M&C 2017 - International Conference on Mathematics & Computational Methods Applied to Nuclear Science & Engineering, Jeju, Korea, April 16-20, 2017, on USB (2017)

Fitting and Analysis Technique for Inconsistent Nuclear Data

Georg Schnabel

Irfu, CEA, Université Paris-Saclay, 91191 Gif-sur-Yvette, France
georg.schnabel@cea.fr

Abstract - *Consistent experiment data are crucial to adjust parameters of physics models and to determine best estimates of observables. However, often experiment data are not consistent due to unrecognized systematic errors. Standard methods of statistics such as χ^2-fitting cannot deal with this case. Their predictions become doubtful and associated uncertainties too small. A human has then to figure out the problem, apply corrections to the data, and repeat the fitting procedure. This takes time and potentially costs money. Therefore, a Bayesian method is introduced to fit and analyze inconsistent experiment data. It automatically detects and resolves inconsistencies. Furthermore, it allows to extract consistent subsets from the data. Finally, it provides an overall prediction with associated uncertainties and correlations less prone to the common problem of too small uncertainties. The method is foreseen to function with a large corpus of data and hence may be used in nuclear databases to deal with inconsistencies in an automated fashion.*

I. INTRODUCTION

Evaluated nuclear data help to push forward the development of novel nuclear facilities. They are needed as input for transport and activation calculations. Any mistake or inconsistency in the data may distort calculation results and, as a consequence, may lead to suboptimal design choices with regard to efficiency and safety.

The acquisition and determination of consistent nuclear data are non-trivial tasks. Not all types of nuclear data required for calculations are available from experiments. Especially reaction data at higher energies are scarce. The remedy is to use predictions of nuclear models or some series expansion and to adjust the parameters to available experimental data. The adjustment of parameters is usually done by a χ^2-fit or the *Generalized Least Squares* (GLS) method, e.g. [1].

The problem with these approaches is the inherent assumption that experiment data are consistent, meaning the data contain what they claim. Consistent experiment data are an ideal. Experiments are complex and acquired raw data has to be corrected for many effects such as background noise and detector efficiency. Already if one such correction has not been done properly, or the associated uncertainties are not estimated well, the experiment data are inconsistent. The visual signature of inconsistent data sets are points whose error bars mutually exclude each other.

Until now, besides using Chauvenet's criterion to remove outliers, e.g. [2, 3], a human had to resolve the inconsistencies by reading the publications and trying to figure out which (if any) data set is better and remove the other one. Unfortunately, the publications are not always accessible or they do not allow a clear statement about which data set is right or wrong. In such a situation, we face the dilemma that rejecting one data set would be arbitrary, but feeding contradicting data sets to conventional fitting methods gives bad results. For instance, the inclusion of two contradicting data sets leads to a larger reduction of uncertainty than if including only one. Common sense suggests that contradictions should increase uncertainties.

The fitting method proposed in this paper resolves these issues. Contradicting experimental data sets can be included at once. The method automatically assigns additional uncertainties to the data sets to achieve consistency. These additional uncertainties enable the segmentation of the experiments into consistent subsets. A human expert can then decide upon which subset is most appropriate and feed it to a conventional fitting method. The possibility to determine several interpretations of the data in form of consistent subsets is an advantage over Chauvenet's criterion, which yields only one interpretation. Furthermore, overall estimates, uncertainties and correlations (in short covariance matrices) including all subsets can be obtained by means of Monte Carlo sampling. As desired, contradictory experiment data sets increase uncertainties. The method is mathematically well founded within the framework of Bayesian statistics.

Technicalities aside, the proposed method is related to the procedures presented in [4, 5, 6]. These papers deal with the problem of estimating an experiment covariance matrix which may then be used to fit a model. In contrast to that, the method introduced in this paper treats the estimation (or correction) of experiment covariance matrices as an integral part of the model fitting procedure.

The improved uncertainty quantification of evaluated nuclear data may be seen as the key feature of the proposed method. The propagation of more realistic uncertainties of nuclear data should lead to a better assessment of simulation results and as a final consequence to safer and more efficient nuclear facilities.

II. METHOD

1. Prototypic Model

To make the discussion of the proposed method more practical, assume that we want to fit some total cross section $\sigma(E)$, which is a function of the incident energy E. We take

as prototypic model the function

$$\boldsymbol{\sigma}_{\text{fit}}(E) = \frac{\sum_{i=1}^{M} y_i \mathcal{N}\left(E \,|\, x_i, \lambda^2\right)}{\sum_{j=1}^{M} \mathcal{N}\left(E \,|\, x_j, \lambda^2\right)} \,. \tag{1}$$

Expressions of this form appear in Nadaraya-Watson kernel regression, which is a non-parametric method for fitting. The function $\mathcal{N}(E \,|\, x_i, \lambda^2)$ gives the probability density at location E of a normal distribution centered at energy x_i with standard deviation λ. The number of grid points x_i, their locations, and the standard deviation λ are fixed. The y_i are the adjustable 'model' parameters defining the shape of the function. For notational convenience, we define the model parameter vector $\boldsymbol{y} = (y_1, \cdots, y_M)^T$. Given enough grid points, the function in eq. (1) can mimic a multitude of possible shapes, which are determined by the choice of $\boldsymbol{y}$. This prototypic model is representative for all models with a linear relationship between model parameters and predictions. Therefore, the subsequent discussion equally applies to e.g. Fourier expansions, Legendre polynomials, and splines. Non-linear models can be replaced by linear approximation or by a surrogate model based on a multivariate normal distribution to make them accessible for the method, e.g. [7]. Of course, real physical models which possess more structure can also be used instead of series expansions.

2. Standard GLS Method and Preliminaries

The proposed method is formulated within the framework of Bayesian statistics, e.g. [8]. We start with outlining the popular GLS method, e.g. [1, 10], for nuclear data evaluation and afterward introduce modifications leading eventually to the new method. The Bayesian update formula reads

$$\rho(\boldsymbol{y} \,|\, \boldsymbol{\sigma}_{\text{exp}}, B) = \frac{\rho(\boldsymbol{\sigma}_{\text{exp}} \,|\, \boldsymbol{y}, B) \times \rho(\boldsymbol{y} \,|\, \boldsymbol{y}_0, A_0)}{\rho(\boldsymbol{\sigma}_{\text{exp}})} \,. \tag{2}$$

The *probability density function* (pdf) $\rho(\boldsymbol{y} \,|\, \boldsymbol{y}_0, A_0)$ reflects the prior knowledge about the model parameters. The standard assumption is that the *prior* pdf for $\boldsymbol{y}$ is given by a multivariate normal distribution with some center vector $\boldsymbol{y}_0$ and covariance matrix A_0, i.e. $\rho(\boldsymbol{y} \,|\, \boldsymbol{y}_0, A_0) = \mathcal{N}(\boldsymbol{y} \,|\, \boldsymbol{y}_0, A_0)$. The *likelihood* $\rho(\boldsymbol{\sigma}_{\text{exp}} \,|\, \boldsymbol{y})$ gives the probability for observing the experimental data set $\boldsymbol{\sigma}_{\text{exp}}$ under the condition that $\boldsymbol{y}$ is the true parameter vector. It is also given by a multivariate normal distribution $\mathcal{N}(\boldsymbol{\sigma}_{\text{exp}} \,|\, S\boldsymbol{y}, B)$. The covariance matrix B is assumed to be known a priori and reflects the statistical and systematic errors of the experiments.

