

Generalized statistical model for fits to parton distributions

Mengshi Yan^{1,*} Tie-Jiun Hou^{2,†} Zhao Li^{3,4,‡} Kirtimaan Mohan^{5,§} and C.-P. Yuan^{5,||}

¹*Department of Physics and State Key Laboratory of Nuclear Physics and Technology, Peking University, Beijing 100871, China*

²*School of Nuclear Science and Technology, University of South China, Hengyang 421001, China*

³*Institute of High Energy Physics, Chinese Academy of Sciences, Beijing 100049, China*
⁴*and School of Physics Sciences, University of Chinese Academy of Sciences, Beijing 100039, China*

⁵*Center for High Energy Physics, Peking University, Beijing 100871, China*

⁵*Department of Physics and Astronomy, Michigan State University, East Lansing, Michigan 48824, USA*

 (Received 1 July 2024; revised 21 May 2025; accepted 21 May 2025; published 13 August 2025)

Parton distribution functions (PDFs) form an essential part of particle physics calculations. Currently, the most precise predictions for these nonperturbative functions are generated through fits to global data. A problem that several PDF fitting groups encounter is the presence of tension in datasets that appear to pull the fits in different directions. In other words, the best fit depends on the choice of dataset. Several methods to capture the uncertainty in PDFs in presence of seemingly inconsistent fits have been proposed and are currently in use. These methods are important to ensure that uncertainty in PDFs are not underestimated. Here we propose a novel method for estimating the uncertainty by introducing a generalized statistical model based on Bayesian hierarchical models, which is implemented via the Gaussian Mixture Model (GMM). The methodology is inspired by unsupervised machine learning techniques and is closely related to the statistical methods of ensemble learning and Bayesian model averaging. Using a toy model of PDFs, we demonstrate how the GMM can be used to faithfully reconstruct the likelihood associated with PDF fits, which can in turn be used to accurately determine the uncertainty on PDFs, especially in the presence of tension in the fitted datasets. We further show how this statistical model reduces to the usual χ^2 likelihood function for a consistent dataset and provide measures to optimize the number of Gaussians in the GMM.

DOI: [10.1103/zcsy-34cc](https://doi.org/10.1103/zcsy-34cc)

I. INTRODUCTION

Parton distribution functions (PDFs) are nonperturbative functions quantifying the probability of finding a quark or gluon with certain momentum fraction x from a nucleon. Taking advantage of a wide variety of experimental data and advanced computing packages for perturbative QCD (and electroweak) calculations, various groups [1–5] provide a sophisticated determination of PDFs. The recently released NNPDF [6] and xFitter [7] fitting codes open opportunities for general users to also conduct global QCD analyses.

These PDFs form a critical component of the theory community's infrastructure [8] and developing the next generation of PDFs are crucial for achieving the physics goals of the Large Hadron Collider (LHC) and the Electron-Ion Collider. An unbiased determination of PDFs is essential for precision phenomenology and the search for new physics, such as the recent measurement of W -boson mass by the CDF Collaboration [9]. A robust understanding of the uncertainties arising from PDFs is necessary to provide precise predictions. PDF uncertainties constitute one of the dominant systematic errors both for direct searches of new physics [10] and for precision tests of the Standard Model (SM), such as determining the properties of the Higgs boson [11]. A faithful determination of PDFs and its uncertainties are thus necessary in order to make precise theory predictions at hadron-hadron and lepton-hadron colliders. Improving PDFs will require new experimental data, new theoretical (higher-order) calculations and a determination of fits to global data in a statistically robust way [8].

Many issues could be encountered when trying to perform global fits to datasets from a variety of experiments. One of these issues, which is the primary focus of

*Contact author: msyan@pku.edu.cn

†Contact author: tjhou@msu.edu

‡Contact author: zhaoli@ihep.ac.cn

§Contact author: kamohan@msu.edu

||Contact author: yuanch@msu.edu

Published by the American Physical Society under the terms of the [Creative Commons Attribution 4.0 International](https://creativecommons.org/licenses/by/4.0/) license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI. Funded by SCOAP³.

this paper, is in dealing with data that appear to be in tension with each other, leading to inconsistencies within the global fit. For example, suppose we have two datasets that pull the fit parameters to different values. In this scenario, the data are in tension with each other. The source of such tension could arise due to deficiencies in theoretical models and calculations or even from incorrect estimation of experimental uncertainties. To keep the discussion simple, we work with an example of the latter, although, the method presented here is agnostic to the source of tension.

An example of tension between datasets would be the measurement of the ATLAS 8 TeV W and Z rapidity distribution [12], which was found to be inconsistent with other data included in the PDF global analysis and favored an enhanced strange quark distribution function [1,2,13]. Although a nonvanishing strangeness asymmetry $(s - \bar{s})(x)$ can partially release this tension [2,13], it is still inadequate to accommodate the ATLAS 8 TeV W/Z data [12] consistently in a PDF global analysis.

Another example is the inconsistency between the E866 NuSea and E906 SeaQuest measurements of the cross-section ratio $\sigma(pd)/2\sigma(pp)$, which approximates the ratio of antiquark PDFs $\sigma(pd)/2\sigma(pp) \approx (1 + \bar{d}_p(x_2)/\bar{u}_p(x_2))/2$. The E866 NuSea and E906 SeaQuest measurements provide valuable information on the PDF ratio $(\bar{d}/\bar{u})(x)$ in the large- x region. However, they are found to be in tension at large- x [13–15]. The effect of tension between certain datasets generally becomes negligible for large datasets, when outliers appear with small probability. However, in this case, E866 NuSea and E906 SeaQuest, the measurements probe a specific kinematic region that is otherwise weakly constrained, and hence the tension between them cannot be ignored. It is thus necessary to develop a prescription to accommodate such inconsistencies in a PDF global analysis in order to properly estimate the PDF uncertainty.

Several methods have been proposed and are currently in use to provide an estimate of uncertainties, especially in the presence of tension [16]. One naive approach would be to simply omit one of the troublesome datasets or perhaps attach a significantly lower weight to this dataset. This approach introduces biases in the fitting procedure that need to be justified with some prior knowledge. In some cases, such a justification may be possible, for example, if perhaps the methodology of the omitted experiment is analyzed and found to be unsatisfactory. In such a scenario, the prior bias in omitting the dataset from this experiment is justified, but this is not always the case, since it is not always straightforward to reanalyze particle physics experiments and their data.

Modern PDFs perform fits to data by defining a likelihood or loss function. For PDFs produced by CTEQ-TEA (CT) [1] and MSHT [2], the likelihood is the usual $\Delta\chi^2$. Under normal circumstances with a consistent set, the

uncertainty can be defined by the 68% confidence level (CL) limit of the χ^2 distribution, which amounts to varying the fit and determining boundaries where $\Delta\chi^2 = 1$. The convention used by the PDF global analysis groups, such as CT and MSHT, is to introduce a tolerance T such that the fits should be varied within $\Delta\chi^2 = T^2$. Values of T are chosen either dynamically [2] or set to a fixed value [1]. Typical values of T are much larger than 1, which increases the uncertainties in the PDFs. For example, $T^2 = 37$ corresponds to the CT tolerance at the 68% CL. NNPDF uses the Monte Carlo method for determining uncertainties, where they produce representative sample of PDFs by sampling their likelihood or loss function with the help of pseudodata, which might be one of the reasons for the uncertainties from NNPDF's analysis being notably different [17]. Although, as pointed out in Refs. [18–20], the level of inconsistency with datasets is found to be negligible in the NNPDF3.1 analyses, further study is needed to understand differences between NNPDF and the other fitting groups. One such task was performed in Ref. [21], which details benchmark comparisons of three global PDF sets—CT18, MSHT20, and NNPDF3.1—and their similarities and differences that have been observed. The end result of that study was a new PDF4LHC21 ensemble of combined PDFs suitable for a wide range of LHC applications.

In this paper, we describe how to apply a more general statistical framework, namely that of Bayesian hierarchical models [22–24], to accommodate potential inconsistencies in PDF data. This rather nontrivial task is accomplished with the aid of the Gaussian Mixture Model (GMM), which is commonly used as an unsupervised machine learning technique as well as a method of combining models. The approach used here can be interpreted as an implementation of the statistical method of ensemble learning [25,26] and the related method of Bayesian model averaging (BMA) [27–31]. BMA is a statistical method used to account for uncertainties arising from model selection. Although it is most often used in the context of linear regression, here we apply it to highly nonlinear fits.¹ On the other hand, ensemble learning, also known as ensemble averaging or Bayesian model combination, is a method to combine predictions of multiple statistical models. Although there are some subtle differences between ensemble learning and Bayesian model averaging, which we discuss briefly in Appendix B 2, we do not dwell on these differences, as they are subject to differences in the interpretation of priors. Ensemble learning is primarily used to combine different theoretical models and their point estimates. On the other hand, Bayesian hierarchical models are primarily applied to cases where there are variations across groups of data, and their use has also been suggested in estimating uncertainty

¹Application of Bayesian model averaging to analyze lattice field theory results can be found in Refs. [32–34].

when there is tension in datasets [24]. Here, we demonstrate how ensemble learning can be used to construct a Bayesian hierarchical model and how it can be used to provide an estimate of the uncertainty in the scenario where there are tensions among datasets. We will not discuss combining or averaging over different theoretical models, which is beyond the scope of this paper. Our major findings are summarized below:

- (i) When fitting inconsistent datasets with tension, the usual least squares (LS) method that employs the χ^2 likelihood underestimates uncertainty. Specifically, the posterior probability distribution for fitted theory parameters should reflect the tension in the datasets and appropriately span the range of possible theory values. We demonstrate how the LS method and its adaptations to address tension in datasets do not appropriately span the range of possible theory values. Another drawback of these methods is that they introduce *ad hoc* criteria to expand uncertainties. These deficiencies are overcome by the use of the GMM.
- (ii) We introduce a new method to fitting inconsistent data, utilizing the Gaussian Mixture Model. The method is closely related to ensemble learning and Bayesian model averaging that are used to estimate uncertainties due to selection of theoretical models. Here, we show how these methods can be repurposed to help determine uncertainties by constructing a Bayesian hierarchical model when there is tension between datasets.
- (iii) The main idea is a modification of the likelihood function used to determine the posterior probability distribution of the theory parameters. This modification, given in Eq. (3.1), accounts for the possibility of fitting with a multimodal distribution rather than a unimodal distribution assumed in the LS method. Estimation of theory parameters and their associated uncertainties proceed through the usual maximum likelihood estimation. We describe methods and provide formulas for the estimate of the mean and variance in Eqs. (3.10) and (3.18), respectively. We demonstrate the use and characteristics of the GMM method using various toy examples where we try to combine various pseudodatasets with varying degrees of tension.
- (iv) The fit presented in Figs. 4(d) and 4(e) is an example of how the GMM method provides a more accurate representation of the uncertainty compared to the LS method, by spanning the range of possible theory predictions, given the tension in the datasets. Specifically, using the GMM, we find that the uncertainty bands are larger when the tension is more prominent and smaller otherwise. On the other hand, more traditional methods of addressing tension tend to expand uncertainties more uniformly, resulting in

larger uncertainty bands in regions where the tension is not large.

- (v) The GMM introduces the number of modes (K) of the distribution as a hyperparameter. We demonstrate the use of Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) not only to determine K in order to avoid overfitting or underfitting, but also to be used as quantitative tests of the consistency of datasets.

The rest of the paper is organized as follows: in Sec. II, we set up a toy PDF model and describe different pseudodatasets, some of which have obvious tension between them. In Appendix A, we describe how the pseudodatasets we use in the fits are generated. Section III introduces the Gaussian Mixture Model as a statistical model that can be used to tackle inconsistencies in datasets. In Sec. IV, the results of the GMM fit are compared with those of conventional fits, using the usual least squares fit. Consistency between the GMM and the commonly used LS method is evaluated using a dataset with no tension in Sec. V. We discuss our results and conclude in Sec. VI. Additionally, we describe a more intuitive Bayesian perspective of the GMM in Appendix B and discuss some of the differences and intricacies related to BMA and ensemble learning in Appendix B 3. In Appendix C, we provide a much simpler and more accessible example using the real world data of W -boson mass measurements.

II. TENSION IN A TOY MODEL

In this section, we introduce a toy model of PDFs and the pseudodatasets to fit the PDFs. We perform LS fits to illustrate the problems encountered when there is tension between datasets.

We begin by considering two sets of data that we refer to as pseudodata #1 and pseudodata #2, containing 50 data points each. These two datasets are not real, but are simulated. However, we treat them as though they are measurements that likely came from two separate experiments. Details of how these pseudodatasets simulated are relegated to Appendix A. The two pseudodatasets are presented in Fig. 1, showing measured values of a pseudo-PDF as a function of x . Both datasets have a larger uncertainty at low and high values of x , but smaller uncertainties at intermediate values of $x \sim 10^{-2}$, to mimic general features of real experimental data. This choice represents the possibility that larger values of the PDF may correspond to larger number of events and therefore a reduced statistical uncertainty. The datasets have an obvious tension between them in the region $10^{-2} \lesssim x \lesssim 10^{-1}$. Such a tension might be due to underestimated systematics or even due to underestimated theoretical uncertainties.

Given the two pseudodatasets, we would like to perform a combined fit to both datasets. We fit the data to a function of the form

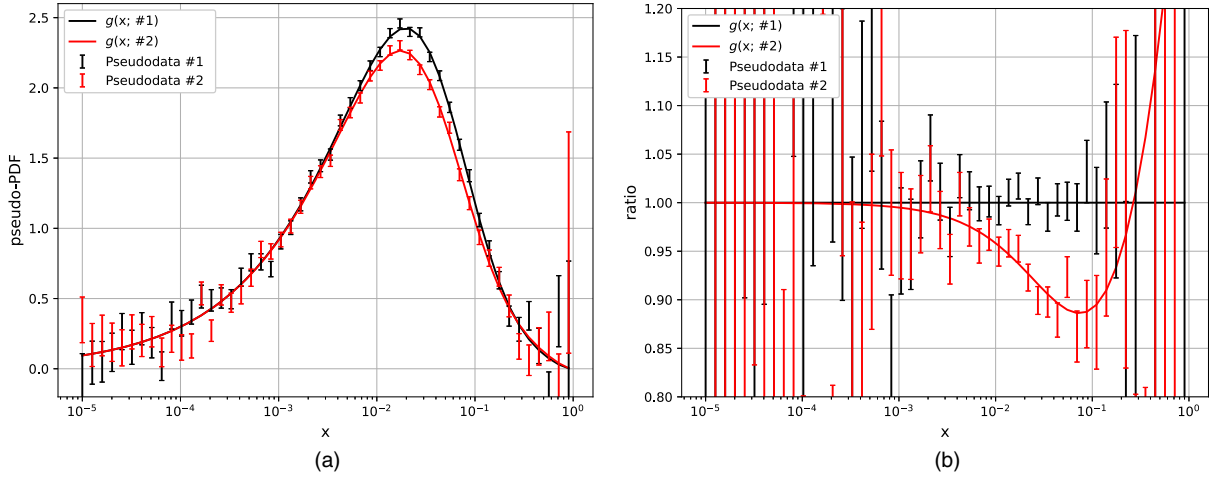


FIG. 1. Two pseudodatasets, pseudodata # 1 (black) and pseudodata # 2 (red) are shown to be in tension with each other. The solid curves $g(x; \#1)$ and $g(x; \#2)$ correspond to the curves given by the function in Eq. (2.1), with parameter values given in Table IV. Panel (a) shows the two pseudo-PDFs as a function of x and panel (b) shows their ratio with respect to $g(x; \#1)$.

$$g(x) = a_0 x^{a_1} (1 - x)^{a_2} e^{x a_3} (1 + x e^{a_4})^{a_5}. \quad (2.1)$$

Here, $\{a_0, a_1, a_2, a_3, a_4, a_5\}$ are the theory parameters that we would like to determine from the pseudodata. In addition, we would also like to estimate the uncertainty on these fitted parameters from the given data. To keep the discussion simple, we only perform fits to a subset of these parameters, while keeping others fixed. We would like to emphasize that although we chose an explicit functional form, the starting point could have also been a deep neural network or a polynomial parametrization.² Here, we restrict ourselves to use only the functional form in Eq. (2.1) for clarity of presentation.

Our purpose is to determine point estimates, i.e., the mean and variance, of the theoretical parameters a , given the experimentally measured data. Note, in the process of generating the two pseudodatasets, we do have knowledge the values of these parameters, c.f., Appendix A. However, when performing the fits, we do so by excluding any prior knowledge of the theory parameters a . Our purpose is not to try to recover parameter values that were used to generate each pseudodataset separately; instead, we would like to determine an estimate of the theory parameters a given a combination of both datasets. Furthermore, to be consistent with QCD, we demand that we have only one set of theoretical parameters, so that there is only one estimate of the mean. We now proceed to first perform LS fits to the two pseudodatasets in Fig. 1.

²The use of a deep neural network is supported by the Universal Approximation Theorem [35], whereas a polynomial is supported by the Stone-Weierstrass theorem [36,37]. These theorems ensure that any smooth function can be well approximated either by neural networks or by polynomials.

A. Least squares fits and tension

Before introducing the GMM method, we first illustrate the method of handling inconsistent data in the least squares fitting framework. Fits are performed by optimizing the log likelihood, which in our case is the χ^2 function defined as

$$\chi^2(\theta) = \sum_{j=1}^{N_{\text{pt}}} \left(\frac{y_i - T(\theta)}{\Delta y_i} \right)^2. \quad (2.2)$$

Given the known N_{pt} measurements y_i associated with uncertainties Δy_i , the χ^2 provides a normalized distance from theoretical predictions $T(\theta)$ to the measurements. Here, θ are the fit parameters, which are some subset of the parameters a_i , and are varied in order to optimize the χ^2 function.

To illustrate the problem with the LS fit when tension exists, three LS fits are performed, which we refer to as SG-A (performed only with pseudodata #1), SG-B (performed only with pseudodata #2) and SG-C (performed by combining pseudodata #1 and pseudodata #2). We label these fits as “SG” to refer to the single-Gaussian nature of the likelihood in the LS method, in contrast to the multi-modal distribution that we will introduce later. Results of these fits are shown in Table I. We fit the data to the function given in Eq. (2.1) by varying and determining the parameters of the function that minimize the χ^2 in Eq. (2.2). To keep the discussion and fits simple in this toy example, we fix parameters $\{a_0, a_1, a_2, a_3\}$ to values which can be found in Table IV. Only two parameters a_4 and a_5 are varied and the values that minimize the χ^2 function are shown in Table I. The values of $\chi^2_{\#1}/N_{\text{pt}}$, $\chi^2_{\#2}/N_{\text{pt}}$, and $\chi^2_{\#1+\#2}/N_{\text{pt}}$, which are approximately equal to the reduced

TABLE I. Best-fit values of a_4 and a_5 , as well as a measure of the goodness of fit χ^2/N_{pt} for three LS fits SG-A, SG-B, and SG-C, performed using pseudodatasets # 1 and # 2. The columns $\chi^2_{\#1}/N_{\text{pt}}$, $\chi^2_{\#2}/N_{\text{pt}}$, and $\chi^2_{\#1+\#2}/N_{\text{pt}}$ indicate the values of these log-likelihood functions evaluated at values of a_4 and a_5 in the given rows. Also shown, for comparison, in the last two rows are the true values of a_4 and a_5 that were used to generate the two pseudodatasets.

Fits	Pseudodata	N_{pt}	Best fit a_4	Best fit a_5	$\chi^2_{\#1}/N_{\text{pt}}$	$\chi^2_{\#2}/N_{\text{pt}}$	$\chi^2_{\#1+\#2}/N_{\text{pt}}$
SG-A	# 1	50	2.32	-3.22	0.88	6.55	3.72
SG-B	# 2	50	2.63	-2.73	7.00	1.02	4.01
SG-C	# 1 and # 2	100	2.48	-2.94	2.27	2.56	2.42

χ^2 , indicate the goodness of fit to individual pseudodatasets # 1, # 2, and the combined pseudodataset, respectively. As is well known, a good fit will have values of a $\chi^2/N_{\text{pt}} \sim 1$, whereas a poor fit has χ^2/N_{pt} values significantly larger than 1.

