?

Akaike Information Criterion (AIC) (Akaike, 1974)

Bayesian Information Criterion (BIC) Schwarz (Ann Stat 1978, 6:461–464)

$$\mathrm{AIC} = N_{\mathrm{parm}} \log N_{\mathrm{pt}} - 2\log L|_{\theta=\hat{\theta}},$$
$$\mathrm{BIC} = 2N_{\mathrm{parm}} - 2\log L|_{\theta=\hat{\theta}}.$$

$$N_{\mathrm{parm}} = 2K + (K-1).$$

Use the lowest values of AIC & BIC to determine the best value of K and avoids over-fitting.

		$K=1$	$K=2$	$K=3$	$K=4$
case-1	AIC	-102.2	**-203.6**	-194.9	-187.9
	BIC	-106.1	**-211.2**	-206.4	-203.2
N_{pt}=100	$-\log L$	-55.0	**-109.6**	-109.2	**-109.6**
case-2	AIC	**-21.2**	-15.4	-7.9	-0.2
	BIC	**-25.0**	-23.0	-19.3	-15.5
N_{pt}=100	$-\log L$	-14.5	-15.5	**-15.7**	**-15.7**
case-3	AIC	-219.3	**-220.2**	-212.8	-205.0
	BIC	-223.2	**-227.8**	-224.3	-220.3
N_{pt}=100	$-\log L$	-113.6	**-117.9**	**-117.9**	-118.1
case-4	AIC	**-117.8**	-109.9	-102.1	-94.3
	BIC	**-121.6**	-117.6	-113.6	-109.6
N_{pt}=50	$-\log L$	**-62.8**	**-62.8**	**-62.8**	**-62.8**
case-5	AIC	**-169.3**	-161.5	-153.6	-145.8
	BIC	**-173.1**	-169.1	-165.1	-161.1
N_{pt}=50	$-\log L$	**-88.6**	**-88.6**	**-88.6**	**-88.6**

Strong tension (case-1)

Weak tension due to large uncertainty (case-2)

Consistent but data fluctuated (case-4)

Consistent - No fluctuation (case-5)

$$\pi(Y|\vec{\theta}) = \prod_{j=1}^{N_{\mathrm{pt}}} \pi(y_j, \Delta y_j|\vec{\theta}) = \prod_{j=1}^{N_{\mathrm{pt}}} \sum_{i=1}^{K} \omega_i \mathcal{N}(y_j, \Delta y_j|\theta_i),$$

$$0 \le \omega_k \le 1 \quad \text{and} \quad \sum_k \omega_k = 1,$$

Summary & Outlook

- Showed how to repurpose the GMM, a well-known machine learning classification tool, as a statistical model to estimate uncertainty in PDF fits
  - Can also be used to classify PDF fitting data and find tensions in data sets – unsupervised machine learning task
- Provides an implementation of Bayesian Model Averaging, to provide statistically robust estimates of uncertainty.
- Can be used in conjunction with both the Hessian and Monte-Carlo method of PDF uncertainty estimation
  - Tools to develop this already exist in machine learning packages like TensorFlow/PyTorch/ scikit-learn
- Here I only showed tension due to experimental inconsistencies, but this also applies to tension resulting from imprecise theoretical predictions.
- Can be used to determine a value of Tolerance in order to connect with existing prescriptions.
- Next steps: Apply to real data and pdf fit.