The sensitivity matrix S maps the model parameters to the observables of the experiments. It equals the Jacobian matrix, which contains the derivatives of the model predictions with respect to the model parameters. For instance, the Jacobian matrix of the prototypic model introduced in eq. (1) is

$$S_{kl} = \frac{\partial}{\partial y_l} \boldsymbol{\sigma}_{\text{fit}}(E_k) = \frac{\mathcal{N}(E_k \,|\, x_l, \lambda)}{\sum_{j=1}^{M} \mathcal{N}(E_k \,|\, x_j, \lambda)} \,, \tag{3}$$

where E_k denotes the energy associated with the k^{th} measurement point in $\boldsymbol{\sigma}_{\text{exp}}$. This matrix is constant with respect to the model parameters $\boldsymbol{y}$, which holds true for linear models in general.

The *marginal likelihood* $\rho(\boldsymbol{\sigma}_{\text{exp}})$ yields the probability density for $\boldsymbol{\sigma}_{\text{exp}}$ under all modelling assumptions and is determined by

$$\rho(\boldsymbol{\sigma}_{\text{exp}}) = \int \rho(\boldsymbol{\sigma}_{\text{exp}} \,|\, \boldsymbol{y}, B) \times \rho(\boldsymbol{y} \,|\, \boldsymbol{y}_0, A_0) \, \mathrm{d}\boldsymbol{y} \,. \tag{4}$$

It rescales the product of likelihood and prior to become a correctly normalized posterior pdf. Due to the form of a multivariate normal pdf, conveniently expressed in terms of its logarithm,

$$\ln \mathcal{N}(\boldsymbol{x} \,|\, \boldsymbol{x}_0, \Sigma) = -\frac{N}{2} \ln(2\pi) - \frac{1}{2} \ln \det \Sigma - \frac{1}{2} (\boldsymbol{x} - \boldsymbol{x}_0)^T \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{x}_0) \tag{5}$$

with the center vector $\boldsymbol{x}_0$ containing N elements and the $N \times N$ covariance matrix Σ, the marginal likelihood can be calculated analytically. The result is (e.g. [9, p. 93])

$$\rho(\boldsymbol{\sigma}_{\text{exp}}) = \mathcal{N}(\boldsymbol{\sigma}_{\text{exp}} \,|\, S\boldsymbol{y}_0, M) \;\text{ with } M = S A_0 S^T + B \,. \tag{6}$$

Consequently, also the posterior pdf in eq. (2) has a closed-form solution. It is given by the multivariate normal distribution,

$$\rho(\boldsymbol{y} \,|\, \boldsymbol{\sigma}_{\text{exp}}, B) = \mathcal{N}(\boldsymbol{y} \,|\, \boldsymbol{y}_1, A_1) \;\text{ with} \tag{7}$$

$$\boldsymbol{y}_1 = \boldsymbol{y}_0 + A_0 S^T (S A_0 S^T + B)^{-1} (\boldsymbol{\sigma}_{\text{exp}} - S\boldsymbol{y}_0) \,, \tag{8}$$

$$A_1 = A_0 - A_0 S^T (S A_0 S^T + B)^{-1} S A_0 \tag{9}$$

with the new center vector $\boldsymbol{y}_1$ and new covariance matrix A_1 for the model parameters. The application of these two formulas is commonly understood as the GLS method. Depending on the dimensions of the matrices, the following equivalent formulas may be preferred:

$$\boldsymbol{y}_1 = A_1 \left(A_0^{-1} \boldsymbol{y}_0 + S^T B^{-1} \boldsymbol{\sigma}_{\text{exp}} \right) , \tag{10}$$

$$A_1 = \left(A_0^{-1} + S^T B^{-1} S \right)^{-1} \,. \tag{11}$$

A derivation of eqs. (8) to (11) can be found in e.g. [10]. The method of χ^2-fitting can be regarded as a special case where $A_0 = \eta Q$ with a matrix Q of full rank and $\eta \to \infty$.

Defining a sensitivity matrix S_{ev} to map to a suitable output grid, the result of $S_{\text{ev}} \boldsymbol{y}_1$ together with the associated covariance matrix $S_{\text{ev}}^T A_1 S_{\text{ev}}$ enters evaluated nuclear data files.

Finally, it has to be noted that $\boldsymbol{\sigma}_{\text{exp}}$ usually bundles data sets from different experiments $\boldsymbol{\sigma}_{\text{exp},i}$. Each measurement vector $\boldsymbol{\sigma}_{\text{exp},i}$ is associated with a covariance matrix B_i. Uncertainties of distinct experiments will be assumed to be uncorrelated, which gives rise to a block diagonal structure of B, with the B_i as blocks. The block diagonal structure of B provides computational benefits and enables the application of the proposed method to large data sets.

3. New Method

A. Uncertainty about the Experiment Covariance Matrix

The standard GLS method assumes that the experiment covariance matrix B is perfectly known a priori. However, in reality it is often very difficult to account exactly for all systematic uncertainties. This circumstance suggests to regard *B itself* as uncertain. In practice, we introduce this additional uncertainty by parameterizing the covariance matrix. More precisely, each block B_i is parametrized individually. Among the many possibilities, an additional normalization uncertainty may be one of the most plausible options,

$$C_i(\kappa_i) = B_i + \kappa_i^2 \boldsymbol{\sigma}_{\mathrm{exp},i} \boldsymbol{\sigma}_{\mathrm{exp},i}^T \,. \tag{12}$$

In an usual evaluation, a human evaluator would try to assign a reasonable value for the parameter κ_i based on his knowledge about the experiment. In the proposed method, most probable values will be determined automatically. Because the uncertainty about κ_i will be expressed in terms of a probability distribution, there is lots of flexibility to account for prior knowledge. For instance, variations of κ_i could be restricted to take place only in a certain interval. Noteworthy, the parametrization in eq. (12) has to be seen as a suggestion and other choices are equally reasonable. For example, if one data set covers a broad range of incident energies, it may suit to introduce an uncertainty component that exhibits only mid-range energy correlation instead of a perfect correlation between the errors at all energies. Such a parameterization will be demonstrated and discussed at the end of section III..