The best fit of SG-A is found with a minimum of $\chi^2_{\#1}/N_{\text{pt}} = 0.88$, with values of $a_4 = 2.32$ and $a_5 = -3.22$. Also shown in the same row of Table I for SG-A are values of $\chi^2_{\#2}/N_{\text{pt}}$ and $\chi^2_{\#1+\#2}/N_{\text{pt}}$. These indicate the values of each of the likelihood functions $\chi^2_{\#2}/N_{\text{pt}}$ and $\chi^2_{\#1+\#2}/N_{\text{pt}}$ determined at the best fit values of a_4 and a_5 for fits SG-A. We see that for SG-A, the values of $\chi^2_{\#2}/N_{\text{pt}} = 6.55$ and $\chi^2_{\#1+\#2}/N_{\text{pt}} = 3.72$ are significantly larger than 1, indicating that there is significant tension between both the combined dataset and dataset # 2.

Likewise, the best fit of SG-B is found with a minimum $\chi^2_{\#2}/N_{\text{pt}} = 1.02$, indicating a reasonably good fit. For SG-B, the values $\chi^2_{\#1}/N_{\text{pt}}$ and $\chi^2_{\#1+\#2}/N_{\text{pt}}$ are determined by using the best fit value of $a_4 = 2.63$ and $a_5 = 2.72$. Once again large values of $\chi^2_{\#1}/N_{\text{pt}} = 7$ and $\chi^2_{\#1+\#2}/N_{\text{pt}} = 4.01$ indicate tension with both the combined dataset and dataset # 1.³

The fit, SG-C, to the combined dataset, is found with a minimum $\chi^2_{\#1+\#2}/N_{\text{pt}} = 2.42$, indicating, as expected, a poor fit to the combined dataset. In this case, the values of $\chi^2_{\#1}/N_{\text{pt}} = 2.27$, $\chi^2_{\#2}/N_{\text{pt}} = 2.42$ indicate that the combined fit still has tension with both individual datasets.

Although the fit to the combined data is poor, the main problem arises when attempting to determine the relevant uncertainty associated with such a fit. This is illustrated in Fig. 2, which presents results of the LS fits SG-A, SG-B, and SG-C. For all three fits, results are presented in the form of replica sets, where the probability distribution of the pseudo-PDF is sampled with the Monte Carlo method as described in Ref. [6]. For the purpose of comparison, we also estimate the uncertainty of the SG-C fit using the Hessian method [1], which is consistent with the

Monte Carlo method. Here we only consider the Hessian uncertainty generated by the variation of χ^2 and determine the uncertainty with the criterion that $\Delta\chi^2 = 1$. We do not include any additional tolerance criterion, such as the two-tier penalty for the CT uncertainty [1] and the dynamic tolerance for the MSHT uncertainty [2]. Later, we will include such tolerance criteria for comparison with the GMM.

In Figs. 2(a) and 2(b), pseudo-PDFs from fits SG-A, SG-B, and SG-C are found to be in disagreement in the range $10^{-3} < x < 0.2$. Even the 1σ uncertainty bands do not overlap in this region. For $x > 0.2$, the tension is weak due to larger uncertainty in the data, which is represented by the overlapping uncertainty bands in this region. Even though the uncertainty bands overlap at large x , the impact of significant tension in the range $10^{-3} < x < 0.2$ still affects the pseudo-PDF at large x , as can be observed by the different shapes of the uncertainty bands for the different LS fits. For the combined LS fit SG-C, the Hessian uncertainty with $\Delta\chi^2 = 1$ is also provided, which deviates from the Monte Carlo uncertainty by a negligible amount. The tension between the different fits, SG-A, SG-B, and SG-C, can also be observed by looking at Fig. 2(c), which shows uncertainty on the parameters a_4 and a_5 . Monte Carlo samples of 1000 replicas are generated for each of the fits SG-A, SG-B, and SG-C. None of the replicas generated for either of the fits overlap, once again indicating the strong tension among the pseudodatasets.

We would like to emphasize that either or both experiments may have underestimated systematic uncertainties and in the absence of further information about the source of systematics, the range of possible theory values should span the theory values predicted by each experimental dataset individually. For example, an uncertainty band that looks like some combination of the uncertainty bands of both SG-A and SG-B should envelope the central values of the fits SG-A and SG-B, or at least come close to doing so. Such an uncertainty band would be more representative of the tension in the data. Figure 2(c) shows that this is not the case for the SG-C fit. The tolerance criteria used by the PDF fitting groups increase the uncertainty, but as will be shown in Sec. IV, this approach is not as precise as using the GMM. Further, in Sec. III, we implement ensemble

³The two pseudodatasets, # 1 and # 2, are generated by starting with different choices of a_4 and a_5 as given in Table IV. As expected, these values of a_4 and a_5 are similar to the fitted values in Table I for SG-A and SG-B and lie within the $1\text{-}\sigma$ confidence interval of the fit.

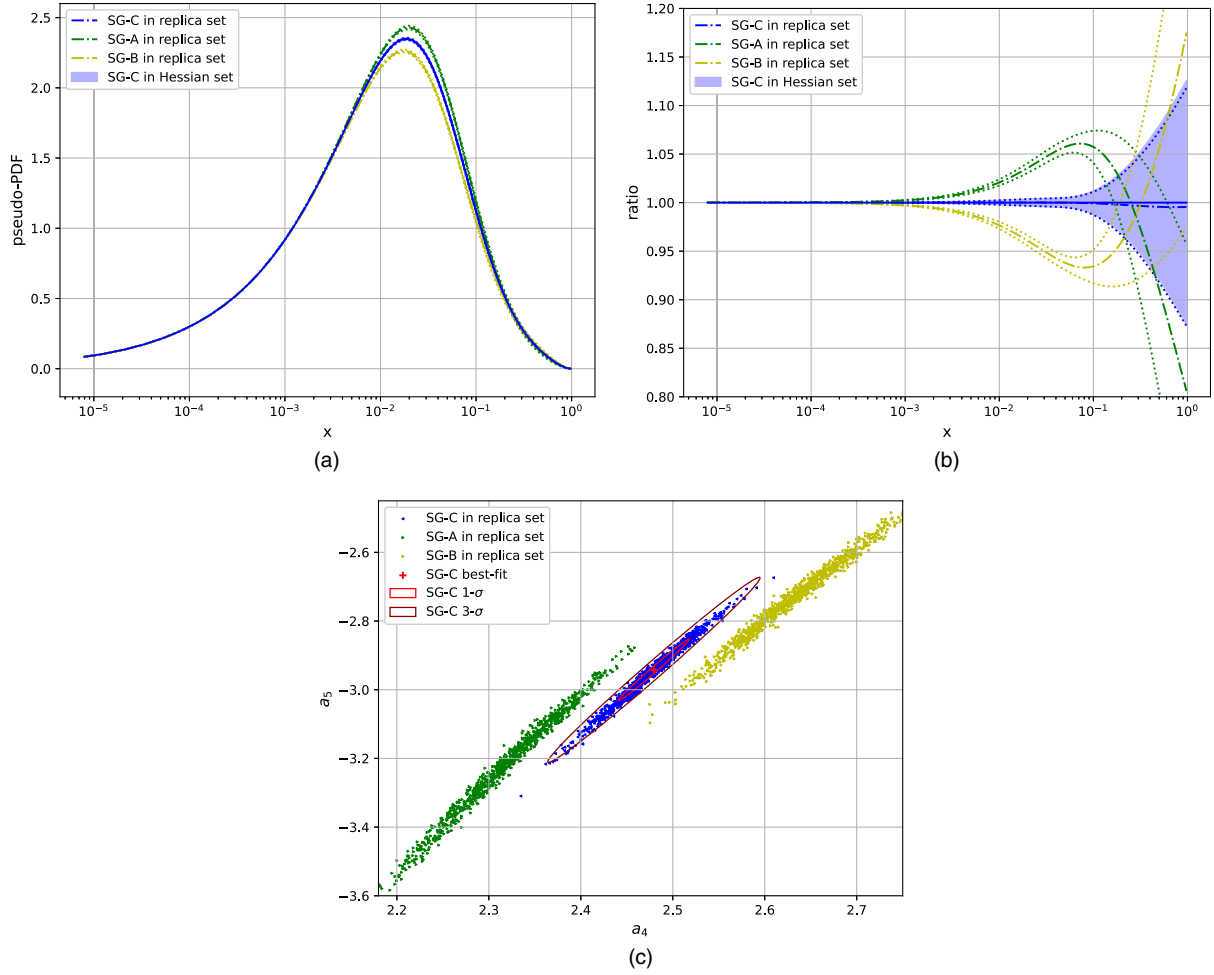


FIG. 2. Panel (a): the pseudo-PDFs determined from each of the fits are shown. The dot-dashed lines refer to the central PDFs for SG-A (green), SG-B (yellow), and the combination SG-C (blue). Uncertainties at 1 standard deviation in the replica sets are shown by the dotted lines for each fit. For the LS fit C, the blue shaded region represents the Hessian uncertainty with $\Delta\chi^2 = 1$, which deviates from the Monte Carlo uncertainty by a negligible amount. The sizes of uncertainties are small with respect to the magnitude of the pseudo-PDFs. Panel (b): similar to panel (a), but for ratios to the central value of the Hessian set of the SG-C fit. Panel (c): the values of parameters a_4 and a_5 in replica sets for SG-A, SG-B, and SG-C are shown along with the 1 and 3 standard deviation uncertainties for SG-C given by the solid curves.

learning with the use of the GMM and show how this allows for the combined likelihood of the pseudodata to be multimodal and results in a uncertainty band that overlaps with the uncertainty bands of SG-A and SG-B. Before we conclude this section, we comment on alternate methods to deal with such inconsistencies in fits. For example, as described in Appendix C, the Particle Data Group (PDG) [38] recommends increasing the uncertainty on datasets by a scale factor $S = \sqrt{\chi^2/(N_{\text{pt}} - 1)}$. For SG-C, this implies scaling the uncertainty on all the data by $S \approx 1.56$, which, for our simple example, translates to setting a tolerance $T = S \approx 1.56$. Increasing the error with the scale factor, or equivalently the tolerance, does not shift the central value of the fit and only increases the uncertainty marginally. In Sec. IV B, we explore even larger values of the tolerance and will see that such a small value of $T \approx 1.56$ is not

sufficient to reproduce uncertainty bands that represent the tension between the datasets.

It is crucial to understand why the uncertainty band of SG-C, even with the increased tolerance ($T = S \approx 1.56$), is unsatisfactory. It is easy to imagine that the true theory value could easily be close to the central values of SG-A or SG-B, i.e., there is a large range of possible values for the true theory parameters, given the data. On the other hand, the central values of SG-A and SG-B lie well outside the uncertainty band of the combined LS fit SG-C, even with $T \approx 1.56$. Hence, the uncertainty bands of SG-C do not provide an accurate representation of the situation. In other words, precision is lost and the uncertainties must be increased in the region of tension in order to avoid biasing the fits and to accommodate our lack of information about the source of tension. We describe a simpler example using W -boson mass in Appendix C, where we show how a naive

combination of data with tension can bias the fits even after increasing the scale factor S .

The method using the GMM, to be presented in the next section, provides an alternate way to combine inconsistent datasets with large tension and provides a more accurate determination of the uncertainties. Furthermore, if desired, it can also provide an alternate estimate for the PDG scale factor S required to make datasets consistent. We describe some of these ideas briefly in Sec. VB and Appendix C.

III. GAUSSIAN MIXTURE MODEL IN PDF FIT

As mentioned in Sec. II, the likelihood from each of the datasets (pseudodata # 1 and pseudodata # 2) are significantly different and need to be combined in a meaningful way. We propose the use of the Gaussian Mixture Model to construct the combined likelihood for data with tension.

The GMM was first used by Karl Pearson in 1894 [39]. It is now used extensively as an unsupervised machine learning technique to classify data [25]. Here, we use it to determine the uncertainty estimate on theoretical fit parameters given a set of discrepant data. In Appendix B, we present the more intuitive Bayesian description of the GMM. Here, we present the frequentist approach by starting with an ansatz for the likelihood function. In general, the GMM is a mixture of K base Gaussian distributions. Given the dataset $Y = \{(y_1, \Delta y_1), (y_2, \Delta y_2), \dots, (y_{N_{\text{pt}}}, \Delta y_{N_{\text{pt}}})\}$, where Δy_j is the uncertainty associated with the measured value y_j ,⁴ we can define the GMM likelihood as

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

$$0 \leq \omega_i \leq 1 \quad \text{and} \quad \sum_i \omega_i = 1, \quad (3.1)$$

where the i th base distribution $\mathcal{N}(y_j, \Delta y_j|\theta_i)$ is associated with a normalized weight ω_i . Here, the index j labels each experimental data point, and index i labels each Gaussian with theory parameters θ_i that are estimated by maximizing the likelihood above. In the example fit that we use, $\theta_i = \{a_4, a_5\}$ where $i = 1, 2, \dots, K$. For example, if $K = 2$, then $\theta_1 = \{a_4, a_5\}$ and $\theta_2 = \{a_4, a_5\}$, and we have effectively introduced four parameters to fit. If $K = 1$, then $\theta_1 = \{a_4, a_5\} = \{a_4, a_5\}$ and the log likelihood reduces to the usual χ^2 likelihood used in Sec. II A. For $K > 1$, in addition to the $(K - 1)$ independent weight parameters w_i , we have $N_{\text{param}} \times K$ parameters that are varied to maximize the likelihood, where N_{param} is the number of parameters in the fitting function $g(x)$ that are being estimated. Since the number of original parameters

N_{param} has increased in the GMM by a factor K , we need to determine not only the central value of $K \times N_{\text{param}}$ parameters but also the uncertainty on each of them. Later, we will describe how this fit to $K \times N_{\text{param}}$ parameters can be combined in a meaningful way to produce a fit and uncertainty on only N_{param} parameters.⁵

To simplify the following expressions, we will abbreviate $\mathcal{N}$ and π for the i th base distribution and the j th measurement as follows:

$$\mathcal{N}(y_j, \Delta y_j|\theta_i) = \mathcal{N}_{ij}, \quad \pi(y_j, \Delta y_j|\vec{\theta}) = \pi_j. \quad (3.2)$$

In the GMM, the logarithm of the likelihood function is

$$-\ln L = -\ln \left(\prod_{j=1}^{N_{\text{pt}}} L(y_j, \Delta y_j|\vec{\theta}) \right) = -\sum_{j=1}^{N_{\text{pt}}} \ln \left(\sum_{i=1}^K \omega_i \mathcal{N}_{ij} \right)$$

$$= -\sum_{j=1}^{N_{\text{pt}}} \ln \left(\sum_{i=1}^K \frac{\omega_i}{\sqrt{2\pi}\Delta y_j} \exp \left[-\frac{1}{2} \left(\frac{y_j - T(\theta_i)}{\Delta y_j} \right)^2 \right] \right). \quad (3.3)$$

For a single Gaussian distribution with $K = 1$, the log-likelihood in Eq. (3.3) is proportional to χ^2 , which reproduces the LS method as shown in the equation below:

$$-\ln L = -\sum_{j=1}^{N_{\text{pt}}} \ln(\mathcal{N}(y_j, \Delta y_j|\theta)) = \frac{1}{2} \chi^2 + \sum_{j=1}^{N_{\text{pt}}} \ln(\sqrt{2\pi}\Delta y_j). \quad (3.4)$$

Notice, in Eq. (3.3), we have chosen to define the likelihood such that each Gaussian has information about all experimental data points, i.e., for each Gaussian, we sum overall values of $j = 1$ to $j = N_{\text{pt}}$. An alternate way of defining the likelihood would be to distribute the experimental data among the various Gaussians, i.e., each dataset would have its own Gaussian. These two definitions have different interpretations. In the latter, we are trying to assign different theoretical predictions to different experiments, whereas in the former we are trying to find the theoretical values that different combinations of all experiments predict. Since we are trying to combine all experiments into a single global fit, we focus on using the definition as given in Eq. (3.3). See Appendix B 2 for a more intuitive explanation of this idea using a Bayesian formalism.

The GMM method allows us to extend the statistical model assumed in the optimization procedure from a single Gaussian distribution to a more generalized configuration,

⁴For clarity of presentation, we use an uncorrelated dataset and present formulas for this scenario. These can easily be generalized when correlated uncertainties are present.

⁵For example, there were two fit parameters (a_4 and a_5) in the LS fits in Sec. II, whereas in the GMM there would be $2 \times K$ parameters. Ultimately, we want to be able determine the uncertainty on two parameters and not $2 \times K$ parameters.

where a multimodal feature is allowed. Importantly, for a consistent dataset, we will see that the multiple Gaussian modes will overlap and yield a result that is equivalent to a single Gaussian, i.e., the LS fit. We will discuss the consistency between the GMM and LS methods in Sec. V.

A. Performing parameter fits with the GMM

In order to determine the best fit parameters, we minimize the log-likelihood defined in Eq. (3.3). We remark here that conventionally the expectation-maximization algorithm [25] is used to learn the best-fit parameters of the mixture model.⁶ The advantage of this algorithm is that it provides an iterative method that does not require information about the derivative of the likelihood. For the purpose of demonstration, we use methods that rely on information about the derivative of the likelihood. Specifically, we use TensorFlow [40] to determine derivatives of the likelihood and minimize the log-likelihood. As described in the previous section, the number of parameters have increased, and for K Gaussians, we have

$$N_{\text{parm}}^{\text{GMM}} = K \times N_{\text{parm}} + (K - 1), \quad (3.5)$$

where the last term, $K - 1$, come from the constraint on the weights ω_i that are required to sum to unity. We carry out minimization of the likelihood by varying all $N_{\text{parm}}^{\text{GMM}}$ parameters.⁷ The gradient of the log-likelihood with respect to the fit parameters that belong to i th Gaussian is

$$\frac{\partial}{\partial \theta_i} \ln L = \sum_{j=1}^{N_{\text{pt}}} \omega_i \frac{\mathcal{N}_{ij} y_j - T(\theta_i)}{\pi_j \Delta y_j^2} \frac{\partial T(\theta_i)}{\partial \theta_i}, \quad (3.6)$$

and that with respect to weight ω_i is

$$\frac{\partial}{\partial \omega_i} \ln L = \sum_{j=1}^{N_{\text{pt}}} \frac{\mathcal{N}_{ij}}{(\sum_{i=1}^K \omega_i \mathcal{N}_{ij})} = \sum_{j=1}^{N_{\text{pt}}} \frac{\mathcal{N}_{ij}}{\pi_j}. \quad (3.7)$$

Optimal values, $\hat{\theta}_i$, for the parameters θ_i are found by extremizing the log-likelihood by solving [41]

$$\left. \frac{\partial}{\partial \theta_i} \ln L \right|_{\theta_i = \hat{\theta}_i} = 0 \quad (3.8)$$

along with a similar equation involving the weights ω_i ,

⁶We remark here that it is possible to use the expectation-maximization algorithm if we produce a large pseudodataset and apply the Gaussian mixture model directly to the theory parameter space.

⁷It is possible to determine the value of the weights ω_i in other more complicated ways, since they indicate the degree of importance to each Gaussian. Here, we take the approach that we learn the values of ω_i directly from the data, and hence it is also varied during the minimization process.

$$\left. \frac{\partial}{\partial \omega_i} \ln L \right|_{\omega_i = \hat{\omega}_i} = 0. \quad (3.9)$$

In this work, we extremize the log-likelihood function by using the Adam optimizer with learning rate being 10^{-6} provided by the TensorFlow API [40]. Note that in Eq. (3.6), the best-fit value of the prediction $y_j(\theta_i)$ is weighted not only by the variance in measurements (Δy_i), but also by the weights ω_i and the factor $\mathcal{N}(y_j|\theta_i)/\pi(y_j, \Delta y_j|\vec{\theta})$ of the i th base distribution, which becomes trivial for a single Gaussian. The Jacobian $\partial y_j(\theta_i)/\partial \theta_i$ transforms information from the data space to the PDF space. The presence of this transformation makes it difficult to use the expectation-maximization algorithm directly.