B. Extended Bayesian Update Formula

The introduction of new variables into the inference procedure necessitates an extension of the Bayesian update formula. For convenience, we combine the variables κ_i associated with different experiments to the vector $\boldsymbol{\kappa}$. The Bayesian update formula now reads

$$\rho(\boldsymbol{y}, \boldsymbol{\kappa} \,|\, \boldsymbol{\sigma}_{\mathrm{exp}}) = \frac{\rho(\boldsymbol{\sigma}_{\mathrm{exp}} \,|\, \boldsymbol{y}, \boldsymbol{\kappa}) \times \rho(\boldsymbol{y} \,|\, \boldsymbol{y}_0, A_0) \times \rho(\boldsymbol{\kappa})}{\rho(\boldsymbol{\sigma}_{\mathrm{exp}})} \,. \tag{13}$$

As in the case of the standard GLS, the likelihood is given by a multivariate normal distribution, $\rho(\boldsymbol{\sigma}_{\mathrm{exp}} \,|\, \boldsymbol{y}, \boldsymbol{\kappa}) = \mathcal{N}(\boldsymbol{\sigma}_{\mathrm{exp}} \,|\, S\,\boldsymbol{y}, C(\boldsymbol{\kappa}))$. Noteworthy, the covariance matrix B is replaced by $C(\boldsymbol{\kappa})$ whose blocks are determined by eq. (12). The specification of the prior for the model parameters $\rho(\boldsymbol{y} \,|\, \boldsymbol{y}_0, A_0) = \mathcal{N}(\boldsymbol{y} \,|\, \boldsymbol{y}_0, A_0)$ mirrors the standard GLS method. We postpone discussing the choice of the prior pdf $\rho(\boldsymbol{\kappa})$ for a moment.

Contrary to the standard GLS approach, the marginal likelihood

$$\rho(\boldsymbol{\sigma}_{\mathrm{exp}}) = \int \left(\int \rho(\boldsymbol{y}, \boldsymbol{\sigma}_{\mathrm{exp}} \,|\, \boldsymbol{\kappa}) \times \rho(\boldsymbol{\kappa}) \,\mathrm{d}\boldsymbol{y} \right) \mathrm{d}\boldsymbol{\kappa}$$
$$\text{with } \rho(\boldsymbol{y}, \boldsymbol{\sigma}_{\mathrm{exp}} \,|\, \boldsymbol{\kappa}) = \rho(\boldsymbol{\sigma}_{\mathrm{exp}} \,|\, \boldsymbol{y}, \boldsymbol{\kappa}) \times \rho(\boldsymbol{y} \,|\, \boldsymbol{y}_0, A_0) \tag{14}$$

has no straight-forward analytical solution. Only the inner integral can be analytically evaluated. Noting that it has the same form as eq. (4), the solution analogous to eq. (6) is

$$\rho(\boldsymbol{\kappa} \,|\, \boldsymbol{\sigma}_{\mathrm{exp}}) \propto \rho(\boldsymbol{\sigma}_{\mathrm{exp}}, \boldsymbol{\kappa}) = \mathcal{N}(\boldsymbol{\sigma}_{\mathrm{exp}} \,|\, S\,\boldsymbol{y}_0,\ M) \times \rho(\boldsymbol{\kappa})$$
$$\text{with } M = S A_0 S^T + C(\boldsymbol{\kappa}) \,. \tag{15}$$

Because this expression is proportional to the posterior pdf $\rho(\boldsymbol{\kappa} \,|\, \boldsymbol{\sigma}_{\mathrm{exp}})$, it is the key to assess the consistency of the experiment data sets. The vector $\boldsymbol{\kappa}'$ that maximizes eq. (15) contains the most probable values for the parameters in the experiment covariance matrix. It tells us which data sets are consistent and which are not, and how wrong the inconsistent ones are estimated to be. In the case of several local maxima, each maximum is associated with a certain interpretation of the experiments. Details concerning the computation and optimization of $\rho(\boldsymbol{\kappa} \,|\, \boldsymbol{\sigma}_{\mathrm{exp}})$ will be discussed in section F.

C. Choice of the Shape of the Prior Distribution

For a full specification of $\rho(\boldsymbol{\kappa} \,|\, \boldsymbol{\sigma}_{\mathrm{exp}})$, we have to define the prior pdf $\rho(\boldsymbol{\kappa})$. Knowledge about correlations between different κ_i is usually limited. Furthermore, the automated detection of most probable adjustments of the experiment covariance matrix is one of the main reasons for the introduction of the new method. Consequently, we want to avoid an informative prior for the covariance matrix parameters.

The information content of a pdf can be characterized in terms of *entropy* (e.g. [11])—the higher the entropy of a pdf, the lower the information content. Given only the marginal pdfs $\rho(\kappa_i), i = 1..N$ with associated entropies $\mathcal{H}[\rho(\kappa_i)]$, the joint pdf $\rho(\boldsymbol{\kappa})$ with highest entropy and compatible with all the marginal pdfs is just the product of the marginal pdfs,

$$\rho(\boldsymbol{\kappa}) = \rho(\kappa_1)\rho(\kappa_2)\ldots\rho(\kappa_N) \,. \tag{16}$$

This result follows from the subadditivity of the entropy [11, p. 28],

$$\mathcal{H}[\rho(\kappa_1), \rho(\kappa_2), \cdots, \rho(\kappa_N)] \leq$$
$$\mathcal{H}[\rho(\kappa_1)] + \mathcal{H}[\rho(\kappa_2)] + ... + \mathcal{H}[\rho(\kappa_N)] \,, \tag{17}$$

with equality only if the variables κ_i are statistically independent, i.e. the joint pdf factorizes into the product of the marginal pdfs.

Concerning the functional form of $\rho(\kappa_i)$, I investigated the Laplace pdf, the normal pdf, and the improper uniform pdf in a schematic evaluation of the neutron-proton total cross section. The term improper refers to the fact that the uniform pdf extends over the complete real line and hence cannot be normalized. Details about the findings will be presented in section III.. However, some results must be already anticipated here in order to provide a complete picture of the method.

In my investigation, I found arguments in favor of the Laplace pdf,

$$\mathcal{L}(\kappa_i \,|\, \delta_i) = \frac{1}{\sqrt{2}\,\delta_i} \exp\left(-\frac{\sqrt{2}\,|\kappa_i|}{\delta_i} \right) . \tag{18}$$

This pdf is symmetric with mean zero and standard deviation δ_i. Experiments believed to be more correct could be

associated with smaller δ_i than those being more distrusted. However, in an automated evaluation without much human involvement, there is no reason to favor one experiment over another a priori. Therefore, I investigated only the case where all δ_i are equal.