In this work, we use the above equation as a constraint that determines the value of ω_i . This is generally the approach taken for other unsupervised machine learning applications. In general, this strategy tends to assign higher weights to experiments with higher precision. If desired, it is possible to control the weights by increasing or decreasing the value of the uncertainties associated with a certain experimental dataset.⁸ Thus, along with θ_i , the weights ω_i are also varied in order to optimize the log-likelihood.

B. Mean and uncertainty in the GMM

It is straightforward to show that the expectation value of parameters is the weighted sum of the optimal parameters for individual base distributions [42],

$$\mathbb{E}[\theta] = \sum_{i=1}^K \omega_i \hat{\theta}_i. \quad (3.10)$$

Here, $\hat{\theta}_i$ are the estimated values of the N_{parm} parameters of the fitted function $g(x)$, and there are K such $\hat{\theta}_i$. Consequently, an experiment with higher precision would be associated with a higher weight and, as a result, the central value of the parameters is pulled towards those values that better agree with higher precision experiments.

There are several ways to determine the uncertainty with the modified log-likelihood. Below, we describe the use of the observed Fisher information matrix to estimate uncertainty [41] on the fit parameters. We note that this is only one way to estimate the uncertainty, and there are other possible methods which will be discussed in detail elsewhere.

When using a single Gaussian or when all the Gaussians have the same mean and covariance, the observed Fisher information matrix reduces to the usual Hessian matrix used

⁸One method of managing inconsistent datasets is to artificially inflate uncertainties until we arrive at a consistent dataset [16]. In Appendix C, we describe how the GMM, in conjunction with certain information criteria, provides a way to determine the extent to which it is necessary to inflate the uncertainty on data until the fits become consistent.

to determine the covariance of the fit parameters in the LS method.⁹ Like the Hessian matrix, the information matrix can also be found with numerical methods. Here, we use algorithmic differentiation [43] within TensorFlow to determine derivatives of the likelihood function. We note that the information matrix is only an estimate of the covariance of the GMM, and it is possible to determine the covariance matrix as well, which will be described in detail elsewhere.

Elements of the information matrix ($\mathcal{I}$) are determined by taking a further derivative on Eq. (3.8), as follows:

$$(\mathcal{I})_{mk} = -\frac{\partial^2}{\partial \theta_m \partial \theta_k} \ln L = \frac{\partial}{\partial \theta_m} \left(-\sum_{j=1}^{N_{\text{pt}}} \frac{\omega_k}{\pi_j} \frac{\partial \mathcal{N}_{kj}}{\partial \theta_k} \right) \\ = \sum_{j=1}^{N_{\text{pt}}} \left[-\frac{\omega_k}{\pi_j} \frac{\partial^2 \mathcal{N}_{mj}}{\partial \theta_m \partial \theta_m} \delta_{mk} + \frac{\omega_k \omega_m}{\pi_j^2} \frac{\partial \mathcal{N}_{mj}}{\partial \theta_m} \frac{\partial \mathcal{N}_{kj}}{\partial \theta_k} \right]. \quad (3.11)$$

Notice that for a fit with N_{parm} parameters per Gaussian, the information matrix is a $(KN_{\text{parm}}) \times (KN_{\text{parm}})$ symmetric matrix, i.e., it spans a space that is a direct product of an N_{parm} parameter and a K parameter space. The second term contributes to the elements of the information matrix ($\mathcal{I}$) that are off diagonal in K space but not in N_{parm} space. In contrast, the off-diagonal terms in N_{parm} space contribute to the covariance matrix. On the other hand, the K -space off-diagonal terms are small since they must add up to the left-hand side of Eq. (3.8), which vanishes at $\theta_i = \hat{\theta}_i$ by definition. This is also verified numerically for all of the fits performed here. For simplicity, we will neglect these particular K -space off-diagonal terms and only consider the $N_{\text{parm}} \times N_{\text{parm}}$ block diagonal terms of the information matrix. Although it is not necessary to neglect these off-diagonal terms, this allows us to quickly approximate the probability distribution function in parameter space (θ) as a mixture of Gaussians. In other words, we are trying to use the likelihood function defined in Eq. (3.3) to approximate the probability distribution function of θ . Thus, we assume that when evaluated at the mean ($\theta_i = \hat{\theta}_i$), the information matrix takes on an approximate block diagonal form, with each $N_{\text{parm}} \times N_{\text{parm}}$ block belonging to the subspace of the i th Gaussian $\mathcal{N}(y_j, \Delta y_j | \theta_i)$. We denote the information matrix for individual base distributions of the GMM as $\mathcal{I}_{\text{GMM},i} = -(\partial^2 \ln L / \partial \theta_i \partial \theta_i) / \omega_i$, which are the unweighted

$N_{\text{parm}} \times N_{\text{parm}}$ diagonal blocks of the full information matrix $\mathcal{I}$. An estimate of the covariance of individual base Gaussian distributions of the GMM is then

$$\text{cov}_{\text{GMM},i} \simeq (\mathcal{I}_{\text{GMM},i})^{-1} \\ = \left(\sum_{j=1}^{N_{\text{pt}}} \left[-\frac{1}{\pi_j} \frac{\partial^2 \mathcal{N}_{ij}}{\partial \theta_i^2} + \frac{\omega_i}{\pi_j^2} \left(\frac{\partial \mathcal{N}_{ij}}{\partial \theta_i} \right)^2 \right] \right)^{-1}. \quad (3.12)$$

We can now define the probability distribution of the parameters of the fit as mixture of Gaussians as follows:

$$\mathcal{P}(x | \theta_i, \text{cov}_{\text{GMM},i}) \\ = \sum_{i=1}^K \frac{\omega_i}{(2\pi)^{N_{\text{parm}}/2} \det(\text{cov}_{\text{GMM},i})^{1/2}} \\ \times \exp \left[-\frac{1}{2} ((x - \theta_i)^T \cdot (\text{cov}_{\text{GMM},i})^{-1} \cdot (x - \theta_i)) \right]. \quad (3.13)$$

With this definition, we can now determine estimates of the uncertainty on the parameters θ . We first note that when $K = 1$, we have a single Gaussian and Eq. (3.11) reduces to the Hessian matrix of the LS method, as follows:

$$-\frac{\partial^2}{\partial \theta \partial \theta} \ln L \Big|_{\theta=\hat{\theta}} = \frac{1}{2} \frac{\partial^2}{\partial \theta \partial \theta} \chi^2 \Big|_{\theta=\hat{\theta}} = \sum_{j=1}^{N_{\text{pt}}} \frac{1}{\Delta y_j^2} \frac{\partial^2 T(\theta)}{\partial \theta^2} \Big|_{\theta=\hat{\theta}}, \quad (3.14)$$

and the covariance is

$$\text{cov}_{\chi^2} = \left(\sum_{j=1}^{N_{\text{pt}}} \frac{1}{\Delta y_j^2} \frac{\partial^2 T(\theta)}{\partial \theta^2} \Big|_{\theta=\hat{\theta}} \right)^{-1}. \quad (3.15)$$

For the GMM, in order to provide a useful estimate of uncertainty, we need to reduce the $(KN_{\text{parm}}) \times (KN_{\text{parm}})$ space to a $(N_{\text{parm}}) \times (N_{\text{parm}})$ space in a meaningful way. To that end, let us consider the covariance of the GMM, which is defined as [42]

$$\text{cov}_{\text{GMM}} = \mathbb{E}[\theta^2] - \mathbb{E}[\theta]^2 \\ = \sum_{i=1}^K \omega_i \hat{\theta}_i^{(2)} - \left(\sum_{i=1}^K \omega_i \hat{\theta}_i \right)^2, \quad (3.16)$$

where $\hat{\theta}_i^{(2)}$ denotes the estimated second moment of parameters for the i th base distribution. It can be reexpressed using the covariance of i th base Gaussian distribution,

$$\hat{\theta}_i^{(2)} = \text{cov}_{\text{GMM},i} + \hat{\theta}_i^2. \quad (3.17)$$

⁹For the GMM with multiple Gaussians, the likelihood function as defined in Eq. (3.3) includes all experimental points for each Gaussian. Thus, if the means of each Gaussian are identical, then this implies that the covariance, which is related to the second-order derivative about the minimum of the likelihood, must also be identical; c.f., Eq. (3.18). In general, the covariance matrices for Gaussians with identical means may be different in the GMM if the likelihoods were to be defined differently.

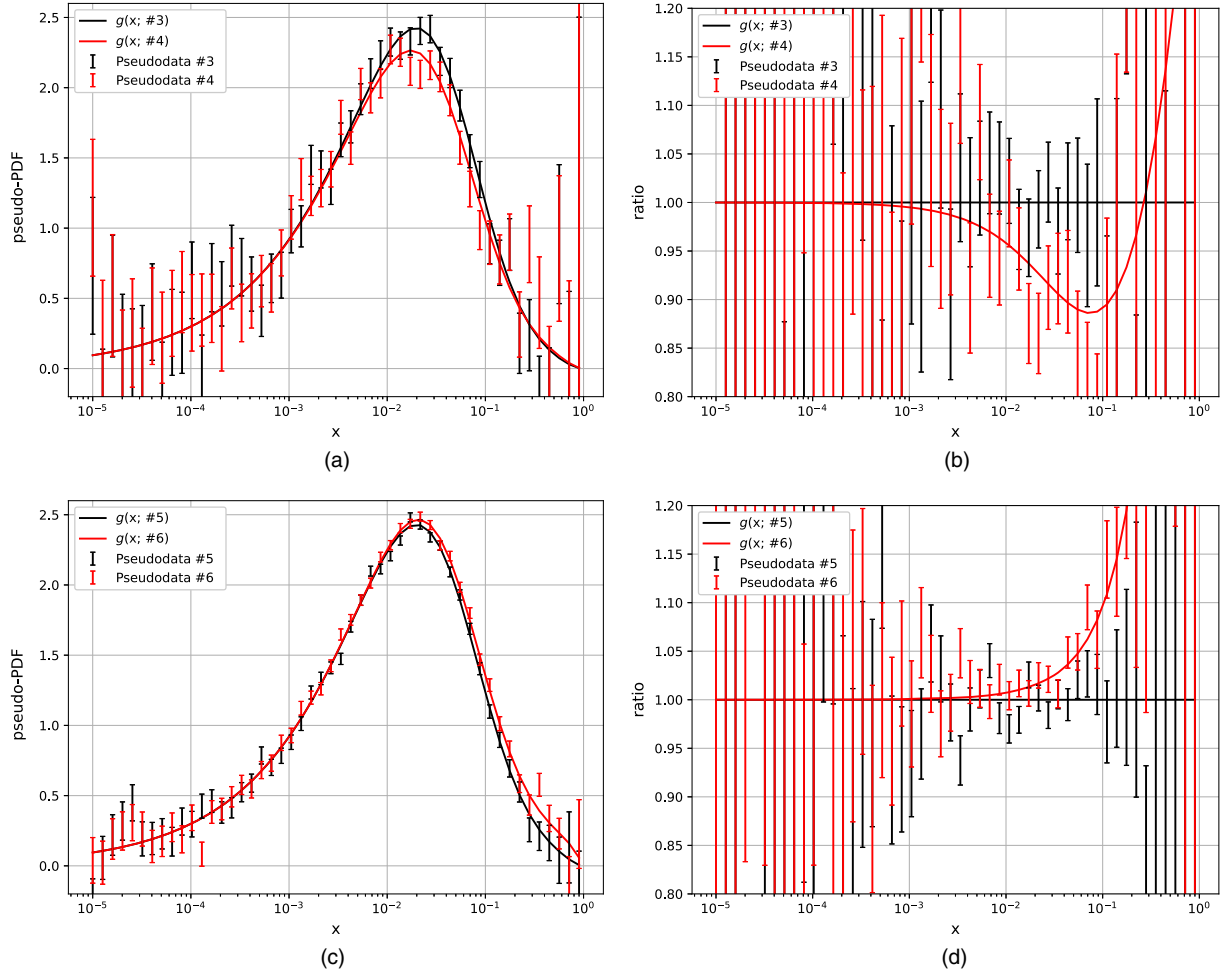


FIG. 3. Similar to Fig. 1, but for Cases 2 and 3. (a) Case-2. (b) Case-2. (c) Case-3. (d) Case-3.

Combining Eqs. (3.16) and (3.17), an estimate of the covariance of the GMM is given by [42]

$$\begin{aligned}
 \text{cov}_{\text{GMM}} &= \sum_{i=1}^K \omega_i \text{cov}_{\text{GMM},i} + \sum_{i=1}^K \omega_i (\mathbb{E}[\theta] - \hat{\theta}_i)^2 \\
 &\simeq \sum_{i=1}^K \omega_i \left(\sum_{j=1}^{N_{\text{pt}}} \left[-\frac{1}{\pi_j} \frac{\partial^2 \mathcal{N}_{ij}}{\partial \theta_i^2} + \frac{\omega_i}{\pi_j^2} \left(\frac{\partial \mathcal{N}_{ij}}{\partial \theta_i} \right)^2 \right] \right)^{-1} + \sum_{i=1}^K \omega_i (\mathbb{E}[\theta] - \hat{\theta}_i)^2.
 \end{aligned} \tag{3.18}$$

The covariance of the GMM shown in the above equation receives two contributions: the first is the weighted sum of covariances for the individual base distributions, and the second term accounts for differences in the mean values of each of the base distributions. When the means of each Gaussian overlap, the second term vanishes and the first term reproduces the covariance in the LS method, Eq. (3.15).

IV. APPLICATION OF THE GMM IN A TOY MODEL OF PDFS

In this section, we compare results obtained by the GMM method, using the likelihood defined in Eq. (3.3), to those

obtained using the LS method. When estimating uncertainty using the LS method, we first diagonalize the Hessian and determine eigenvalue directions. We then determine values of the parameters that satisfy the $\Delta\chi^2 = 1$ condition and define the uncertainty on the fit parameters, which corresponds to the 68% CL interval. Here, we also implement the CTEQ-TEA [1] and MSHT [2] tolerance criteria, which provide a more conservative estimate of uncertainty under the potential imperfections of fits. Below we present the results of three different combinations of pseudodatasets, including the two pseudodatasets that were discussed in Sec. II.

A. Applying GMM to different scenarios

In this section, we consider three cases of combining different pairs of pseudodatasets. Details of how each of these pseudodatasets are generated can be found in Appendix A. Here, we list and describe each of these cases:

- (i) Case 1: here we use pseudodatasets # 1 and # 2, which were introduced in Sec. II. The pseudodatasets are plotted in Fig. 1. This case represents the scenario when there is strong tension between datasets. With this case, we will demonstrate how the GMM can provide a more accurate estimate of uncertainty.
- (ii) Case 2: here we combine pseudodatasets # 3 and # 4, which are shown in Figs. 3(a) and 3(b). These two datasets are similar to datasets # 1 and # 2, except that their uncertainties are tripled and thus have less tension. With this combination, we will be able to show how the GMM behaves when the errors of the datasets are large enough so that there is less tension. We can connect this to the alternative suggestions of dealing with uncertain data by artificially increasing the errors of the datasets [16].
- (iii) Case 3: this case contains pseudodatasets # 5 and #6, which are shown in Figs. 3(c) and 3(d). The tension between these two pseudodatasets resides in the large- x region, where the uncertainty is large. We fit this case with parameters a_2 and a_4 . With this combination, we demonstrate overfitting with the GMM method, and later, in Sec. V B, we demonstrate how to address overfitting.

The best fit obtained by the GMM method for Case 1 is shown in Figs. 4(a) and 4(b) and in Table II. We choose to fit the GMM likelihood function with a mixture of two Gaussian distributions. As a result, the mean values of each Gaussian, can describe the two distinct pseudodatasets well and are in agreement with the fits SG-A and SG-B. We find the weights ω_i to be roughly equal, which is a result of the errors in each of the datasets # 1 and # 2 being almost equal and also having the same number of data points. In Figs. 4(c) and 4(d), the pseudo-PDF uncertainty estimated by the GMM is compared to uncertainties predicted using the LS method. For $10^{-3} < x < 0.1$, the (small) error band of SG-C pseudo-PDF excludes those of SG-A and SG-B fits, and the error band estimated by the GMM is consistent and includes areas covered by both SG-A and SG-B pseudo-PDFs. For $x > 0.1$, where the error in the datasets is large, uncertainties from SG-C and the GMM overlap. Figure 4(e) shows the results of the different fits in the fit parameter space. The $1\text{-}\sigma$ correlation ellipse estimated by SG-C is small, as has been mentioned before, and does not overlap with SG-A and SG-B replicas. On the other hand, the GMM correlation ellipse is large enough and, importantly, also possesses different axes orientations that enable the GMM covariance to overlap with some of the replicas of both SG-A and SG-B. The essential feature we see here

is that the tension in the pseudodatasets lies in the region $10^{-2} < x < 10^{-1}$ [see, for example, Fig. 4(d)]. The GMM expands errors in this region, whereas the χ^2 method with tolerance displays a smaller uncertainty band in this region and expands the error in the region $x > 10^{-1}$. The GMM therefore provides a better estimate of the probability distribution of the theory parameters, given the discrepancies in the data. Note that we have represented the GMM by a single covariance matrix, but this should not be confused with the covariance arising from a single Gaussian, since the shape of the GMM likelihood is multimodal and not unimodal.

For Case 2, shown in Fig. 5, the tension in the dataset has been reduced by increasing the error on the data, so that SG-C describes the combined pseudodataset with $\chi^2_{\#3+\#4}/N_{\text{pt}} = 1.04$. Even for such a mild tension, the fits SG-A and SG-B do not overlap with SG-C. On the other hand, the GMM fit overlaps with both SG-A and SG-B. The difference between the axes orientations of the correlation ellipses between SG-C and the GMM is small. Although the GMM method produces a larger uncertainty, which is more consistent with SG-A and SG-B, the differences with SG-C are small, and we need a way to determine whether we should use SG-C or the GMM method. We will return to a discussion of this in Sec. V B.

The performances for Case 3 of the GMM and LS methods are similar to Case 1. The LS method cannot fit both pseudodatasets # 5 and # 6 well; c.f., the bottom rows of Table II, even though the uncertainties in both datasets are large. The GMM is able to learn about the tension in these datasets and shows a strong inclination to two modes. In terms of the fitted pseudo-PDFs, the GMM in Fig. 6 systematically enlarges the uncertainty and envelopes both datasets # 5 and # 6 simultaneously.

B. The GMM method compared to different tolerance criteria

Various factors that determine the precision of PDFs include theoretical uncertainties, tension between experimental data, and the sampling of PDF functional forms and fitting methodologies, often referred to as epistemic uncertainties; c.f., Ref. [44]. Global PDF analysis groups are aware of the shortcomings of using a fixed tolerance $\Delta\chi^2 = 1$ and have proposed alternate methods. Specifically, CTEQ-TEA [1] and MSHT [2] global PDF analyses adopt certain tolerance criteria, known as the two-tier tolerance and the dynamic tolerance criteria, respectively, on top of the basic Hessian method.

In Sec. IV A, we compared results of the GMM with the usual LS method of determining $1\text{-}\sigma$ contours by setting $\Delta\chi^2 = 1$. In this subsection, we compare the GMM results presented in Sec. IV A to the Hessian results with the CTEQ-TEA and MSHT tolerance criteria.

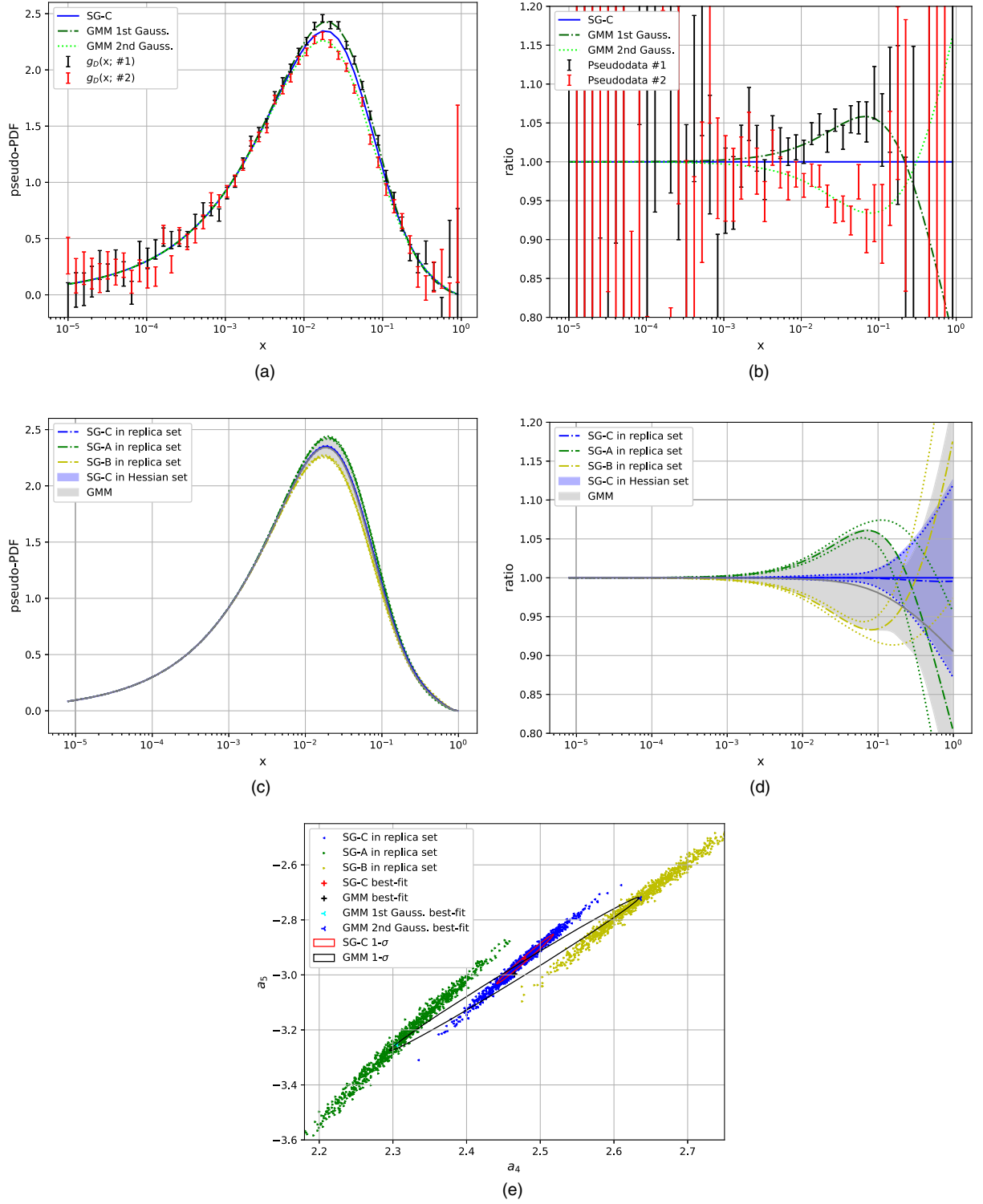


FIG. 4. Comparison of pseudo-PDFs obtained by the GMM method and the LS method for Case 1. Panel (a): pseudodatasets compared to best-fit pseudo-PDFs. Although shown separately, the two best-fit pseudo-PDFs from the GMM method are eventually combined according to weights shown in Table II. Panel (b): similar to panel (a), but presented as ratios of the best fit obtained from the SG-C fit. Panel (c): absolute values of pseudo-PDFs for all SG-A, -B, -C, and GMM fits. Panel (d): similar to panel (c), but presented as ratios to the best fit obtained from the SG-C Hessian set. Panel (e): correlation ellipse from the GMM method compared to replicas from SG-A, -B, and -C fits. The best-fit points are located at their corresponding values in Table II, while the GMM best fit corresponds to the weighted average of best fit of two individual Gaussian distributions as given by Eq. (3.10).

TABLE II. Best-fit values of parameters, and χ^2/N_{pt} values for individual and combined pseudodatasets obtained by the LS method and the GMM fit.

Case 1	Fits	Pseudodata	a_4	a_5	$\chi^2_{\#1}/N_{\text{pt}}$	$\chi^2_{\#2}/N_{\text{pt}}$	$\chi^2_{\#1+\#2}/N_{\text{pt}}$	Weight
	SG-A	# 1	2.32	-3.22	0.88	6.55	3.72	...
	SG-B	# 2	2.63	-2.73	7.00	1.02	4.01	...
	SG-C	# 1 and # 2	2.48	-2.94	2.27	2.56	2.42	...
	GMM Gauss.-1	# 1 and # 2	2.31	-3.25	0.88	6.55	3.72	0.52
	GMM Gauss.-2	# 1 and # 2	2.64	-2.72	7.00	1.02	4.01	0.48
Case 2	Fits	Pseudodata	a_4	a_5	$\chi^2_{\#3}/N_{\text{pt}}$	$\chi^2_{\#4}/N_{\text{pt}}$	$\chi^2_{\#3+\#4}/N_{\text{pt}}$	Weight
	SG-A	# 3	2.20	-3.50	0.76	1.68	1.22	...
	SG-B	# 4	2.74	-2.54	1.61	0.89	1.25	...
	SG-C	# 3 and # 4	2.48	-2.98	0.96	1.11	1.04	...
	GMM Gauss.-1	# 3 and # 4	2.40	-3.10	0.81	1.36	1.09	0.69
	GMM Gauss.-2	# 3 and # 4	2.66	-2.73	1.68	0.90	1.29	0.31
Case 3	Fits	Pseudodata	a_2	a_4	$\chi^2_{\#5}/N_{\text{pt}}$	$\chi^2_{\#6}/N_{\text{pt}}$	$\chi^2_{\#5+\#6}/N_{\text{pt}}$	Weight
	SG-A	# 5	2.27	2.40	1.16	2.33	1.75	...
	SG-B	# 6	1.25	2.41	2.46	0.82	1.64	...
	SG-C	# 5 and # 6	1.60	2.41	1.54	1.14	1.34	...
	GMM Gauss.-1	# 5 and # 6	2.06	2.40	1.19	1.96	1.58	0.58
	GMM Gauss.-2	# 5 and # 6	1.13	2.41	2.58	0.82	1.70	0.42

We remind the reader that CTEQ-TEA Hessian eigenvector PDFs are constructed using a two-tier prescription [45,46] that prevents unexpectedly large increments to either the global χ^2 or to the χ^2_E values of individual experiments. Hence, the CTEQ-TEA PDF error sets are determined using a two-tier (global + dynamic) tolerance (T) that accounts for theory, experimental, parametrization, and methodological uncertainties and results in error bands that are wider than the $\Delta\chi^2 = 1$ criterion. The CTEQ-TEA Hessian uncertainty, at the 68% CL, nominally corresponds to $T^2 \equiv \Delta\chi^2_{\text{total}} \approx 37$, where χ^2_{total} is the sum of the global χ^2 and the so-called Tier 2 (dynamic) penalty [45,46]. The MSHT PDF fits by default apply the “dynamic tolerance” procedure, as detailed in Refs. [2,47,48]. This enlarges the uncertainties beyond the $\Delta\chi^2 = 1$ definition to account for datasets with tension, as well as potential mismatch of data and theory due to imprecise theory or parametrization. The impact of different tolerance criteria on the determination of PDF-induced uncertainties has been examined in Ref. [49]. In this study, we compare the results of applying the CTEQ-TEA and MSHT-like tolerance criteria with $T^2 = 37$ to the ones with $\Delta\chi^2 = 1$.

The Hessian uncertainties based on the single Gaussian LS method with a tolerance criteria of $T^2 = 37$ for all three cases are compared to the GMM method in Figs. 7–9.

- (i) Case 1: both Hessian sets with CTEQ and MSHT tolerance criteria estimate a smaller pseudo-PDF uncertainty for $10^{-3} < x < 0.1$, compared to the GMM. Recall that for Case 1, the GMM produced a more faithful representation of uncertainties and overlapped with SG-A and SG-B. We see that the alternate tolerance criteria is not able to account for the

larger uncertainties in this x range. Further, the uncertainties given by the modified tolerance criteria are much larger than required when $x > 0.1$. The same can also be observed in the fit parameter space. Here, the SG-C correlation ellipses with tolerance cover larger regions than the default Hessian set, but still cannot predict a correlation that spans regions of parameter space that cover portions of SG-A and SG-B.

- (ii) Case 2: as shown in Fig. 8, the larger tolerance leads to an overestimation of pseudo-PDF uncertainty, particularly for $x > 0.1$. In fact, the correlation ellipse for the larger tolerance, seen in the fit parameter space, has regions that are not populated by any of the replica sets.
- (iii) Case 3: as shown in Fig. 9(b), the GMM uncertainty (gray-shaded regions of the pseudo-PDF ratios) shows better agreement with the uncertainty bands of the higher tolerance criteria. However, when comparing results in the fit parameter space, shown in Fig. 9(a), there appears to be significant mismatch between the higher tolerance ellipse and the GMM ellipse.

The tolerance criteria, currently in use by PDF collaborations, was introduced to capture larger uncertainties when there is tension in the fits. However, these methods are still based on the assumption that the likelihood behaves like a single Gaussian and is unimodal. The tolerance criteria are not able to faithfully capture the effect of tension in the data. In some regions, the higher tolerance underestimates the uncertainty, and in others it overestimates the uncertainty. The GMM, on the other hand, is able to provide a more faithful representation of the combined likelihoods of two datasets with tension.

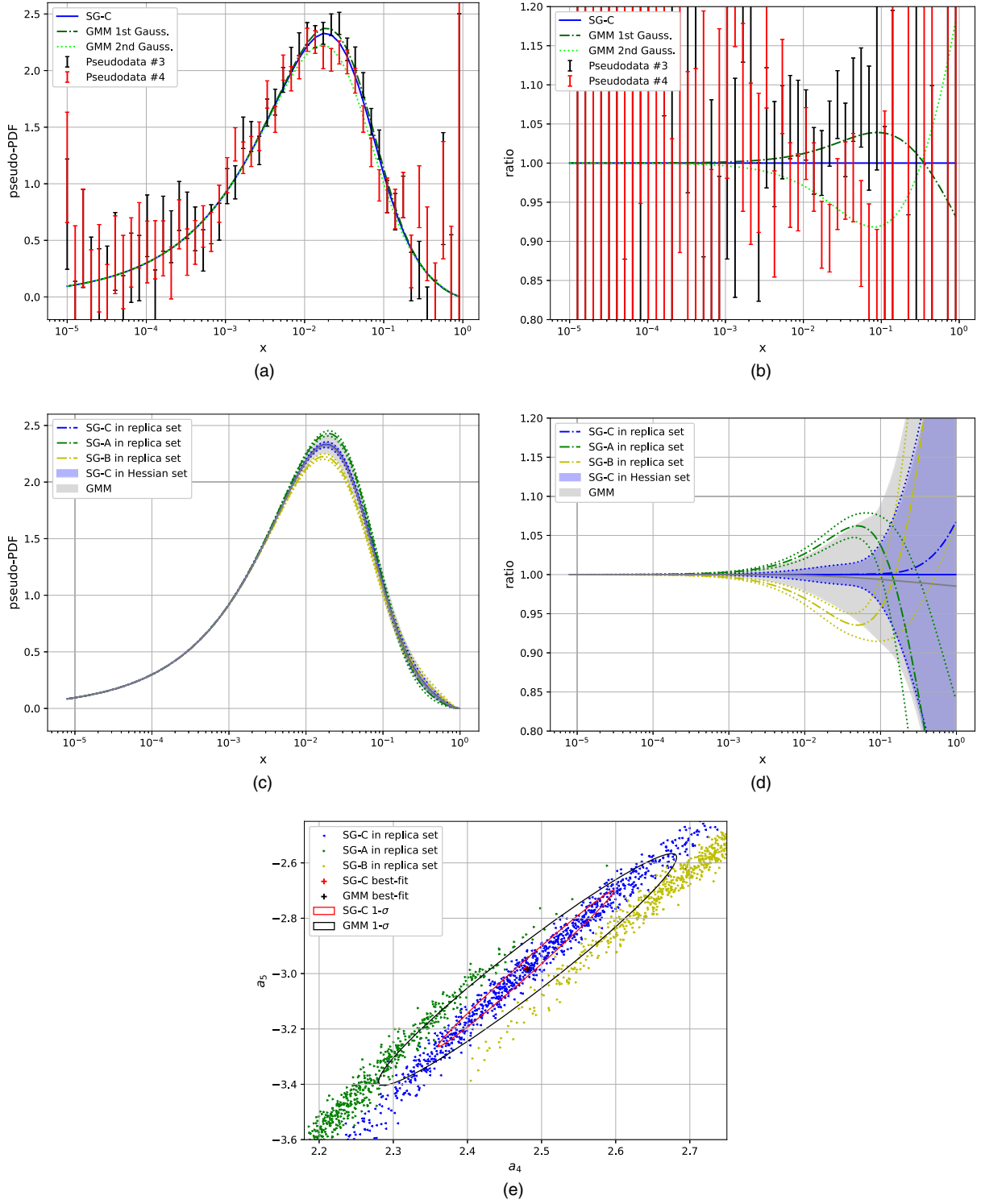


FIG. 5. Similar to Fig. 4, but for Case 2. (a): Pseudodatasets compared to best-fit pseudo-PDFs, (b): Similar to panel (a), but presented as ratios of the best fit obtained from the SG-C fit, (c) absolute values of pseudo-PDFs for all SG-A, -B, -C, and GMM fits, (d): Similar to panel (c), but presented as ratios to the best fit obtained from the SG-C Hessian set, and (e) Correlation ellipse from the GMM method compared to SG-C correlation ellipses, and to replicas from SG-A, -B, and -C fits.

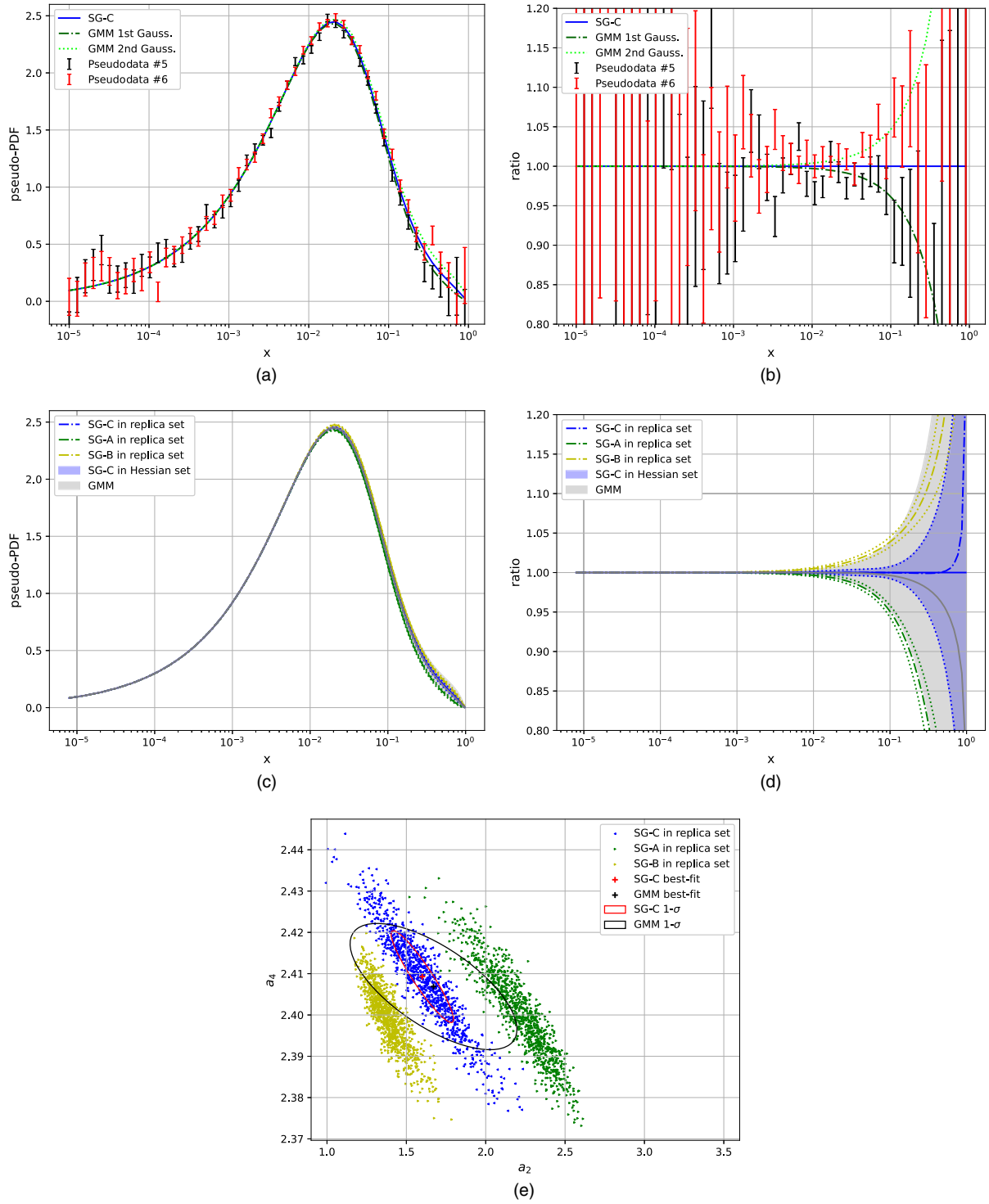


FIG. 6. Similar to Fig. 4, but for Case 3. (a): Pseudodatasets compared to best-fit pseudo-PDFs, (b): Similar to panel (a), but presented as ratios of the best fit obtained from the SG-C fit, (c) absolute values of pseudo-PDFs for all SG-A, -B, -C, and GMM fits, (d): Similar to panel (c), but presented as ratios to the best fit obtained from the SG-C Hessian set, and (e) Correlation ellipse from the GMM method compared to SG-C correlation ellipses, and to replicas from SG-A, -B, and -C fits.

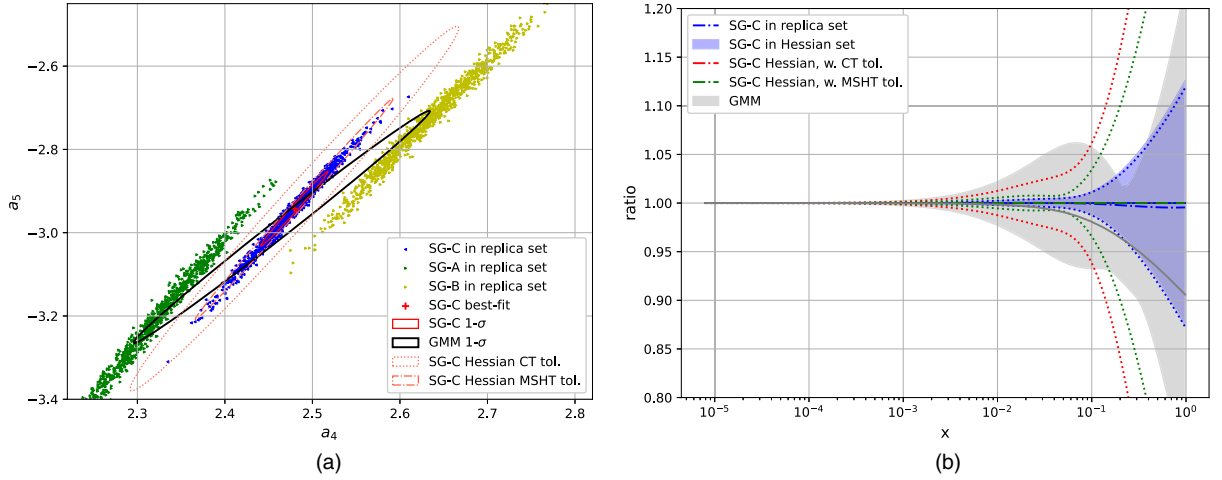


FIG. 7. Pseudo-PDF uncertainty estimated by the GMM method in Case 1 compared to pseudo-PDF uncertainties predicted by the LS method along with the CT and MSHT tolerance criteria. (a) GMM compared to tolerance criteria. (b) Pseudo-PDF ratio.