In the studied scenario, the experiment data in combination with the uniform distribution did not sufficiently constrain the posterior pdf. Even though the most probable assignments $\boldsymbol{\kappa}'$ were usually reasonable, the relative standard deviations of the κ_i exceeded thousand percent—unreasonably large. In contrast to that, both the normal pdf and the Laplace pdf with a reasonable standard deviation δ restricted sufficiently the spread of $\rho(\boldsymbol{\kappa}\,|\,\boldsymbol{\sigma}_{\text{exp}})$. Noteworthy, the Laplace pdf tended to set more κ_i to zero at the cost of slightly increased values of non-zero parameters. I regard this behavior to favor sparse solutions as beneficial. If this behavior is not desired, the normal distribution should be prefered.

In fact, the logarithm of the product of identical Laplace pdfs appears (up to a constant) as penalty term in LASSO regression [12] where it serves the exact purpose of variable elimination. To understand this behaviour, consider the set $\{\boldsymbol{\kappa}\,|\,\tau = \rho(\boldsymbol{\kappa})\}$ for a fixed positive real number τ and with $\rho(\boldsymbol{\kappa})$ being specified as a product of identical Laplace pdfs. The vectors in this set define a hypercube whose corners are aligned with the parameter axes. The gradient perpendicular to the surface of this hypercube does almost everywhere *not* point exactly to the origin of the coordinate system, as it would be the case for the product of identical normal distributions. Instead, following locally the direction of steepest ascent leads to a vector with one component being zero, say $\kappa_1 = 0$. Continuing on the path of steepest ascent leads stepwise to the elimination of more and more parameters. The center of the distribution is reached only at the very end of the path. The process is visualized in fig. 1. This theoretical argument explains why the preference for sparse solutions is a general feature when using a product of Laplace pdfs as prior pdf.

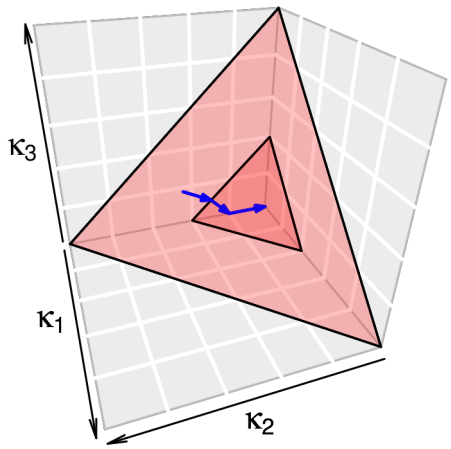

Fig. 1. Surfaces of equal probability for a product of three identical Laplace pdfs. Displayed is the octant where all parameters have positive values. The blue arrows show an exemplary path of steepest ascent.

D. Choice of the Parameter δ

The prior pdf $\rho(\boldsymbol{\kappa})$ of the covariance matrix parameters $\boldsymbol{\kappa}$ depends itself on parameters. In the last section we encountered the standard deviation δ as the parameter defining the shape of the identical Laplace pdfs. We can make this dependence explicit by writing $\rho(\boldsymbol{\kappa}\,|\,\delta)$ instead of $\rho(\boldsymbol{\kappa})$. Having introduced a new parameter, which value should we assign to it?

Again, we anticipate some results from section III. If the κ_i denote normalization uncertainties, then δ should be set to a plausible value for the normalization uncertainty. So if we think it is quite probable that some experiments have normalization errors between 5% and 10% which were not considered in the original experiment covariance matrix B, then δ should be also in that range. It appears that evaluation results are only mildly dependent on the exact choice of δ. This approach, however, is rather subjective. Next we discuss data-driven approaches to alleviate the problem of subjectivity.

A sufficient criterion to determine whether the value of δ is large enough is to calculate the generalized χ^2-value

$$\chi^2(\boldsymbol{\kappa}') = (\boldsymbol{\sigma}_{\text{exp}} - S\,\boldsymbol{y}_0)^T \left[M(\boldsymbol{\kappa}')\right]^{-1} (\boldsymbol{\sigma}_{\text{exp}} - S\,\boldsymbol{y}_0) \tag{19}$$

where $\boldsymbol{\kappa}'$ maximizes $\rho(\boldsymbol{\kappa}\,|\,\boldsymbol{\sigma}_{\text{exp}})$, see eq. (15). The quantity $\chi^2(\boldsymbol{\kappa}')/N$, with N being the number of measurement points, should be close to one. Otherwise one or more of the following statements is true: 1) the value of δ is too low, 2) the model to fit the data is misspecified, 3) a normalization uncertainty is not enough to correct the misspecified experiment covariance matrix. At the end of section III. we discuss besides a normalization uncertainty also a more flexible energy-dependent uncertainty.

An approach that directly aims at the determination of δ is the maximization of the marginal likelihood $\rho(\boldsymbol{\sigma}_{\text{exp}})$ defined in eq. (14). Because the prior $\rho(\boldsymbol{\kappa}\,|\,\delta)$ is conditioned on δ, we more appropriately write $\rho(\boldsymbol{\sigma}_{\text{exp}}\,|\,\delta)$ instead of $\rho(\boldsymbol{\sigma}_{\text{exp}})$. The resulting value represents the probability density to obtain the measurement vector $\boldsymbol{\sigma}_{\text{exp}}$ given a certain value of δ. Of course, this probability density is also conditioned on the assumption of the model, the parameterization of the adjusted experiment covariance matrix C, and the prior specifications of all occurring parameters Selecting a value for δ that maximizes $\rho(\boldsymbol{\sigma}_{\text{exp}}\,|\,\delta)$ is a sensible choice, effectively removing subjectivity.

Unfortunately, it seems as there is no analytical expression for $\rho(\boldsymbol{\sigma}_{\text{exp}}\,|\,\delta)$. We can approximately solve the integral by using Monte Carlo integration in combination with *importance sampling*, e.g. [13, p. 131]. The idea is to identically rewrite eq. (14) as

$$\rho(\boldsymbol{\sigma}_{\text{exp}}\,|\,\delta) = \int \frac{\rho(\boldsymbol{\sigma}_{\text{exp}}\,|\,\boldsymbol{\kappa}) \times \rho(\boldsymbol{\kappa}\,|\,\delta)}{\phi(\boldsymbol{\kappa})} \times \phi(\boldsymbol{\kappa})\,\mathrm{d}\boldsymbol{\kappa}\,. \tag{20}$$

Given that $\phi(\boldsymbol{\kappa})$ is a pdf from which we can draw a sample $\boldsymbol{\kappa}_1, \boldsymbol{\kappa}_2, \cdots, \boldsymbol{\kappa}_K$, an estimate of this integral is

$$\rho(\boldsymbol{\sigma}_{\text{exp}}\,|\,\delta) \approx \frac{1}{K} \sum_{i=1}^{K} \frac{\rho(\boldsymbol{\sigma}_{\text{exp}}\,|\,\boldsymbol{\kappa}_i)}{\phi(\boldsymbol{\kappa}_i)} \times \rho(\boldsymbol{\kappa}_i\,|\,\delta)\,. \tag{21}$$

The choice of $\phi(\boldsymbol{\kappa})$ will be discussed in a moment.