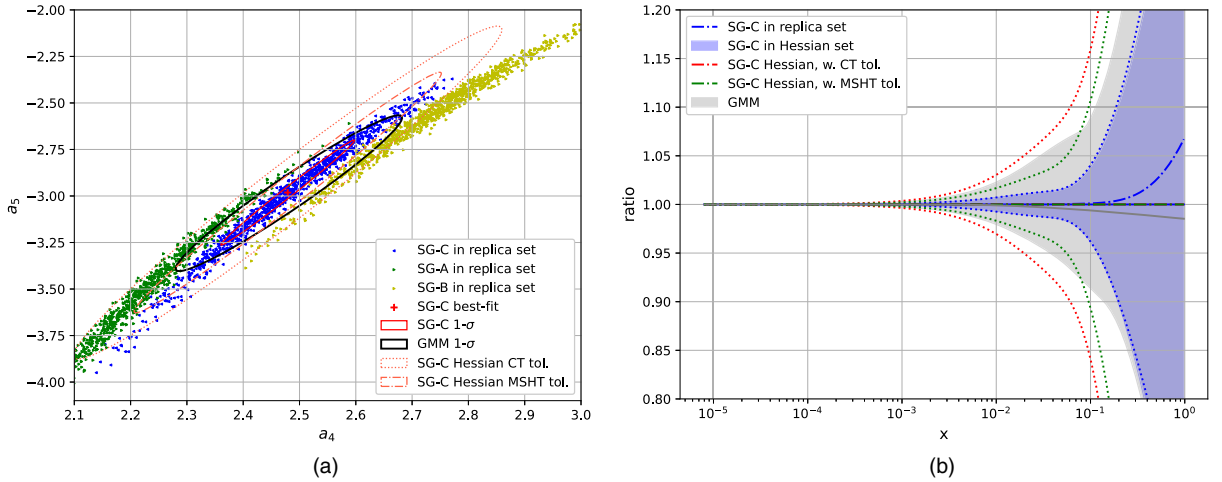


FIG. 8. Similar to Fig. 7, but for Case 2. (a) GMM compared to tolerance criteria. (b) Pseudo-PDF ratio.

V. CONSISTENCY CHECKS FOR THE GMM

In Sec. IV, we demonstrated how the GMM with two base Gaussian distributions provides a better description of the combined likelihood of two pseudodatasets with tension and provides more accurate uncertainties. In this section, we show how the GMM reduces to the usual single Gaussian LS method when data are consistent. Further, as with any present day machine learning technique, when implementing the GMM, one must contend with overfitting. In this section, we also remark on the consequences of overfitting with the GMM.

A. Consistent results from consistent data

In Sec. III, we have shown how the GMM likelihood function reduces to the LS method when the number of Gaussians used in the GMM is $K = 1$. Another possibility for reducing to the usual χ^2 fit is when all base Gaussians

end up with the same mean and covariance, i.e.,

$$\mathbb{E}[\theta] \sim \hat{\theta}_i \sim \hat{\theta}_k, \quad \text{cov}_{\text{GMM},i} \sim \text{cov}_{\text{GMM},k}, \quad \text{and} \\ \mathcal{N}(y_j, \Delta y_j | \theta_i) \sim \mathcal{N}(y_j, \Delta y_j | \theta_k).$$

Consequently, the second term in Eq. (3.18), which accounts for the difference of means, vanishes. In this scenario, the covariance of the GMM reduces to

$$\begin{aligned} \text{cov}_{\text{GMM}} &\simeq \sum_{i=1}^K \omega_i \left(\sum_{j=1}^{N_{\text{pt}}} \left[-\frac{1}{\pi_j} \frac{\partial^2 \mathcal{N}_{ij}}{\partial \theta_i^2} + \frac{\omega_i}{\pi_j^2} \left(\frac{\partial \mathcal{N}_{ij}}{\partial \theta_i} \right)^2 \right] \right)^{-1} \\ &\quad + \sum_{i=1}^K \omega_i (\mathbb{E}[\theta] - \hat{\theta}_i)^2 \\ &\simeq \left(\sum_{j=1}^{N_{\text{pt}}} \frac{1}{\Delta y_j^2} \right)^{-1}, \end{aligned} \quad (5.1)$$

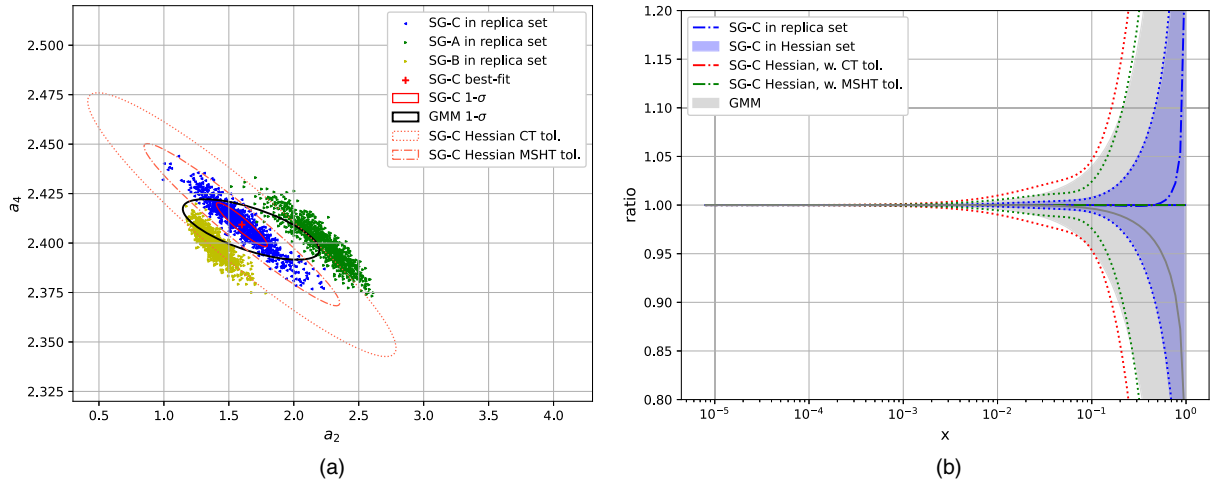


FIG. 9. Similar to Fig. 7, but for Case 3. (a) GMM compared to tolerance criteria. (b) Pseudo-PDF ratio.

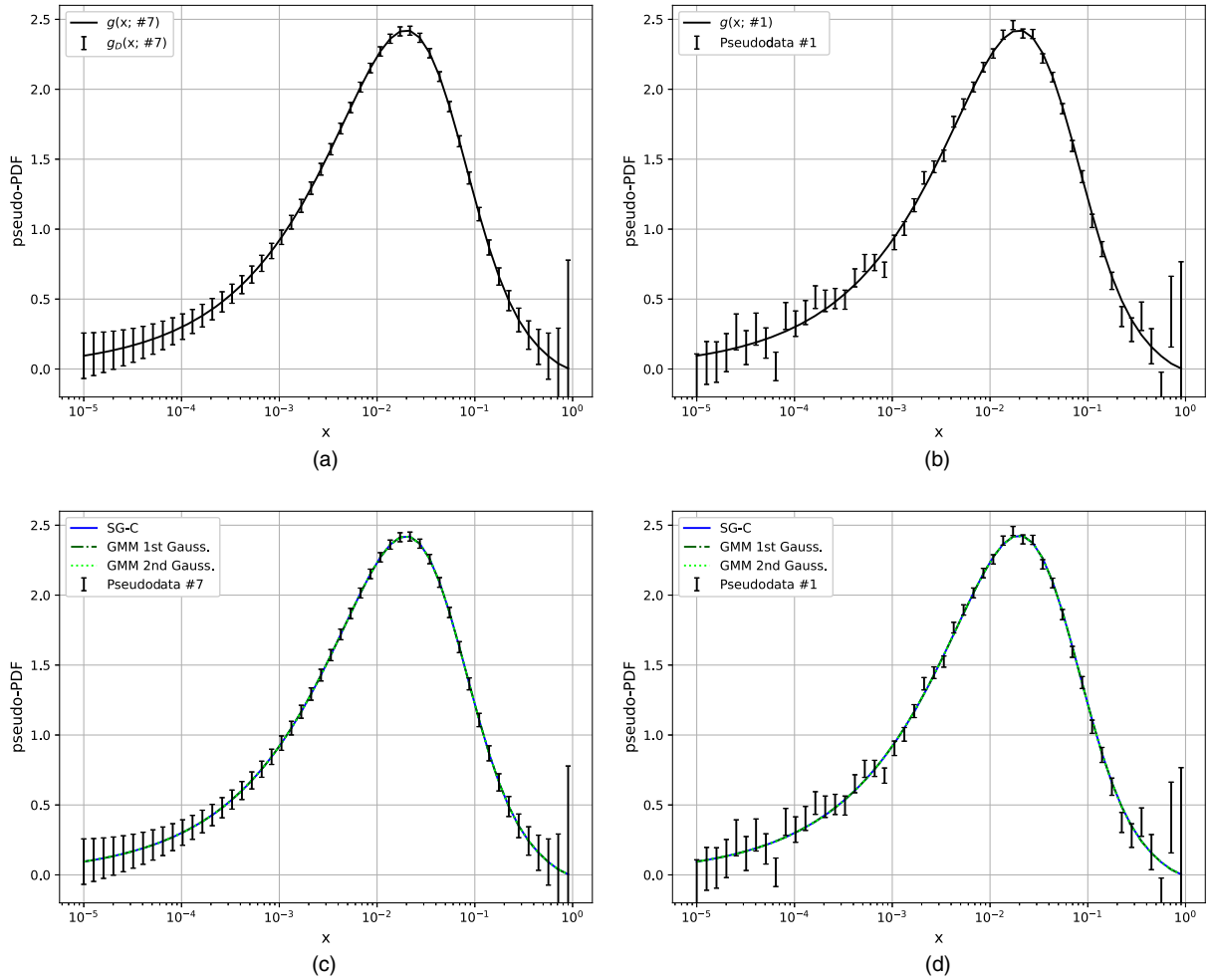


FIG. 10. Upper panels: the truths are compared to pseudodata for Cases 4 and 5. Lower panels: the best-fit pseudo-PDFs are compared to pseudodata for both cases. (a) Truth for Case-4. (b) Truth for Case-5. (c) Best-fits for Case-4. (d) Best-fits for Case-5.

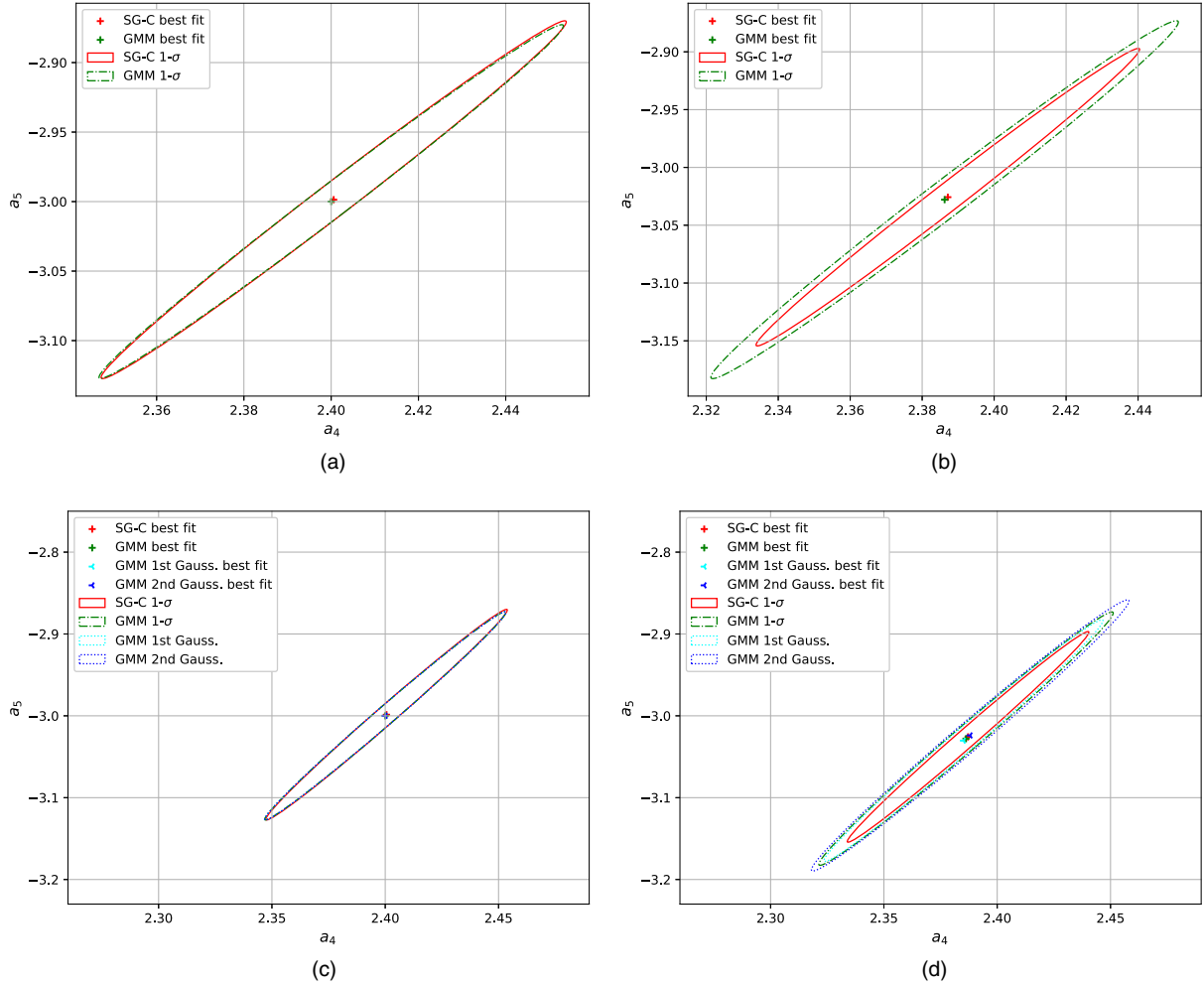


FIG. 11. Upper panels: comparison of covariances obtained by the GMM method and the LS method, for Cases 4 and 5. Lower panels: similar to upper panels, but include the covariances of individual base Gaussian distributions of the GMM. (a) Correlations for Case-4. (b) Correlations for Case-5. (c) Correlations for Case-4. (d) Correlations for Case-5.

which is the covariance found by the LS method in Eq. (3.15). When fitting a consistent dataset, consistency between the GMM method and LS method is expected, as shown analytically in Eq. (5.1). In practice, due to overfitting, we will see that the GMM reduces to the LS method only if no local minima exist in the log-likelihood, i.e., there is only one minimum of the log-likelihood.

Here, we verify Eq. (5.1) by performing two more fits on pseudodatasets, as shown in Fig. 10:

- (i) Case 4: this case concerns a fit to the pseudodata # 7, which is the same as pseudodata # 1, but without any statistical fluctuations. In this case, only a single extremum of the log-likelihood of the GMM exists, i.e., $\theta_i = \theta_k$, and we shall see that numerically as well, the GMM reduces to the LS method fit.
- (ii) Case 5: here, we fit only to a single dataset, namely pseudodataset # 1. Due to overfitting, we will see that although $\theta_i \simeq \theta_k$, the GMM tends to find multiple possible values of θ_k . This in turn implies that the results

from the GMM method are different from the LS method.

We begin by looking at results from the fit in Case 4. The uncertainties predicted by both methods are shown in Figs. 11(a), 11(c), 12(a), and 12(c). The figures indicate, as expected, that the central value of the fit for both the GMM and the LS methods are in good agreement. In Figs. 11(b) and 11(c), we see that the ellipses from the covariance matrix estimated using both the GMM and the LS methods overlap and are identical. On the other hand, for Case 5, we notice in Figs. 11(b) and 11(d) that the covariance predicted by the GMM method cov_{GMM} is marginally larger than the covariance given by the LS method cov_{χ^2} . Figure 11(d) shows that the covariance of individual base Gaussian distributions of the GMM $\text{cov}_{\text{GMM},i}$, which is larger than that obtained by the LS method cov_{χ^2} , and that the optimal values of parameters $\hat{\theta}_i$ for both base Gaussian distributions of the GMM differ by a small amount. The overestimation of uncertainty is a result of overfitting in GMM, where the GMM ends up learning

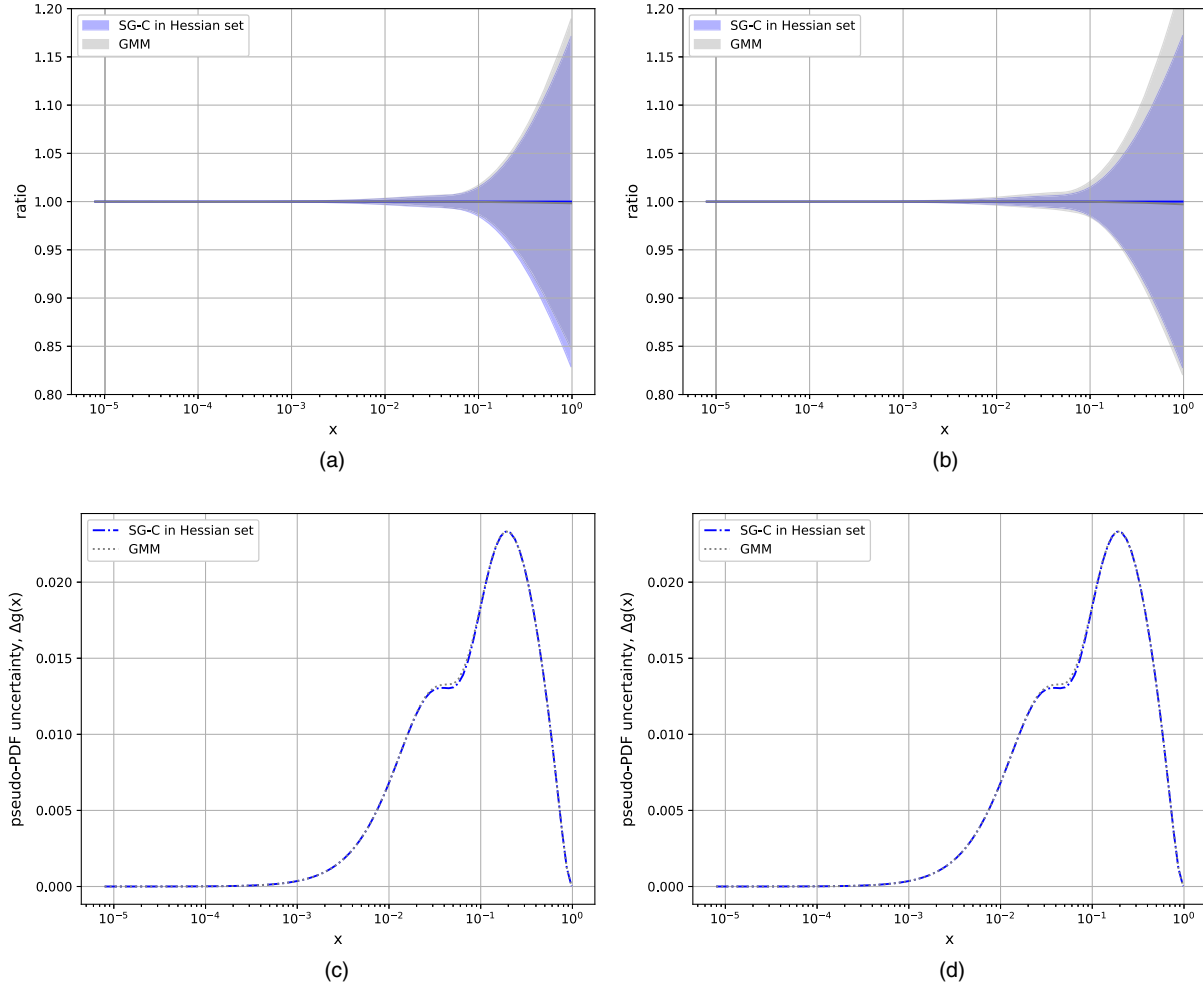


FIG. 12. Upper panels: comparison of pseudo-PDF ratio obtained by both methods for Cases 4 and 5, given the reference being the central value obtained by the LS method. Lower panels: the comparison of the sizes of pseudo-PDF uncertainty obtained by the GMM method and the LS method, for Cases 4 and 5. (a) Pseudo-PDF ratio for Case-4. (b) Pseudo-PDF ratio for Case-5. (c) Size of pseudo-PDF uncertainty for Case-4. (d) Size of pseudo-PDF uncertainty for Case-5.

statistical fluctuations rather than real tension between datasets.

The GMM method looks for tension and tries to categorize data into two or more different Gaussians. In general, for the GMM, overfitting gives rise to larger uncertainties, and in that sense, the GMM always provides a conservative estimate of the uncertainty. However, one would like to provide a precise PDF, which is essential for the physics goals of the community. In order to determine if there is overfitting in the GMM, we describe certain information criteria in the next section that can be used to determine the optimal number of Gaussians, i.e., the value of K . The result of Case 5 suggests that applying the GMM to a consistent dataset will lead to larger uncertainties, which although conservative, should also be avoided.

B. Determining the number of Gaussian modes using information criteria

The results shown in Sec. VA indicate that the GMM method overestimates the uncertainty for consistent

datasets, especially if one includes more base Gaussian distributions than needed. Although the GMM automatically clusters data into sets with tension and those that are consistent, a method to quantify the level of consistency and to determine the optimal number of base Gaussian distributions (K) is required to avoid overfitting and therefore overestimation of uncertainties.

Here, we propose the use of the AIC [50] and BIC [51] as indicators that will help determine optimal values for K , i.e., the number of Gaussians, and as a test of the consistency of fits. The values of AIC and BIC, defined in Eq. (5.2), increase as more parameters are introduced. They also take into account the change in goodness of fit and reduce as the fit improves. Both AIC and BIC are thus able to account for the trade-off between introducing additional parameters and an improvement to the fit. The smallest values of the AIC and BIC indicate to us the best choice of parameters, in this case, the best choice for the value of K . The AIC and BIC are defined as

TABLE III. AIC, BIC, and $-\ln L$ for Cases 1–5 obtained by the GMM with various numbers of base distributions. The lowest values of AIC, BIC, and $-\ln L$ for each case are boldfaced.

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

$$\begin{aligned} \text{AIC} &= N_p \ln N_{\text{pt}} - 2 \ln L|_{\theta=\hat{\theta}}, \\ \text{BIC} &= 2N_p - 2 \ln L|_{\theta=\hat{\theta}}. \end{aligned} \quad (5.2)$$

We count the number of independent parameters N_p for the GMM as follows:

$$N_p = N_{\text{parm}}^{\text{GMM}} = K \times N_{\text{parm}} + (K - 1), \quad (5.3)$$

where N_{parm} is the number of independent parameters for each base Gaussian distribution. The first term in Eq. (5.3) is the total number of fitting parameters over K copies of base distributions. The second term corresponds to the number of independent weights. In our pseudodata study, there are always two free variables in the pseudo-PDF parametrization. Hence the total number of independent parameters for the GMM method in our current fits is

$$N_{\text{parm}}^{\text{GMM}}|_{N_{\text{parm}}=2} = 2K + (K - 1). \quad (5.4)$$

We scan the AIC and BIC of the GMM up to $K = 4$ for all five cases considered in this work, as summarized in Table III. For Cases 1 and 3, both AIC and BIC suggest two base Gaussian distributions are optimal. This confirms the results shown in Sec. IV, particularly in Figs. 4(e) and 6(e), where two base Gaussian distributions can estimate a correlation that covers the two pseudodatasets in tension. The AIC and BIC suggest a single Gaussian distribution suffices for Case 2. In Fig. 5(e), the LS method predicts a correlation, whose slope is close to the correlation estimated with two base Gaussian distributions. This is consistent with the fact that Case 2 has minimal tension due to the large values of data errors. At this point, we would like to point out that the GMM in conjunction with

AIC and BIC can be used to determine how much the errors in the data should be inflated in order to make them consistent, so that we can use the LS method, as suggested in Ref. [16,24,38,52,53].¹⁰

For Cases 4 and 5, the AIC and BIC, as expected, suggest that there is no need of using $K > 1$, since by doing so one is overfitting the data. We reiterate that the consequence of overfitting is to increase our uncertainty and therefore leads to a conservative estimate of uncertainty, i.e., using the optimum value of K provides the most precise PDFs, whereas larger values of K will lead to PDFs with larger uncertainty. We also notice in Table III that a higher number of base distributions K sometimes leads to equal or slightly lower values of $-\ln L$. Thus, if we were to only use $-\ln L$ as the measure of the goodness of fit, and not use the AIC and BIC, we would end up overfitting and choosing larger values of K .

To conclude, although the GMM method is expected to reduce to the LS method for a consistent dataset, one should avoid using it without detailed assessment of the level of tension. The combination of AIC and BIC indicators with the generalized statistical model of the GMM method could effectively reproduce the correct correlation in the presence of inconsistency in data. At the same time, we can determine optimal values of K and provide a formalism to determine more precise PDFs.

VI. CONCLUSION

Accurate PDF uncertainty estimation is a theoretical necessity for precision measurements as well as new physics searches at the LHC and future colliders. Several sources of PDF uncertainty exist, including theoretical uncertainty, experimental systematic errors, and sampling of PDF functional forms and fitting methodologies; see, for example, Ref. [44]. The existence of tension among different experimental datasets can reduce the precision of PDFs and necessitates the use of alternate statistical models that provide more precise estimates of uncertainty. In the presence of tension, expansion of uncertainty is necessary to make fits consistent, and methods to do so have been proposed and used by various PDF fitting groups. The nature and amount of expansion of uncertainty differs for different methodologies available in literature, many of which introduce *ad hoc* criteria.

In this work, we have introduced a novel method to estimate uncertainties for inconsistent datasets that avoids the use of *ad hoc* criteria, checks consistency, and produces expanded uncertainties in a statistically meaningful way. Specifically, for inconsistent data, uncertainties need to be expanded so that the central values of the separate fits to

¹⁰Specifically, one possible measure is to increase the errors in the datasets until the AIC and BIC indicate that $K = 1$ is the optimal value of K . This provides an alternate way to determine the scale factor used in Ref. [38]; c.f., Appendix C.

each consistent dataset lie close to or within the uncertainty bands of the theory parameters for the combined fit. This ensures a more precise representation of the range of possible theory values, given the inconsistent data, and avoids biasing the combined fit. The proposed generalized likelihood, modeled via the Gaussian Mixture Model, is able to appropriately map uncertainties by including the possibility that the likelihood can be either multimodal or unimodal.

In this work, we demonstrated the shortcomings of the LS model when combining datasets that are in tension, using a toy model of PDFs. For LS fitting, the Monte Carlo method and the Hessian method provide a narrow uncertainty band that is not representative of the tension in the datasets. Increasing the uncertainty, either by scaling the uncertainty on each measurement or equivalently, by introducing a static tolerance criteria, does address the problem but is not very precise. For example, using such criteria may increase uncertainty in regions where it is not necessary to do so, whereas the uncertainty in regions of parameter space that do have tension is not suitably adjusted. Hence, the tension is not appropriately represented by the uncertainty bands of the LS fit.

The novel GMM pseudo-PDF fit overcomes these shortcomings of the LS method by parametrizing the likelihood as a mixture of Gaussian distributions and by allowing for a multimodal feature of the likelihood. We have described how this unsupervised machine learning technique can be leveraged to learn about tension in the fits as well as be used to determine precisely the uncertainties on the PDFs. We have shown how to use the AIC and BIC to avoid overfitting, determine the number of Gaussian modes of the GMM, and check for consistency of the fits. Specifically, it is possible to use the GMM for performing the unsupervised machine learning task of clustering the data in order to identify and learn about the tension in the datasets. Moreover, we have demonstrated how the GMM reduces to the usual LS method in the absence of tension. Finally, we have also provided a more intuitive Bayesian interpretation of the GMM method and discussed its connection to ensemble averaging and Bayesian model averaging.

ACKNOWLEDGMENTS

M. Y. thanks Yandong Liu for the discussions. We thank Pavel Nadolsky and Aurore Courtoy for useful discussions and insights. M. Y. has left PKU at the time of this paper being posted. This work was supported in part by the National Science Foundation of China under Grant No. 12075251, and the Natural Science Foundation of Hunan Province of China under Grant No. 2023JJ30496. It is also supported in part by the U.S. National Science Foundation under Grants No. PHY-2310291 and No. PHY-2310497.

APPENDIX A: GENERATING PSEUDODATASETS

We utilize the toy model of pseudo-PDF suggested in Ref. [54] and choose the functional form given in Eq. (2.1) to generate pseudodatasets. For easy reference, we reproduce the function given in Eq. (2.1):

$$g(x) = a_0 x^{a_1} (1-x)^{a_2} e^{x a_3} (1+x e^{a_4})^{a_5}. \quad (\text{A1})$$

Here, $\{a_0, a_1, a_2, a_3, a_4, a_5\}$ are parameters of the function. We generate pseudodatasets by setting values of the parameters (a_i) to certain fixed values. Then, for each fixed value of x , we generate a pseudodata point by assuming the data follow a normal distribution with uncertainty given by

$$\Delta g(x) = \frac{\alpha}{\sqrt{g(x)}}. \quad (\text{A2})$$

Here, the parameter α in Eq. (A2) is used to scale the size of uncertainty. In the region where the magnitude of pseudo-PDF is small, the size of uncertainty is large, mimicking the case of real world data, which has larger statistical uncertainty in regions where event numbers are small. We generate pseudodatasets using

$$g_D(x) = (1 + r \times \Delta g(x))g(x), \quad (\text{A3})$$

where the random variable r follows a Gaussian distribution with unit standard deviation. Although it is possible to include correlated systematic uncertainties or even full likelihoods from experimental data, for simplicity, we treat the pseudo-PDF uncertainty to be uncorrelated and purely statistical.

In this work, we create six pseudodatasets that are summarized in Table II. The pseudodatasets are paired into three cases where we fit the combination of these datasets. The choice of parameter values used to generate each of the pseudodatasets are presented in Table IV. The

TABLE IV. Choice of parameters (a_i) of the function $g(x)$ defined in Eq. (A1) used to generate the six different pseudodatasets # 1–# 6. N_{pt} refers to the number of data points in the set and α determines the error via Eq. (A2). The pseudodatasets are paired and combined into a single fit. These fits are referred to as Cases 1, 2, and 3. We also refer to the figures where a plot of each of these cases can be found.

Case	Pseudodata	α	N_{pt}	a_0	a_1	a_2	a_3	a_4	a_5
Case 1 Fig. 1	# 1	0.05	50	30	0.5	2.4	4.3	2.4	−3.0
	# 2	0.05	50	30	0.5	2.4	4.3	2.6	−2.8
Case 2 Fig. 3(a)	# 3	0.15	50	30	0.5	2.4	4.3	2.4	−3.0
	# 4	0.15	50	30	0.5	2.4	4.3	2.6	−2.8
Case 3 Fig. 3(c)	# 5	0.05	50	30	0.5	2.4	4.3	2.4	−3.0
	# 6	0.05	50	30	0.5	1.4	4.3	2.4	−3.0

figures with plots of the different pseudodatasets are also referenced in the table.

APPENDIX B: BAYESIAN VIEW OF GMM FITS

Here we describe the Bayesian view of the Gaussian Mixture Model as applied to fitting parton distribution functions. We begin by first describing the Bayesian view of the usual χ^2 method before describing the same for the GMM. We follow closely the presentation given in Refs. [16,24].

1. Bayesian view of the χ^2 likelihood

Our objective is to be able to fit theory parameters a_α , $\alpha = \{1, \dots, N_P\}$ given N_D data points D_k , $k = 1, \dots, N_D$. Data that are Gaussian distributed may be represented as

$$D_i = \langle D_i \rangle + \sigma_i \Delta_i. \quad (\text{B1})$$

Here, $\langle D_i \rangle$ can be thought of as the central value of the i th data point, σ_i the uncertainty on the data, and Δ_i is a random variable with a Gaussian distribution with mean 0 and standard deviation of 1. Δ_i represents statistical fluctuations. Note that we have left out the correlated errors to keep the discussion simple and that it is straightforward to reintroduce such errors.

By definition, our data are Gaussian distributed and if f is some function of Δ_i , then its expectation value is

$$\langle f \rangle = (2\pi)^{N_D/2} \int f(\Delta) \prod_{i=1}^{N_D} d\Delta_i \exp\left(-\frac{1}{2} \Delta_i^2\right). \quad (\text{B2})$$

From this, we see immediately that the expectation value of the data is $\langle D_i \rangle$, and it is also straightforward to show by calculating the second moment $\langle D_i D_j \rangle$ that the covariance is

$$C_{ij} = \langle (D_i - \langle D_i \rangle)(D_j - \langle D_j \rangle) \rangle = \delta_{ij} \sigma_i \sigma_j. \quad (\text{B3})$$

In this case, the covariance is diagonal, because we have not considered correlated uncertainties. If g is some function of the data D_i , then it is straightforward to show that the expectation value of g is

$$\begin{aligned} \langle g \rangle &= \frac{1}{\sqrt{(2\pi)^{N_D} \det C}} \int g(D) \prod_{i,j=1}^{N_D} dD_i \\ &\times \exp\left(-\frac{1}{2} (D_i - \langle D_i \rangle)(D_j - \langle D_j \rangle) C_{ij}^{-1}\right). \end{aligned} \quad (\text{B4})$$

If we now make the identification that $\langle D_i \rangle = T_i(a)$, where $T_i(a)$ is the theoretical prediction for the datum D_i , then we get the conditional probability to obtain the experimental results D if the theory is represented by $T(a)$ is

$$\begin{aligned} P(D|T(a)) &= \frac{1}{\sqrt{(2\pi)^{N_D} \det C}} \prod_{i,j=1}^{N_D} dD_i \\ &\times \exp\left(-\frac{1}{2} (D_i - T_i(a))(D_j - T_j(a)) C_{ij}^{-1}\right). \end{aligned} \quad (\text{B5})$$

This can be recast into the more familiar form

$$\begin{aligned} P(D|T(a)) &= \frac{1}{\sqrt{(2\pi)^{N_D} \det C}} dD \\ &\times \exp\left(-\frac{1}{2} \sum_{i,j=1}^{N_D} (D_i - T_i(a))(D_j - T_j(a)) C_{ij}^{-1}\right), \end{aligned} \quad (\text{B6})$$

where the sum represents the usual χ^2 function that is minimized and $dD = \prod_{i=1}^{N_D} dD_i$. $P(D|T(a))$ is the likelihood function. Applying Bayes' theorem, it is possible to determine the conditional probability that the theory $T(a)$ is correct given the data D , as follows:

$$P(T(a)|D) = \frac{P(D|T(a))P(T(a))}{P(D)}. \quad (\text{B7})$$

Knowledge of $P(D)$ and $P(T(a))$ is limited, making determination of this probability difficult. On the other hand, the typical fitting problem requires that we make an optimal choice of parameters that maximize $P(T(a)|D)$. This is easily achieved through the ratio of likelihoods. For example, consider the ratio of the conditional probabilities for two different choices of the theory parameters a_1 and a_2 :

$$\begin{aligned} \frac{P(T(a_1)|D)}{P(T(a_2)|D)} &= \frac{P(D|T(a_1))P(T(a_1))}{P(D|T(a_2))P(T(a_2))} \\ &= \exp\left(-\frac{\chi^2(D, a_1) - \chi^2(D, a_2)}{2}\right) \\ &\times \left(\frac{P(T(a_1))}{P(T(a_2))}\right). \end{aligned} \quad (\text{B8})$$

Here, $P(T(a_1))/P(T(a_2))$ is the ratio of our prior probabilities that we begin with and which is updated as we collect more data.¹¹

2. Hierarchical models and Bayesian view of the GMM

Consider the scenario where we collect data from K different experiments. The data from each of these experiments are represented as

¹¹See Ref. [16] for a more detailed discussion on updating prior and posterior probabilities.

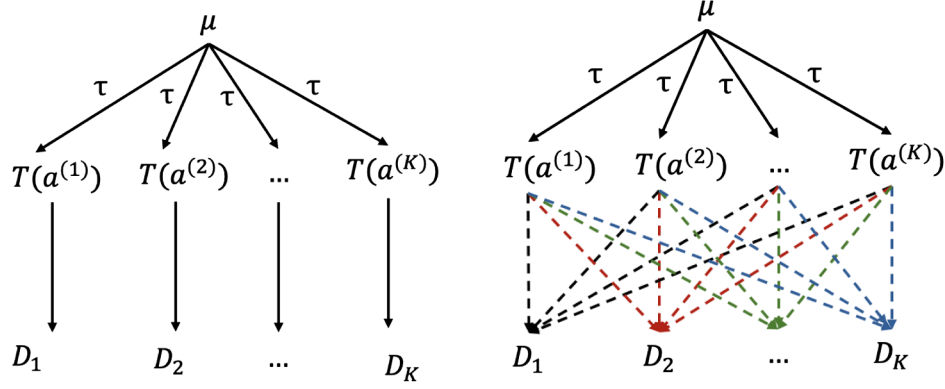


FIG. 13. Left: hierarchical model, as proposed in Ref. [24]. Right: GMM model. Here, μ and τ correspond to hyperparameters from which the theory models ($T(a^{(k)})$) are drawn.

$$D_i^{(k)} = \langle D_i^{(k)} \rangle + \sigma_i^{(k)} \Delta_i^{(k)} = T_i(a^{(k)}) + \sigma_i^{(k)} \Delta_i^{(k)}. \quad (\text{B9})$$

Here $k = \{1, 2, \dots, K\}$. Note we have assumed, for simplicity, that the data between experiments is uncorrelated. This is not a necessary assumption and can be relaxed in general cases. Suppose each dataset from each experiment produces a different central value $\langle D_i^{(k)} \rangle$, indicating a different preferred theory $T_i(a^{(k)})$ for each experiment. In reality, such differences may be due to an underestimation of experimental uncertainties or the use of an inaccurate theoretical model. Therefore, one should investigate the theory and data more rigorously to identify the source of these discrepancies. However, this is often difficult or even impossible to do with particle physics data. Instead, we can use the Gaussian Mixture Model to estimate uncertainty, given the gaps in our understanding of the source of discrepancy between experiments. For example, factorization in QCD requires there to be only one set of theory parameters of parton distribution functions. The GMM allows for the possibility that several values of our theory parameters $a^{(k)}$ can exist simultaneously. We are not abandoning QCD factorization; rather, we introduce multiple possibilities of the same theory fit parameters in order to capture the discrepancies present in the data. As mentioned earlier, such discrepancies could arise due to imprecise theoretical calculations and/or unaccounted-for experimental biases. The GMM is able to provide an estimate of the probability distribution of the theory, given these unresolved discrepancies in the data.