Finding the value of δ that maximizes $\rho(\boldsymbol{\sigma}_{\text{exp}} \,|\, \delta)$ means to evaluate the integral for many possible values of δ. In order to scan the parameter space in a systematic way, we can evaluate the integral on a grid of reasonable values $\delta_1, \delta_2, \cdots, \delta_M$. In fact, we can use the same sequence $\boldsymbol{\kappa}_1, \boldsymbol{\kappa}_2, \cdots, \boldsymbol{\kappa}_K$ drawn from $\phi(\boldsymbol{\kappa})$ to estimate all the integrals $\phi(\boldsymbol{\sigma}_{\text{exp}} \,|\, \delta_i), i = 1..M$ in parallel. The ratios $\rho(\boldsymbol{\sigma}_{\text{exp}} \,|\, \boldsymbol{\kappa}_i)/\phi(\boldsymbol{\kappa}_i)$ in eq. (21) are the same for all integrals. Only the last factor $\rho(\boldsymbol{\kappa}_i \,|\, \delta)$ has to be recomputed for each value δ_i. Because it does not depend on the experiment data sets and is of a simple form, e.g. a Laplace pdf, it can be evaluated quickly.

How should the sampling distribution $\phi(\boldsymbol{\kappa})$ be chosen? A sampling distribution that declines at a faster rate in the tails than the other part of the integrand destroys convergence in importance sampling. In order to protect against this case, we define a mixture of possible posterior pdfs associated with the values δ_j on the grid,

$$\phi(\boldsymbol{\kappa}) = \mathcal{I} \times \mathcal{N}(\boldsymbol{\sigma}_{\text{exp}} \,|\, S\,\boldsymbol{y}_0,\ M(\boldsymbol{\kappa})) \times \sum_j \rho(\boldsymbol{\kappa} \,|\, \delta_j)\,. \tag{22}$$

We emphasize the dependence of M on $\boldsymbol{\kappa}$, see eq. (15). The normalization constant $\mathcal{I}$ is not required to generate samples if using the *Metropolis-Hastings* (MH) algorithm [14] (see also the appendix).

The fact that the pdf $\phi(\boldsymbol{\kappa})$ enters eq. (21) and hence estimates of $\rho(\boldsymbol{\sigma}_{\text{exp}} \,|\, \delta_m)$ depend on $\mathcal{I}$ is not important. Since the normalization $\mathcal{I}$ affects each estimate in the same way, its value does not influence the relative likelihoods.

Due to the form of eq. (22) and due to $\rho(\boldsymbol{\sigma}_{\text{exp}} \,|\, \boldsymbol{\kappa}_i) = \mathcal{N}(\boldsymbol{\sigma}_{\text{exp}} \,|\, S\,\boldsymbol{y}_0,\ M(\boldsymbol{\kappa}))$, the estimate in eq. (21) simplifies to

$$\rho(\boldsymbol{\sigma}_{\text{exp}} \,|\, \delta_m) \approx \frac{1}{\mathcal{I}K} \sum_{i=1}^{K} \omega_m(\kappa_i) \tag{23}$$

with the abbreviation

$$\omega_m(\kappa_i) = \frac{\rho(\boldsymbol{\kappa}_i \,|\, \delta_m)}{\sum_j \rho(\boldsymbol{\kappa}_i \,|\, \delta_j)}\,. \tag{24}$$

The value $\delta_{m'}$ associated with the biggest value $\rho(\boldsymbol{\sigma}_{\text{exp}} \,|\, \delta_{m'})$ should be selected for the analysis.

E. Overall Prediction and Covariance Matrix

Besides finding consistent subsets of experiments, which is a question of maximizing $\rho(\boldsymbol{\sigma}_{\text{exp}} \,|\, \delta)$ given in eq. (15), we may be interested in an overall prediction and the associated covariance matrix by averaging over all possible interpretations. Technically, to find the overall prediction, we have to solve

$$\begin{aligned} \hat{\boldsymbol{y}} = \mathbb{E}[\boldsymbol{y}] &= \int\int \boldsymbol{y}\, \rho(\boldsymbol{y}, \boldsymbol{\kappa} \,|\, \boldsymbol{\sigma}_{\text{exp}})\, \mathrm{d}\boldsymbol{y}\, \mathrm{d}\boldsymbol{\kappa} \\ &= \int \left(\int \boldsymbol{y}\, \rho(\boldsymbol{y} \,|\, \boldsymbol{\kappa}, \boldsymbol{\sigma}_{\text{exp}})\, \mathrm{d}\boldsymbol{y} \right) \rho(\boldsymbol{\kappa} \,|\, \boldsymbol{\sigma}_{\text{exp}})\, \mathrm{d}\boldsymbol{\kappa}\,. \end{aligned} \tag{25}$$

The inner integral yields the expectation of $\boldsymbol{y}$ under $\rho(\boldsymbol{y} \,|\, \boldsymbol{\kappa}, \boldsymbol{\sigma}_{\text{exp}})$. This conditioned posterior pdf has the same functional form as the posterior pdf of the standard GLS method in eq. (7). For the latter distribution we know the result of the integral, which is eq. (8). Therefore, the result of the inner integral in eq. (27) is given by

$$\boldsymbol{y}_1(\boldsymbol{\kappa}) = \boldsymbol{y}_0 + A_0 S^T \left(S A_0 S^T + C(\boldsymbol{\kappa}) \right)^{-1} (\boldsymbol{\sigma}_{\text{exp}} - S\,\boldsymbol{y}_0)\,. \tag{26}$$

If the computation of the inverse matrix is infeasible, the form of eq. (10) can be used. Using this analytic expression, the integral in eq. (25) takes the form

$$\hat{\boldsymbol{y}} = \mathbb{E}[\boldsymbol{y}] = \int \boldsymbol{y}_1(\boldsymbol{\kappa})\, \rho(\boldsymbol{\kappa} \,|\, \boldsymbol{\sigma}_{\text{exp}})\, \mathrm{d}\boldsymbol{\kappa}\,. \tag{27}$$

Likely no analytic solution exists for this remaining integral and we have to take recourse to Monte Carlo integration.

The simplification of the integral for the overall covariance matrix follows analogous steps. The overall covariance matrix can be written as

$$\begin{aligned} \hat{\Sigma} &= \mathbb{E}[\boldsymbol{y}\boldsymbol{y}^T] - \mathbb{E}[\boldsymbol{y}]\mathbb{E}[\boldsymbol{y}^T] \\ &= \int \left(\boldsymbol{y}\boldsymbol{y}^T - \hat{\boldsymbol{y}}\hat{\boldsymbol{y}}^T \right) \rho(\boldsymbol{y}, \boldsymbol{\kappa} \,|\, \boldsymbol{\sigma}_{\text{exp}})\, \mathrm{d}\boldsymbol{y}\, \mathrm{d}\boldsymbol{\kappa} \\ &= \int \left(\int \boldsymbol{y}\boldsymbol{y}^T\, \rho(\boldsymbol{y} \,|\, \boldsymbol{\kappa}, \boldsymbol{\sigma}_{\text{exp}})\, \mathrm{d}\boldsymbol{y} \right) \rho(\boldsymbol{\kappa} \,|\, \boldsymbol{\sigma}_{\text{exp}})\, \mathrm{d}\boldsymbol{\kappa} - \hat{\boldsymbol{y}}\hat{\boldsymbol{y}}^T\,. \end{aligned} \tag{28}$$

Using the identity

$$A_1 = \int \boldsymbol{y}\boldsymbol{y}^T\, \rho(\boldsymbol{y} \,|\, \boldsymbol{\kappa}, \boldsymbol{\sigma}_{\text{exp}})\, \mathrm{d}\boldsymbol{y}\ - \boldsymbol{y}_1(\boldsymbol{\kappa})\boldsymbol{y}_1(\boldsymbol{\kappa})^T \tag{29}$$

whose solution is analogous to the standard GLS method, see eq. (9),

$$A_1(\boldsymbol{\kappa}) = A_0 - A_0 S^T \left(S A_0 S^T + C(\boldsymbol{\kappa}) \right)^{-1} S A_0\,, \tag{30}$$

we can express eq. (28) as

$$\hat{\Sigma} = \int \left(A_1(\boldsymbol{\kappa}) + \boldsymbol{y}_1(\boldsymbol{\kappa})\boldsymbol{y}_1(\boldsymbol{\kappa})^T \right) \rho(\boldsymbol{\kappa} \,|\, \boldsymbol{\sigma}_{\text{exp}})\, \mathrm{d}\boldsymbol{\kappa} - \hat{\boldsymbol{y}}\hat{\boldsymbol{y}}^T\,. \tag{31}$$

As for the overall prediction, also this integral likely has no analytic solution.

Consequently, the integrals for the overall prediction in eq. (27) and the overall covariance matrix in eq. (31) have to be solved by means of Monte Carlo integration. One possibility is to obtain a sample $\boldsymbol{\kappa}_1, \cdots, \boldsymbol{\kappa}_K$ from $\rho(\boldsymbol{\kappa} \,|\, \boldsymbol{\sigma}_{\text{exp}})$ using the Metropolis-Hastings algorithm (see the appendix) and to approximate the integrals in terms of mean values. The approximations for the overall prediction and covariance matrix are then

$$\hat{\boldsymbol{y}} \approx \frac{1}{K} \sum_{i=1}^{K} \boldsymbol{y}_1(\boldsymbol{\kappa}_i)\,, \text{ and} \tag{32}$$

$$\hat{\Sigma} \approx \frac{1}{K} \sum_{i=1}^{K} \left(A_1(\boldsymbol{\kappa}) + \boldsymbol{y}_1(\boldsymbol{\kappa})\boldsymbol{y}_1(\boldsymbol{\kappa})^T \right) - \hat{\boldsymbol{y}}\hat{\boldsymbol{y}}^T\,. \tag{33}$$

However, if we have determined the most likely standard deviation δ for the multivariate Laplace prior according to the sampling procedure outlined in section D., there is an alternative route. We can reuse the samples $\kappa_1, \cdots, \kappa_K$ drawn

from the mixture pdf $\phi(\boldsymbol{\kappa})$ specified in eq. (22). Using the notation $\omega_m(\kappa_i)$ introduced in eq. (24), the approximations are given by

$$\hat{\boldsymbol{y}} \approx \frac{1}{\mathcal{J}K} \sum_{i=1}^{K} \omega_m(\kappa_i)\, \boldsymbol{y}(\boldsymbol{\kappa}_i)\,, \text{ and} \tag{34}$$

$$\hat{\Sigma} \approx \frac{1}{\mathcal{J}K} \sum_{i=1}^{K} \omega_m(\kappa_i) \left(A_1(\boldsymbol{\kappa}) + \boldsymbol{y}_1(\boldsymbol{\kappa})\boldsymbol{y}_1(\boldsymbol{\kappa})^T\right) - \hat{\boldsymbol{y}}\hat{\boldsymbol{y}}^T\,. \tag{35}$$

The index m refers to the value δ_m that has been selected as the most likely candidate. The unknown normalization constant $\mathcal{J}$ can be estimated by

$$\mathcal{J} = \frac{1}{K} \sum_{i=1}^{K} \omega_m(\kappa_i)\,. \tag{36}$$

Please note that this normalization constant is not identical to $\mathcal{I}$ of eq. (24) because it is also determined by the unknown normalization of the posterior pdf $\rho(\boldsymbol{\kappa}\,|\,\boldsymbol{\sigma}_{\text{exp}})$.

The described scheme of approximation is known as *self-normalized importance sampling* in the statistics literature, e.g. [13, p. 131]. The approach to solve some integrations of a multi-dimensional integral analytically and to use Monte Carlo sampling to evaluate the remaining integrals is termed as *Conditional Monte Carlo* in [13, p. 125].

F. Efficient Computation

The identification of plausible covariance matrix parameters $\boldsymbol{\kappa}$ (e.g. normalization uncertainties) is a question of maximizing $\rho(\boldsymbol{\kappa}\,|\,\boldsymbol{\sigma}_{\text{exp}}) \propto \rho(\boldsymbol{\sigma}_{\text{exp}}\,|\,\boldsymbol{\kappa}) \times \rho(\boldsymbol{\kappa})$ given in eq. (15). Also determining an overall prediction and the associated covariance matrix involves the evaluation of this pdf. The prior pdf $\rho(\boldsymbol{\kappa})$ can be calculated quickly if opting for a product of Laplace or normal pdfs. Contrary to that, the computation of the likelihood $\rho(\boldsymbol{\sigma}_{\text{exp}}\,|\,\boldsymbol{\kappa}) = \mathcal{N}(\boldsymbol{\sigma}_{\text{exp}}\,|\,S\,\boldsymbol{y}_0,\ M)$ may be computationally expensive. Inspecting the form of this multivariate normal pdf,

$$\begin{aligned}\ln \mathcal{N}(\boldsymbol{\sigma}_{\text{exp}}\,|\,S\,\boldsymbol{y}_0, M(\boldsymbol{\kappa})) = &-\frac{N}{2}\ln(2\pi) - \frac{1}{2}\ln\det M(\boldsymbol{\kappa}) \\ &- \frac{1}{2}(\boldsymbol{\sigma}_{\text{exp}} - S\,\boldsymbol{y}_0)^T\,(M(\boldsymbol{\kappa}))^{-1}\,(\boldsymbol{\sigma}_{\text{exp}} - S\,\boldsymbol{y}_0)\end{aligned} \tag{37}$$