A similar approach to the GMM proposed in this work has been used in Ref. [24]. As shown in the diagram on the left in Fig. 13, here each datum (D_k) is viewed as measuring different theory model parameters $T(a^{(k)})$, and each $T(a^{(k)})$ originates from some parent distribution governed by hyperparameters μ and τ . The probability $P(T(a^{(k)}))$ is then given by

$$P(T(a^{(k)})) = \int d\mu d\tau P(T(a^{(k)})|\mu, \tau) p(\mu, \tau) \equiv w_k. \quad (\text{B10})$$

Here, we introduce $w_k = P(T(a^{(k)}))$ so that $\sum_{k=1}^K w_k = 1$. The diagram on the right in Fig. 13 shows the approach used in the GMM. We assume that each datum $D^{(i)}$ could arise due to a random draw from each of the possible theories $T(a^{(k)})$, and so the probability of measuring datum $D^{(i)}$ given the theory $T(a^{(k)})$ is

$$\begin{aligned} P(D_i|T(a^{(k)}))P(T(a^{(k)})) &= w_k P(D_i|T(a^{(k)})) \\ &= P(T(a^{(k)})|D_i)P(D_i). \end{aligned} \quad (\text{B11})$$

Here, we have used Bayes' theorem in the last equality. Marginalizing over the different possible theories $T(a^{(k)})$, we get

$$\left(\sum_{k=1}^K P(T(a^{(k)})|D_i) \right) P(D_i) = \sum_{k=1}^K w_k P(D_i|T(a^{(k)})). \quad (\text{B12})$$

The term on the left-hand side tells us the probability of getting all $T(a^{(k)})$ given the datum D_i , whereas the right-hand side is the probability of drawing the datum D_i given the set of $T(a^{(k)})$. The objective for a fitting problem using the GMM is to determine the choices of $T(a^{(k)})$ and w_k that maximize the probability on the left-hand side. Thus, the right-hand side of the equation corresponds to the modified likelihood of the GMM. Specifically, assuming Gaussian probability distributions, we have

$$\prod_{i=1}^{N_D} \left(\sum_{k=1}^K P(T(a^{(k)})|D_i) \right) \propto \prod_{i=1}^{N_D} \left(\sum_{k=1}^K w_k \mathcal{N}(D_i|T(a^{(k)}), \sigma_i) \right), \quad (\text{B13})$$

which is the likelihood function used in the Gaussian Mixture Model; c.f., Eq. (3.1). Here $\mathcal{N}(D_i|T(a^{(k)}), \sigma_i)$ is the normal distribution with mean $T(a^{(k)})$ and standard deviation σ_i . Note that in Fig 13, it appears as though for N_D data points one must choose $K = N_D$. This is not

required in the GMM, and in fact, K is a hyperparameter that can be determined with the help of information criteria; c.f., Sec. V B. Typically, if all the data are consistent, we find $K = 1$, which reduces to the usual χ^2 likelihood function. Hence, in Eq. (B13), K and N_D are different with $K \ll N_D$ in typical applications of the GMM.

One major advantage of using the GMM to construct the Bayesian hierarchical model is in overcoming a combinatorial problem of partitioning the data into different sets. The GMM partitions the data into consistent datasets, and in the limit of large tension, the partitioned datasets approximate the usual hierarchical model shown in the left of Fig. 13, since the probability of generating the dataset D_i from $(T(a^{(j)}))$ is large when $i = j$ and negligibly small when $i \neq j$. A further advantage is that it is not necessary to assume a hyperprior distribution $p(\mu, \tau)$ since the weights w_k are determined through likelihood extremization. Starting with the model in the left of Fig. 13 would necessarily require us to specify a hyperprior distribution or determine a way to partition the datasets. Both of these concerns are directly addressed by the use of the GMM. Finally, we note in passing the interesting possibility of introducing and modifying the hyperprior distribution in the GMM, which can be achieved by modifying the weights, which could lead to different interpretation of the weights and the statistical model in use.

3. A note on Bayesian model averaging and ensemble learning

In Appendix B 2, we described our approach as a hierarchical model assuming the weights w_k are determined through some parent distribution $p(\mu, \tau)$. However, in practice we do not need to know the parent distribution and, in fact, the weights w_k are determined through a maximum likelihood estimate of the likelihood function in Eq. (B13). The use of the GMM makes the approach taken here similar to the method of ensemble learning used frequently in predictive machine learning tasks such as weather forecasting [42]. The main idea in ensemble learning is to define a combined probability distribution as a convex combination of a basis of models, as follows:

$$\begin{aligned} P(y|D, w_1 \dots w_K, T(a^{(1)}) \dots T(a^{(K)})) \\ = \sum_{k=1}^K w_k P(y|D, T(a^{(k)})), \end{aligned} \quad (\text{B14})$$

where y is the quantity to be predicted. While several algorithms, such as stacking [55,56], have been proposed to estimate the weights w_k , for our purpose, since each of our models $T(a^{(k)})$ are equally complex, a straightforward maximization of the likelihood suffices.

This method is sometimes referred to as the committee method [25] or as a method of model combination [26]. In the statistics literature, a distinction is made between

Bayesian model combination and another related and popular method known as Bayesian model averaging [27,29,30,57]. The former is a convex combination of a basis of models that produces a single combined model that describes the probability distribution of data, such that data could arise from any of the theoretical model distributions. On the other hand, Bayesian model averaging assumes that there is just one model $T(a^{(k)})$ responsible for generating the data and w_k represents the probability for it to be the best model. For a more in-depth discussion of these differences, see, for example, Refs. [58,59].¹² If our goal is to find the best model as is done in Bayesian model averaging, then we may interpret the w_k as the posterior probability that $T(a^{(k)})$ is the best model to describe the data. Then what we compute through the use of Eq. (B13) is in fact an implementation of Bayesian model averaging, see for example Ref. [42]. The formulas we provide for the estimate of the mean and variance in Eqs. (3.10) and (3.18), respectively, coincide with the formulas used for these quantities in Bayesian model averaging [29,42]. Further, one could use values of w_k to determine Bayes factors [29] and even perform model selection to select the best model out of the mixture of models. Here, we are more interested in a determination of the uncertainty and do not undertake such a program. Importantly, model averaging and combination have been shown to not only provide better predictive power of machine learning tasks, but also provide better estimates of uncertainty [25].

APPENDIX C: A SIMPLE ONE-DIMENSIONAL EXAMPLE OF W-BOSON MASS COMBINATION

The purpose of this appendix is to demonstrate the use and characteristics of the GMM method with the help of a simple one-dimensional example. Note that a one-dimensional fit does not allow us to demonstrate one major advantage that the GMM provides in estimating uncertainties through appropriate modification of eigenvector directions of the Hessian. Nevertheless, the example provided here is simple and uses real world data.

Consider the measurement of the W -boson mass by various experiments around the world. The measured values of the W -boson mass are shown in Table V. CDF and ATLAS recently provided updated measurements of the W -boson mass. The reported value from CDF is in tension with the ATLAS measurement. In this situation, one might argue that one or both of the analyses are inaccurate. Alternately, either or both experiments could be imprecise, and have underestimated uncertainties.¹³ Here,

¹²For more discussion on the use of mixture models as a way to implement Bayesian model averaging and combination, see, for example, Refs. [60,61].

¹³Understanding the source of this discrepancy is a complicated task and studies in this regard are still being undertaken [62].

TABLE V. Measured values of the W -boson mass.

Experiment	W -boson mass (GeV)	$\pm$ Error (GeV)
D0-I [63]	80.483	0.084
CDF-I [64]	80.433	0.079
LEP [65]	80.376	0.033
D0-II [66]	80.375	0.023
LHCB [67]	80.354	0.032
CDF-II [9]	80.4335	0.0094
ATLAS23 [68]	80.36	0.016

we take an agnostic view on the source of the tension between experiments, which could likely be a result of theoretical inadequacies as well. In this scenario, when we are uncertain about both the source of uncertainty as well as its estimate, how should we combine these disparate measurements in a meaningful way? In other words, in the presence of such tension in the data, how should the measurements be combined to report a mean and uncertainty (point estimates) of the W -boson mass? In the following subsections, we discuss various possibilities to perform such a combination. Although it is possible to incorporate any correlated uncertainties between the measurements, to keep our presentation simple, we treat each measurement as independent and ignore any correlated uncertainties.

1. Least squares fit

In the standard χ^2 approach, a Gaussian likelihood function from each experiment is multiplied to provide a combined likelihood. Taking the logarithm of the likelihood yields the usual χ^2 log-likelihood up to constant terms, as follows:

$$\chi^2 = \sum_j \frac{(y_j - m_W)^2}{\Delta y_j^2}. \quad (\text{C1})$$

Here, y_j corresponds to the measured values of the W -boson mass and Δy_j the uncertainty on the measurement, which can be found in Table V. By using the χ^2 , we have implicitly assumed that the likelihood from each measurement is Gaussian with a standard deviation given by the uncertainty.¹⁴ Our goal is to learn the distribution of the random variable m_W . Although full information about the distribution is available in the log-likelihood, it is usually useful to report the characteristics of the distribution through point estimates such as moments of the distribution. Typically, that means reporting the mean or expectation value ($\bar{m}_W$) as well as the variance (σ_{m_W}), which is sufficient to fully specify a Gaussian distribution, as follows:

¹⁴Alternately, it is possible to take a product of the full experimental likelihoods. However, the assumption that each experimental likelihood is Gaussian usually works well.

$$\pi(x; \bar{m}_W, \sigma_{m_W}) = \frac{1}{\sigma_{m_W} \sqrt{2\pi}} e^{-\frac{1}{2} \left(\frac{x - \bar{m}_W}{\sigma_{m_W}} \right)^2}. \quad (\text{C2})$$

The mean, $\bar{m}_W$, is determined by minimizing the χ^2 function and the variance can be found from the Hessian. Alternately, since the m_W is a random variable, it follows that the log-likelihood follows a χ^2 distribution. Thus, the variance, σ_{m_W} , can be obtained by varying the χ^2 away from the minimum up to $\Delta\chi^2 = 1$.

Minimizing the χ^2 function, we find the best fit value of the W -boson to be

$$\bar{m}_W|_{\chi^2} = 80.4065 \pm 0.0072 \text{ GeV}. \quad (\text{C3})$$

The uncertainty of 7.2 MeV is determined with the condition $\Delta\chi^2 = 1$, which in this case is identical to the estimate of the standard error using the inverse of the information matrix or the Hessian matrix, which is related to the covariance of the combined likelihood of m_W . In this simple example, rather than minimizing numerically, it is possible to derive exact formulas, where the mean and uncertainty of the m_W combination is given by

$$\bar{m}_W|_{\chi^2} = \frac{\sum_j \frac{y_j}{\Delta y_j^2}}{\frac{1}{\sigma_{m_W}^2}}, \quad \text{with} \quad \frac{1}{\sigma_{m_W}^2} = \sum_j \frac{1}{\Delta y_j^2}. \quad (\text{C4})$$

As a result of the tension between the ATLAS and CDF measurements, the fit turns out to be quite poor, with $\chi^2/\text{d.o.f} \simeq 3.8$.

One way to proceed with an alternative combination is to inflate the experimental uncertainties by a scale factor (S_{PDG}), as proposed by the Particle Data Group [69], using the conditions described below.

- (1) If the $\chi^2/\text{d.o.f} < 1$, the results of the fit are accepted and there is no scaling.
- (2) If the $\chi^2/\text{d.o.f} \gg 1$, either a combined mean is not produced or an educated guess of the uncertainty is made.
- (3) If the $\chi^2/\text{d.o.f} > 1$, then:
 - (a) If all errors are comparable, then all errors are rescaled by a common scale factor $S_{\text{PDG}} = \sqrt{\chi^2/\text{d.o.f}}$. This, by definition, forces the $\chi^2/\text{d.o.f}$ to approach 1. Note, for the simple examples used here, this is equivalent to introducing a global tolerance criteria with $T = S_{\text{PDG}}$.
 - (b) If some of the individual experimental uncertainties are significantly smaller than others, then S_{PDG} is calculated using only the most precise experiments. The choice of cutoff to include or exclude experiments by their precision is *ad hoc*. PDG uses an arbitrary choice of only scaling those experimental uncertainties if the uncertainty for a measurement is less than $3N^{1/2}\bar{\sigma}$,

where N is the number of measurements, $\bar{\sigma}$ is the unscaled uncertainty computed from Eq. (C4).

This idea has been further explored and elucidated in both the Bayesian and frequentist approaches [24,52,53]. Using the above recommendations, we find for the dataset in Table V, all uncertainties are rescaled by a factor of $S_{\text{PDG}} \simeq \sqrt{3.8} \simeq 1.95$. This is equivalent to introducing a tolerance of $T \simeq 1.95$, without rescaling uncertainties. Applying the second condition to the fit involving only the CDF-II and ATLAS data, we find

$$\bar{m}_W|_{\chi^2} = 80.4065 \pm 0.0141 \text{ GeV}, \quad (\text{C5})$$

with $\chi^2/\text{d.o.f} = 1$. The uncertainty, by definition, also increases by a factor of S_{PDG} .

A more sophisticated combination of the global measurement of the W -boson mass [70] reveals that the CDF measurement is incompatible with the rest of the measurements, and dropping this measurement in the combination improves the probability of compatibility by about $\sim 90\%$.¹⁵ A refit after leaving out the CDF-II measurement yields

$$\bar{m}_W|_{\chi^2}^{\text{no CDF-II}} = 80.3683 \pm 0.0112, \quad (\text{C6})$$

with $\chi^2/\text{d.o.f} \simeq 0.62$.

In this situation of combining the W -boson mass, we not only have ample data to compare, but also theoretical prediction of $m_W|_{\text{SM}} = 80.355 \pm 0.006 \text{ GeV}$ [71,72]. On the other hand, PDF fits do not have strong theoretical information, and often data can be sparse so that a determination of compatibility is not straightforward and in some cases not even possible. To draw an analogy with the W -boson mass combination, let us consider the hypothetical situation in which we have only the CDF-II and ATLAS measurements of the W -boson mass and no theoretical insight on this quantity, as is the case for nonperturbative parton distribution functions.¹⁶ If we redo the fits with only the CDF-II and ATLAS data, we get a very poor fit with $\chi^2/\text{d.o.f} \simeq 15.7$ and

$$\bar{m}_W|_{\chi^2}^{\text{ATLAS23+CDF-II}} = 80.4146 \pm 0.0081 \text{ GeV}. \quad (\text{C7})$$

For such a poor fit, PDG prescription suggests either not doing a combination at all or making an educated guess for the uncertainty. If we were to rescale the uncertainties, the combined uncertainty for this fit would increase approximately by a factor of 4 to about ± 0.0321 . Note that, in spite of the large fourfold increase in uncertainty, the ATLAS measurements central value does not fall within the $1\text{-}\sigma$ uncertainty of this fit; neither does the theoretical prediction of

$m_W|_{\text{SM}} = 80.355 \pm 0.006 \text{ GeV}$ [71,72]. Hence, as suggested by PDG, an educated guess for the uncertainty is necessary to account for the bias introduced by the combination. Rather than making an educated guess, we can use the GMM method proposed in this work to determine the uncertainty.

An important question to ask is why the uncertainty should increase as more data are collected? This appears to be contrary to what we learn in statistics, where the statistical uncertainty reduces with increasing sample size. One way to try and understand this is to recall that experiments usually need to unbiased their measurements through the use of some procedure. For example, proper calibration is necessary to create a map ($f: \theta_M \rightarrow \theta_{\text{Th}}$) that translates the numbers read out from the experimental device (θ_M), such as a data acquisition system, to the quantity of interest θ_{Th} . The procedure to determine this map is usually statistical in nature and introduces systematic uncertainties. When fitting to a consistent dataset, we assume that the map ($f: \theta_M \rightarrow \theta_{\text{Th}}$) has been accurately determined and any uncertainty is accounted for by the systematic uncertainties provided by the experiment. If a second experiment produces data that are in tension with the first one, then this provides evidence that the map $f: \theta_M \rightarrow \theta_{\text{Th}}$ has not been determined accurately for either or both experiments. In this situation, our initial assumption that the maps have been determined accurately is incorrect, and we must incorporate a new uncertainty that accounts for the presence of a systematic bias. One way to do this is to marginalize over the space of possible maps ($f: \theta_M \rightarrow \theta_{\text{Th}}$), or in other words, we must marginalize over the possible distributions of these maps given the experimental data. The hyperprior distribution of the hierarchical model ($p(\mu, \tau)$) in Eq. (B10) approximates the possible distribution of such maps. When all experiments are consistent, then it is consistent to use a Dirac delta distribution for the hyperprior. On the other hand, information about the possible inaccuracy of such maps changes the hyperprior away from a Dirac delta distribution, which in turn necessarily increases the uncertainty. This can be avoided in the ideal situation if there is a way to choose between experiments, or there is a way to reanalyze and determine the maps more accurately.

We reiterate that, in this simple example, we are able to compare with several other W -boson mass measurements as well as one from electroweak fits, to understand the compatibility of various experimental measurements. On the other hand, parton distribution functions do not have a theoretical counterpart to be compared to and several portions of the fit might also have very sparse amounts of data. Hence, uncertainties on the PDFs should be more representative of the data from which they are extracted. In the next subsection, we describe how to improve the probability distribution on measured quantities and hence provide more precise estimates of uncertainty by using a mixture model and allowing for the likelihood to be multimodal.

¹⁵The probability of compatibility is estimated using the χ^2 distribution. See Ref. [70] for more details.

¹⁶Recently, there have been some attempts to determine the PDFs using lattice QCD; see, for example, Ref. [73].

2. GMM fit

Here, we apply the GMM method to the simple case of W -boson mass combination. We begin by defining the likelihood as follows:

$$\begin{aligned} -\log L &= -\log \left(\prod_{j=1}^{N_{\text{pt}}} L(y_j, \Delta y_j | \vec{\theta}) \right) \\ &= -\sum_{j=1}^{N_{\text{pt}}} \log \left(\sum_{i=1}^K \omega_i \mathcal{N}(y_j, \Delta y_j | \theta_i) \right) \\ &= -\sum_{j=1}^{N_{\text{pt}}} \log \left(\sum_{i=1}^K \frac{\omega_i}{\sqrt{2\pi}\Delta y_j} \exp \left[-\frac{1}{2} \left(\frac{y_j - y_j(\theta_i)}{\Delta y_j} \right)^2 \right] \right). \end{aligned} \quad (\text{C8})$$

By doing so, we have introduced K different random variables θ_i , and we would like to learn about the probability distribution of these K random variables. While information about the probability distribution of the K variables is available in the likelihood function above, we would like to report the point estimates of the distribution, namely the mean and variance of the distribution.¹⁷ The best fit is determined through maximum likelihood estimation by varying and finding the K different values of m_W as well as the weights ω_i of each Gaussian that maximize the full log-likelihood. Once we have minimized the loss function, determination of an estimate of the mean is straightforward and simply given by

$$\bar{m}_W|_{\text{GMM}} = \mathbb{E}[\theta] = \sum_{i=1}^K \omega_i \hat{\theta}_i. \quad (\text{C9})$$

Here, $\hat{\theta}_i$ are the K different values of m_W that minimize the loss function defined in Eq. (C8). Determination of the variance requires more work. We begin with the ansatz that the probability distribution of m_W is, to a good approximation, given by a Gaussian mixture model of the form

$$\pi(x | \hat{\theta}_i, \sigma_i) = \sum_i^K \frac{\omega_i}{\sigma_i \sqrt{2\pi}} e^{-\frac{1}{2} \left(\frac{x - \hat{\theta}_i}{\sigma_i} \right)^2}. \quad (\text{C10})$$

The variance for the probability distribution function defined above is

$$\sigma^2 = \text{cov}_{\text{GMM}} = \sum_{i=1}^K \omega_i \sigma_i^2 + \sum_{i=1}^K \omega_i (\mathbb{E}[\theta] - \hat{\theta}_i)^2. \quad (\text{C11})$$

¹⁷Other measures of uncertainty can be used; however, using the mean and variance makes it straightforward to compare with χ^2 fits.

TABLE VI. Mean ($\bar{m}_W$), uncertainty (σ), AIC, and BIC values for different values of K . The numbers in boldface correspond to the smallest values of AIC and BIC.

K	$\bar{m}_W$	σ	AIC	BIC
1	80.4065	0.0072	-11.01	-10.96
2	80.3850	0.03081	-22.89	-22.73
3	80.3850	0.03091	-19.00	-18.73
4	80.3850	0.03091	-15.10	-14.73
5	80.3850	0.03095	-11.21	-10.73

While the second term in the equation above is already determined with the help of Eq. (C9), the first term requires knowledge of the variance (σ_i^2) for each Gaussian. When using the expectation-maximization algorithm, the individual variances are calculated automatically. Here, we approximate the variance by calculating the observed Fisher information matrix.¹⁸ The results of minimizing the likelihood defined in Eq. (C8) for different values of K are shown in Table VI. The mean ($\bar{m}_W$), uncertainty (σ), as well as the values of the two information criteria (AIC, BIC) are presented. We see that the mean and variances for $K = 2$ through $K = 5$ are similar and this is a result of the values of m_W for each Gaussian settling either around the latest measured values of CDF-II or ATLAS, and thus creating a degeneracy for values $K > 2$. The fact that adding additional Gaussians and going beyond $K = 2$ is unnecessary can also be seen by looking at the values of AIC and BIC in Table VI. Importantly, we see that the AIC and BIC are minimized for $K = 2$ and thus this model provides the best fit to the data. As expected, the uncertainty suggested by this model is 30.8 MeV and is significantly larger than the 7.2 MeV suggested by the $K = 1$ model, which is identical to the LS fit in Appendix C 1, which provided a poor fit to the data.