with $M(\boldsymbol{\kappa}) = S A_0 S^T + C(\boldsymbol{\kappa})$, we see that the expensive operations are the calculation of the determinant and the inversion of the matrix M. The dimension of this matrix is determined by the total number of experiment data points. The time to invert a $10^4 \times 10^4$ matrix may be tens of seconds on a contemporary personal computer. In addition, numerical maximization and the generation of a Monte Carlo chain require at least thousands of function evaluations. Clearly, to *efficiently* compute $\rho(\boldsymbol{\sigma}_{\text{exp}}\,|\,\boldsymbol{\kappa})$ is important. This section explains therefore the efficient computation of $\ln \mathcal{N}(\boldsymbol{\sigma}_{\text{exp}}\,|\,S\,\boldsymbol{y}_0,\ M)$ and its gradient $\mathrm{d}(\ln \rho(\boldsymbol{\kappa}\,|\,\boldsymbol{\sigma}_{\text{exp}}))/\,\mathrm{d}\boldsymbol{\kappa}$. Having an analytic expression for the gradient offers great benefits in numerical maximization.

Inverse matrices will often appear in the discussion, so we use the notation $\tilde{X}$ instead of X^{-1} to save space. Further, we just write C and M from now on, but the dependence on $\boldsymbol{\kappa}$ should be kept in mind. Using the Woodbury identity (eq. (B.15) in the appendix), we express the inverse of M as

$$\tilde{M} = \tilde{C} - \tilde{C}S\left(\tilde{A}_0 + S^T\tilde{C}S\right)^{-1} S^T\tilde{C}\,. \tag{38}$$

The measurement vector $\boldsymbol{\sigma}_{\text{exp}}$ is partitioned into subvectors $\boldsymbol{\sigma}_{\text{exp},i}$ associated with different experiments. For each $\boldsymbol{\sigma}_{\text{exp},i}$ there is a sensitivity matrix S_i to map from model parameters to the respective predictions. Therefore, the sensitivity matrix is partitioned into $S = (S_1^T, \cdots, S_N^T)^T$. Exploiting this partitioned form and the the block diagonal structure of $\tilde{C}$ allows us to write

$$S^T\tilde{C}S \;=\; \sum_k S_k^T\tilde{C}_k S_k\,. \tag{39}$$

Because the number of data points in one data set is usually limited, say less than hundred, the computation of the inverse matrices $\tilde{C}_k$ can be performed fast on contemporary personal computers. The sum of matrices in the brackets in eq. (38) leads to a matrix of the same dimension as $\tilde{A}_0$, hence it is determined by the number of model parameters. I expect models or series expansions not to have more than hundreds of adjustable parameters.

The inverse matrix $\tilde{M}$ appears only in the matrix product $\boldsymbol{u}^T\tilde{M}\boldsymbol{u}$ with $\boldsymbol{u} = \boldsymbol{\sigma}_{\text{exp}} - S\,\boldsymbol{y}_0$. Also $\boldsymbol{u}$ is partitioned into subvectors $\boldsymbol{u}_i = \boldsymbol{\sigma}_{\text{exp},i} - S_i\boldsymbol{y}_0$. Noting that

$$\tilde{C}\boldsymbol{u} = \left(\left(\tilde{C}_1\boldsymbol{u}_1\right)^T, \cdots, \left(\tilde{C}_N\boldsymbol{u}_N\right)^T\right)^T \tag{40}$$

is a vector and considering the form of eq. (38), we see that $\boldsymbol{u}^T\tilde{M}\boldsymbol{u}$ can be completely evaluated in terms of computationally cheap matrix-vector products.

To tackle the determinant, we use the matrix determinant lemma (eq. (B.14) in the appendix) to obtain

$$\begin{aligned}\ln|M| &= \ln|\tilde{A}_0 + S^T\tilde{C}S| + \ln|C| + \ln|A_0| = \\ &= \ln\left|\tilde{A}_0 + \sum_k S_k^T\tilde{C}_k S_k\right| + \sum_k \ln|\tilde{C}_k| + \ln|\tilde{A}_0|\,,\end{aligned} \tag{41}$$

with $|X|$ being the short-hand notation for $\det X$. To get from the first to the second line, we used eq. (39) and the fact that the determinant of a block diagonal matrix is the product of the block determinants. As elaborated above, determinants have to be taken only from comparatively low dimensional matrices.

Finally, we briefly discuss how to calculate the gradient of $\ln\rho(\boldsymbol{\kappa}\,|\,\boldsymbol{\sigma}_{\text{exp}})$. Blocks of $\tilde{M}$ are given by

$$\tilde{M}_{ij} = \delta_{ij}\tilde{C}_i - \tilde{C}_i S_i\left(\tilde{A}_0 + \sum_k S_k^T\tilde{C}_k S_k\right)^{-1} S_j^T\tilde{C}_j\,, \tag{42}$$

where $\delta_{ij} = 1$ for $i = j$ and $\delta_{ij} = 0$ for $i \neq j$. The derivative of $\ln|M|$ in eq. (37) can be written as (eq. (B.10) in the appendix)

$$\frac{\partial \ln|M|}{\partial\kappa_i} = \mathrm{Tr}\left[\tilde{M}\frac{\partial C}{\partial\kappa_i}\right] = \mathrm{Tr}\left[\tilde{M}_{ii}\frac{\partial C_i}{\partial\kappa_i}\right]. \tag{43}$$

The partial derivatives of the matrix product with respect to the parameters κ_i are (eq. (B.5) in the appendix)

$$\frac{\partial \boldsymbol{u}^T \tilde{M} \boldsymbol{u}}{\partial \kappa_i} = -\boldsymbol{u}^T \tilde{M} \frac{\partial C}{\partial \kappa_i} \tilde{M} \boldsymbol{u} = -\boldsymbol{u}^T \tilde{M}_{\cdot i} \frac{\partial C_i}{\partial \kappa_i} \tilde{M}_{i \cdot} \boldsymbol{u} \,. \tag{44}$$

A point in the index of a matrix denotes the inclusion of all rows or columns. Again, exploiting the partitioned form in eq. (40), the evaluation of this quantity only involves computationally inexpensive matrix-vector products. The inner derivative completes the determination of the gradient. For the normalization uncertainty defined in eq. (12), we get

$$\frac{\partial C_i}{\partial \kappa_i} = 2\kappa_i \left(\boldsymbol{\sigma}_{\mathrm{exp},i} \boldsymbol{\sigma}_{\mathrm{exp},i}^T \right) . \tag{45}$$