One could also consider the possibility of performing the fits without the latest CDF-II [9] data that are in tension with the ATLAS [68] data. When excluding the CDF-II data, we find that the AIC and BIC are minimized for $K = 1$, suggesting that the data are consistent and that the usual χ^2 approach should work. The result of the combination for $K = 1$, which is identical to the LS methods, is given in Eq. (C6). On the other hand, excluding only the ATLAS23 result and including CDF-II in the dataset does not lead to a consistent fit for $K = 1$. Moreover, the AIC and BIC are minimized for $K = 2$, suggesting that there is lingering tension with the rest of the data, especially since LEP, D0-II, and LHCb are in tension with CDF-II. The result of the fit is

$$\bar{m}_W|_{\text{no ATLAS23}}^{K=2} = 80.393 \pm 0.031 \text{ GeV}. \quad (\text{C12})$$

¹⁸It is also possible to sample the likelihood function in Eq. (C8) to produce a Monte Carlo set on which we can run the expectation-maximization algorithm.

The uncertainty suggested by $K = 2$ model when excluding ATLAS23 is about 31 MeV, which is similar to the uncertainty found with the full dataset.

We now consider the effect of artificially increasing the measured uncertainties on the GMM fits, which, for this simple example, is equivalent to introducing a global tolerance criteria. The GMM approach described here provides alternate measures for determining the amount by which the uncertainties should be rescaled to make the fit consistent. There are many ways to increase the uncertainty. For the purpose of simple demonstration, we choose to increase the uncertainty on all measurements given in Table V. We track the value of the AIC and BIC, for $K = 1$ and $K = 2$, as the uncertainty on each measurement is increased by the scale factor (S). The result of doing so is shown in Fig. 14.

The plot on the left in Fig. 14 shows the variation of the AIC for $K = 1$ and $K = 2$ as a function of the scale factor (S). Initially, when $S = 1$, the AIC is smaller for $K = 2$, indicating that this is a preferred model. As S increases, the AIC increases for $K = 2$ and decreases for $K = 1$ until they intersect at around $S \simeq 1.6$, suggesting that increasing the uncertainties of all datasets by at least this factor makes the data consistent. Also shown on the figure is $S = S_{\text{PDG}} = 1.95$.

The plot on the right shows the mean and estimates of uncertainty of the fits for $K = 1$ and $K = 2$, for various values of S . The mean value for $K = 2$ below $S \simeq 2.8$ is lower than the mean for $K = 1$, since most of the data are more compatible with the ATLAS23 measurement. Without knowledge of what the true value of the W -boson mass should be, the larger uncertainty and lower value of the mean for $K = 2$ and $S = 1$ is more representative of the

measured values across all experiments, whereas the result for $K = 1$ underestimates the uncertainty and also prefers a higher value of the mean, indicating the presence of a bias. Importantly, the bias introduced by combining all the data shifts the W -boson mass to larger values and this shift is larger than the (green) uncertainty band, even after scaling the uncertainties to make the data consistent. If it is not possible to unbiased the fit by omitting inconsistent datasets, then the uncertainty should be expanded to account for this bias.

Another way to say this is that if we were to predict the outcome of future experiments, based on the combined probability distribution for $K = 1$, then we should expect the central value to be closer to 80.4 with a high probability even after scaling the uncertainty. Furthermore, we should not expect future measurements to be close to the SM theory prediction (80.358 ± 0.008 GeV) [74]. On the other hand, the result of using $K = 2$ correctly indicates that there is a much higher probability for predicting lower values of the W -boson mass, since most past experiments are compatible with a lower value of the W -boson mass. Further, the larger uncertainty accommodates the discrepancy between experimental measurements and allows us to predict larger values for future measurements of the W -boson mass as well.

We end with a reminder that in the ideal situation, we would be able to go back to the experimental measurements and identify possible reasons for the discrepancy before combining the fits. While this may be possible for the W -boson mass, it is not typically the case for fits to parton distribution functions. We must therefore rely almost entirely on the existing data to provide a better estimate of the uncertainty of PDF fits.

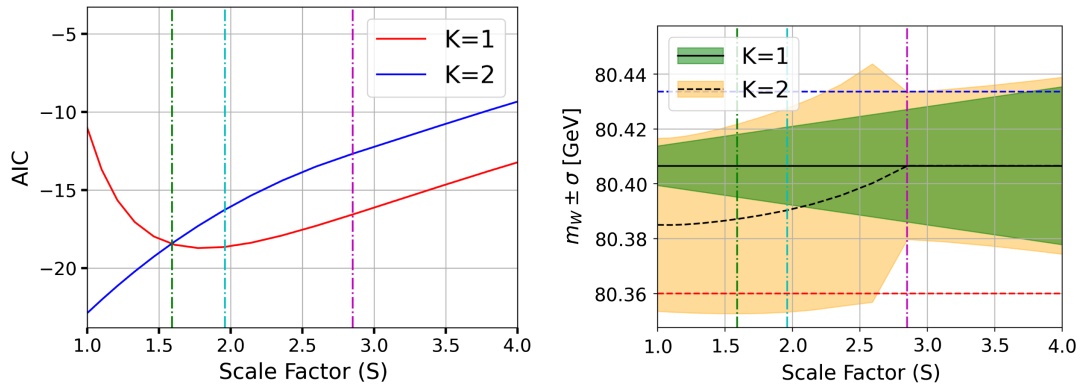


FIG. 14. Left: values of AIC as a function of the Scale factor (S) by which the uncertainties of experimental measurements of the W -boson mass in Table V are multiplied. The red and blue lines represent the AIC values for fits with $K = 1$ and $K = 2$, respectively. The three vertical dot-dashed lines represent the values of S when the AIC for $K = 1$ and $K = 2$ are equal (green), the value S calculated according to the PDG recommendation (cyan) and the value of S when the mean for $K = 1$ and $K = 2$ coincide (magenta). Right: mean and uncertainty of the combination of the data in Table V as a function of the Scale factor that multiplies the uncertainties of experimental measurements of the W -boson mass. The horizontal red dashed line represents the central value of the ATLAS23 measurement and the horizontal blue dashed line represents the central value of the CDF-II measurement. The three vertical lines are identical to the plot on the left and represent the same thing.

- [1] T.-J. Hou *et al.*, New CTEQ global analysis of quantum chromodynamics with high-precision data from the LHC, *Phys. Rev. D* **103**, 014013 (2021).
- [2] S. Bailey, T. Cridge, L. A. Harland-Lang, A. D. Martin, and R. S. Thorne, Parton distributions from LHC, HERA, Tevatron and fixed target data: MSHT20 PDFs, *Eur. Phys. J. C* **81**, 341 (2021).
- [3] R. D. Ball *et al.* (NNPDF Collaboration), The path to proton structure at 1% accuracy, *Eur. Phys. J. C* **82**, 428 (2022).
- [4] S. Alekhin, J. Blümlein, S. Moch, and R. Placakyte, Parton distribution functions, α_s , and heavy-quark masses for LHC Run II, *Phys. Rev. D* **96**, 014011 (2017).
- [5] G. Aad *et al.* (ATLAS Collaboration), Determination of the parton distribution functions of the proton using diverse ATLAS data from pp collisions at $\sqrt{s} = 7, 8$ and 13 TeV, *Eur. Phys. J. C* **82**, 438 (2022).
- [6] R. D. Ball *et al.* (NNPDF Collaboration), An open-source machine learning framework for global analyses of parton distributions, *Eur. Phys. J. C* **81**, 958 (2021).
- [7] S. Alekhin *et al.*, HERAFitter, *Eur. Phys. J. C* **75**, 304 (2015).
- [8] S. Amoroso *et al.*, Snowmass 2021 whitepaper: Proton structure at the precision frontier, *Acta Phys. Pol. B* **53**, A1 (2022).
- [9] T. Aaltonen *et al.* (CDF Collaboration), High-precision measurement of the W boson mass with the CDF II detector, *Science* **376**, 170 (2022).
- [10] M. Cepeda *et al.*, Report from Working Group 2: Higgs physics at the HL-LHC and HE-LHC, *CERN Yellow Rep. Monogr.* **7**, 221 (2019).
- [11] D. de Florian *et al.* (LHC Higgs Cross Section Working Group Collaboration), *Handbook of LHC Higgs Cross Sections: 4. Deciphering the Nature of the Higgs Sector* (CERN, Geneva, 2016), 10.23731/CYRM-2017-002.
- [12] M. Aaboud *et al.* (ATLAS Collaboration), Precision measurement and interpretation of inclusive W^+ , W^- and Z/γ^* production cross sections with the ATLAS detector, *Eur. Phys. J. C* **77**, 367 (2017).
- [13] T.-J. Hou, H.-W. Lin, M. Yan, and C. P. Yuan, Impact of lattice $s(x) - \bar{s}(x)$ data in the CTEQ-TEA global analysis, *arXiv:2204.07944*.
- [14] M. Guzzi *et al.*, NNLO constraints on proton PDFs from the SeaQuest and STAR experiments and other developments in the CTEQ-TEA global analysis, *SciPost Phys. Proc.* **8**, 005 (2022).
- [15] T.-J. Hou, M. Yan, J. Liang, K.-F. Liu, and C. P. Yuan, Connected and disconnected sea partons from the CT18 parametrization of PDFs, *Phys. Rev. D* **106**, 096008 (2022).
- [16] K. Kovarik, P. M. Nadolsky, and D. E. Soper, Hadronic structure in high-energy collisions, *Rev. Mod. Phys.* **92**, 045003 (2020).
- [17] R. D. Ball, A. Candido, J. Cruz-Martinez, S. Forte, T. Giani, F. Hekhorn, K. Kudashkin, G. Magni, and J. Rojo (NNPDF Collaboration), Evidence for intrinsic charm quarks in the proton, *Nature (London)* **608**, 483 (2022).
- [18] R. Stegeman, The negligible impact of experimental inconsistencies in the NNPDF4.0 global dataset, *Proc. Sci. ICHEP2022 (2022)* 787 [arXiv:2212.07703].
- [19] N. Lambri, Optimized regression models for parton distribution functions determination using deep learning models, Master's thesis, Università degli Studi di Milano, 2020.
- [20] L. Talon, Optimization of parton density uncertainties, Master's thesis, Università degli Studi di Milano, 2019.
- [21] T. Cridge (PDF4LHC21 Combination Group Collaboration), PDF4LHC21: Update on the benchmarking of the CT, MSHT and NNPDF global PDF fits, *SciPost Phys. Proc.* **8**, 101 (2022).
- [22] A. Gelman, J. Carlin, H. Stern, D. Dunson, A. Vehtari, and D. Rubin, *Bayesian Data Analysis, Third Edition*, Chapman & Hall/CRC Texts in Statistical Science (Taylor & Francis, London, 2013), 10.1201/b16018.
- [23] A. E. McGlothlin and K. Viele, Bayesian hierarchical models, *JAMA, J. Am. Med. Assoc.* **320**, 2365 (2018).
- [24] J. Erler and R. Ferro-Hernández, Alternative to the application of PDG scale factors, *Eur. Phys. J. C* **80**, 541 (2020).
- [25] K. P. Murphy, *Machine Learning: A Probabilistic Perspective* (MIT Press, Cambridge, MA, 2012).
- [26] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*, 1st ed. (Springer, New York, 2007).
- [27] E. E. Leamer and E. E. Leamer, *Specification Searches: Ad Hoc Inference with Nonexperimental Data* (Wiley, New York, 1978), Vol. 53.
- [28] E. I. George and R. E. McCulloch, Variable selection via gibbs sampling, *J. Am. Stat. Assoc.* **88**, 881 (1993).
- [29] R. E. Kass and A. E. Raftery, Bayes factors, *J. Am. Stat. Assoc.* **90**, 773 (1995).
- [30] J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky, Bayesian model averaging: A tutorial (with comments by M. Clyde, David Draper and E. I. George, and a rejoinder by the authors, *Stat. Sci.* **14**, 382 (1999).
- [31] A. E. Raftery and Y. Zheng, Discussion: Performance of bayesian model averaging, *J. Am. Stat. Assoc.* **98**, 931 (2003).
- [32] W. I. Jay and E. T. Neil, Bayesian model averaging for analysis of lattice field theory results, *Phys. Rev. D* **103**, 114502 (2021).
- [33] E. T. Neil and J. W. Sitison, Improved information criteria for Bayesian model averaging in lattice field theory, *Phys. Rev. D* **109**, 014510 (2024).
- [34] E. T. Neil and J. W. Sitison, Model averaging approaches to data subset selection, *Phys. Rev. E* **108**, 045308 (2023).
- [35] G. Cybenko, Approximation by superpositions of a sigmoidal function, *Math. Control Signals Syst.* **2**, 303 (1989).
- [36] M. H. Stone, Applications of the theory of boolean rings to general topology, *Trans. Am. Math. Soc.* **41**, 375 (1937).
- [37] M. H. Stone, *Mathematics Magazine* (Mathematical Association of America, Taylor & Francis, Ltd., 1948), Vol. 21, pp. 167–184, 10.2307/3029750.
- [38] R. L. Workman *et al.* (Particle Data Group Collaboration), Review of particle physics, *Prog. Theor. Exp. Phys.* **2022**, 083C01 (2022).
- [39] K. Pearson, Contributions to the mathematical theory of evolution, *Phil. Trans. R. Soc. A* **185**, 71 (1894).
- [40] M. Abadi *et al.*, TensorFlow: Large-scale machine learning on heterogeneous distributed systems, *arXiv:1603.04467*.

- [41] G. J. McLachlan and D. Peel, *Finite Mixture Models*, Wiley Series in Probability and Statistics (John Wiley & Sons, 2000), [10.1002/0471721182](#).
- [42] A. E. Raftery, T. Gneiting, F. Balabdaoui, and M. Polakowski, Using Bayesian model averaging to calibrate forecast ensembles, *Mon. Weather Rev.* **133**, 1155 (2005).
- [43] H. M. Bücker and G. F. Corliss, A bibliography on automatic differentiation, in *Automatic Differentiation*, Lecture Notes in Computational Science and Engineering Vol. 50, edited by H. M. Bücker, G. F. Corliss, P. D. Hovland, U. Naumann, and B. Norris (Springer, New York, 2005), pp. 321–322, [10.1007/3-540-28438-9_28](#).
- [44] A. Courtoy, J. Huston, P. Nadolsky, K. Xie, M. Yan, and C. P. Yuan, Parton distributions need representative sampling, *Phys. Rev. D* **107**, 034008 (2023).
- [45] H.-L. Lai, M. Guzzi, J. Huston, Z. Li, P. M. Nadolsky, J. Pumplin, and C. P. Yuan, New parton distributions for collider physics, *Phys. Rev. D* **82**, 074024 (2010).
- [46] J. Gao, M. Guzzi, J. Huston, H.-L. Lai, Z. Li, P. Nadolsky, J. Pumplin, D. Stump, and C. P. Yuan, CT10 next-to-next-to-leading order global analysis of QCD, *Phys. Rev. D* **89**, 033009 (2014).
- [47] A. D. Martin, W. J. Stirling, R. S. Thorne, and G. Watt, Parton distributions for the LHC, *Eur. Phys. J. C* **63**, 189 (2009).
- [48] L. A. Harland-Lang, A. D. Martin, P. Motylinski, and R. S. Thorne, Parton distributions in the LHC era: MMHT 2014 PDFs, *Eur. Phys. J. C* **75**, 204 (2015).
- [49] X. Jing *et al.*, Quantifying the interplay of experimental constraints in analyses of parton distributions, *Phys. Rev. D* **108**, 034029 (2023).
- [50] H. Akaike, A new look at the statistical model identification, *IEEE Trans. Autom. Control* **19**, 716 (1974).
- [51] G. Schwarz, Estimating the dimension of a model, *Ann. Stat.* **6**, 461 (1978).
- [52] G. Cowan, Statistical models with uncertain error parameters, *Eur. Phys. J. C* **79**, 133 (2019).
- [53] G. D’Agostini, Sceptical combination of experimental results: General considerations and application to epsilon-prime / epsilon, [arXiv:hep-ex/9910036](#).
- [54] H. Paukkunen and P. Zurita, PDF reweighting in the Hessian matrix approach, *J. High Energy Phys.* **12** (2014) 100.
- [55] D. H. Wolpert, Stacked generalization, *Neural Netw.* **5**, 241 (1992).
- [56] Y. Yao, A. Vehtari, D. Simpson, and A. Gelman, Using stacking to average Bayesian predictive distributions (with discussion), *Bayesian Anal.* **13**, 917 (2018).
- [57] L. Wasserman, Bayesian model selection and model averaging, *J. Math. Psychol.* **44**, 92 (2000).
- [58] T. P. Minka, Bayesian model averaging is not model combination, Available electronically at <http://www.stat.cmu.edu/minka/papers/bma.html> (2000), p. 1.
- [59] K. Monteith, J. L. Carroll, K. Seppi, and T. Martinez, Turning bayesian model averaging into bayesian model combination, in *The 2011 International Joint Conference on Neural Networks* (2011), pp. 2657–2663, [10.1109/IJCNN.2011.6033566](#).
- [60] K. Kamary, K. Mengersen, C. P. Robert, and J. Rousseau, Testing hypotheses via a mixture estimation model, [arXiv:1412.2044](#).
- [61] M. Keller and K. Kamary, Bayesian model averaging via mixture model estimation, [arXiv:1711.10016](#).
- [62] A. V. Kotwal, Model for the curvature response of the CDF II drift chamber, *Phys. Rev. Res.* **7**, 013128 (2025).
- [63] V. M. Abazov *et al.* (D0 Collaboration), Improved W boson mass measurement with the D0 detector, *Phys. Rev. D* **66**, 012001 (2002).
- [64] T. Affolder *et al.* (CDF Collaboration), Measurement of the W boson mass with the collider detector at Fermilab, *Phys. Rev. D* **64**, 052001 (2001).
- [65] S. Schael *et al.* (ALEPH, DELPHI, L3, OPAL, LEP Electroweak Collaborations), Electroweak measurements in electron-positron collisions at W -boson-pair energies at LEP, *Phys. Rep.* **532**, 119 (2013).
- [66] V. M. Abazov *et al.* (D0 Collaboration), Measurement of the W boson mass with the D0 detector, *Phys. Rev. Lett.* **108**, 151804 (2012).
- [67] R. Aaij *et al.* (LHCb Collaboration), Measurement of the W boson mass, *J. High Energy Phys.* **01** (2022) 036.
- [68] G. Aad *et al.* (ATLAS Collaboration), Measurement of the W -boson mass and width with the ATLAS detector using proton–proton collisions at $\sqrt{s} = 7$ TeV, *Eur. Phys. J. C* **84**, 1309 (2024).
- [69] S. Navas *et al.* (Particle Data Group Collaboration), Review of particle physics, *Phys. Rev. D* **110**, 030001 (2024).
- [70] S. Amoroso *et al.* (LHC-TeV MW Working Group Collaboration), Compatibility and combination of world W -boson mass measurements, *Eur. Phys. J. C* **84**, 451 (2024).
- [71] J. de Blas, M. Ciuchini, E. Franco, A. Goncalves, S. Mishima, M. Pierini, L. Reina, and L. Silvestrini, Global analysis of electroweak data in the standard model, *Phys. Rev. D* **106**, 033003 (2022).
- [72] J. Haller, A. Hoecker, R. Kogler, K. Mönig, and J. Stelzer, Status of the global electroweak fit with Gfitter in the light of new precision measurements, *Proc. Sci. ICHEP2022* (2022) 897 [[arXiv:2211.07665](#)].
- [73] H.-W. Lin *et al.*, Parton distributions and lattice QCD calculations: A community white paper, *Prog. Part. Nucl. Phys.* **100**, 107 (2018).
- [74] M. Baak, J. Cúth, J. Haller, A. Hoecker, R. Kogler, K. Mönig, M. Schott, and J. Stelzer (Gfitter Group Collaboration), The global electroweak fit at NNLO and prospects for the LHC and ILC, *Eur. Phys. J. C* **74**, 3046 (2014).