Now, equipped with an analytic expression for the gradient $(\mathrm{d}/\mathrm{d}\boldsymbol{\kappa})(\ln \mathcal{N}(\boldsymbol{\sigma}_{\mathrm{exp}} \,|\, S\,\boldsymbol{y}_0, M(\boldsymbol{\kappa})))$, the full gradient $\mathrm{d}(\ln \rho(\boldsymbol{\kappa} \,|\, \boldsymbol{\sigma}_{\mathrm{exp}}))/\,\mathrm{d}\boldsymbol{\kappa}$ is straight-forward to compute. Considering the complete log-posterior pdf

$$\ln \rho(\boldsymbol{\kappa} \,|\, \boldsymbol{\sigma}_{\mathrm{exp}}) \overset{C}{=} \ln \rho(\boldsymbol{\sigma}_{\mathrm{exp}} \,|\, \boldsymbol{\kappa}) + \ln \rho(\boldsymbol{\kappa}) \,, \tag{46}$$

we just have to add $\mathrm{d}(\ln \rho(\boldsymbol{\kappa}))/\,\mathrm{d}\boldsymbol{\kappa}$. In the case of identical Laplace pdfs, see eq. (18), the components of the latter gradient are

$$\frac{\partial \ln \rho(\boldsymbol{\kappa})}{\partial \kappa_i} = -\frac{\sqrt{2}}{\delta} \mathrm{sign}(\kappa_i) \tag{47}$$

with $\mathrm{sign}(\kappa_i)$ being either -1 or $+1$ according to the sign of κ_i.

In summary, this section elaborated on the efficient computation of $\ln \rho(\boldsymbol{\kappa} \,|\, \boldsymbol{\sigma}_{\mathrm{exp}})$ and its gradient. The inversion and the determinant of the potentially large matrix M have been transformed to the same operations on the comparatively small matrix $\left(\tilde{A}_0 + S^T \tilde{C} S\right)$. The size of the latter matrix is determined by the number of model parameters. Due to the block-diagonal structure of C, the inversion can be performed fast and yields another block-diagonal matrix. The resulting matrix $\tilde{C}$ only enters an inexpensive matrix-vector product whose evaluation profits again from the block-diagonal structure of $\tilde{C}$. For the same reasons, the analytic expression of the gradient can be also computed quickly.

Finally, the availability of the gradient enables the application of gradient-based optimization algorithms, such as the *BFGS* [15] or *L-BFGS* algorithm [16]. Especially the latter algorithm is very memory efficient and hence suited for a scenario with many experiment data sets. As another useful feature, it allows the specification of parameter boundaries.

III. DEMONSTRATION AND DISCUSSION

The method will be demonstrated at the example of eleven data sets taken from [17] with measurements of the proton-neutron total cross section. I selected the data points with incident momenta (in the laboratory frame) between 0.5 GeV/c and 5 GeV/c because many discrepant data sets are available in this range. I assume that only statistical uncertainties are present, which leads to a diagonal matrix B. Correlated errors, such as an uncertainty about the detector efficiency, could also be included, but the respective information is not always available in nuclear databases. The experiment data are shown

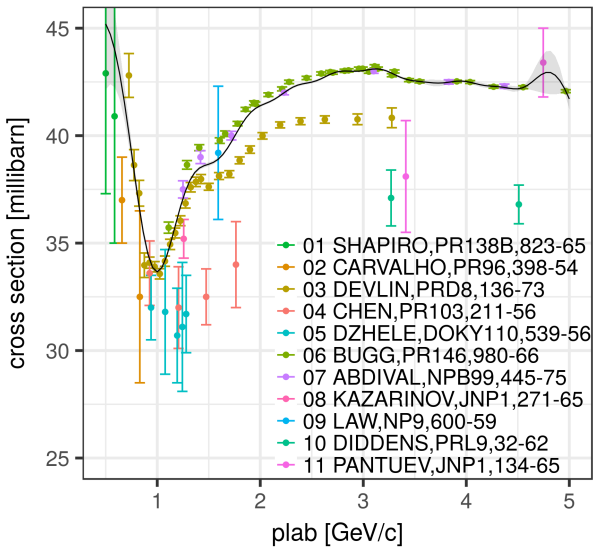

Fig. 2. Experiment data used in the schematic evaluation. The black line is the resulting prediction from a GLS fit of the original data without any correction of the uncertainty assumptions.

in fig. 2. Considering the extent of the error bars indicating the 68% confidence interval, the data are clearly inconsistent.

The series expansion introduced in eq. (1) with fifty expansion terms provides the model to fit the data. The exact specification employed in this section is given by

$$\boldsymbol{\sigma}_{\mathrm{fit}}(E) = \frac{\sum_{i=1}^{50} y_i \mathcal{N}\left(E \,|\, x_i, \lambda^2\right)}{\sum_{i=1}^{50} \mathcal{N}\left(E \,|\, x_i, \lambda^2\right)} \tag{48}$$

with $\lambda = 0.2$ and $x_j = 0.2 + i \times (5 - 0.5)/50$. This model imposes a certain degree of smoothness on the cross section curve but besides that can adapt flexibly to the data.

The prior $\rho(\boldsymbol{y} \,|\, \boldsymbol{y}_0, A_0)$ for the parameter vector $\boldsymbol{y}$ is a multivariate normal distribution $\mathcal{N}(\boldsymbol{y} \,|\, \boldsymbol{y}_0, A_0)$ with all elements in $\boldsymbol{y}_0$ equal forty. The associated prior covariance matrix A_0 is diagonal with all elements equal thousand. This prior covers well the experiment data.

Using the standard GLS method to fit the model yields the curve illustrated in fig. 2. The 68% error band is hardly visible at most energies. Further, the fit runs in between the data sets around 1.5 GeV/c and the associated uncertainty band excludes them. This observation is associated with the presence of inconsistent data. The result of the GLS method serves as a reference to which the results of the proposed method can be compared.

In the first part of the demonstration, blocks of the adjusted covariance matrix $C(\boldsymbol{\kappa})$ are parameterized as

$$C_i(\kappa_i) = B_i + \kappa_i^2 \boldsymbol{\sigma}_{\mathrm{exp},i} \boldsymbol{\sigma}_{\mathrm{exp},i}^T \,. \tag{49}$$

This parameterization introduces an additional normalization uncertainty κ_i for each experiment data set. Therefore, the vector $\boldsymbol{\kappa}$ contains eleven variables. Afterward, the method will be also applied with a more flexible energy-dependent parameterization.

Because normalization uncertainties κ_i are themselves uncertain, we need to specify a prior pdf $\rho(\boldsymbol{\kappa}\,|\,\delta)$. I tested the method with the following three prior specifications:

$$\rho_{\mathrm{U}}(\boldsymbol{\kappa}) = \mathrm{const}\,, \tag{50}$$

$$\rho_{\mathrm{N}}(\boldsymbol{\kappa}\,|\,\delta) = \prod_{i=1}^{11} \frac{1}{\sqrt{2\pi}\,\delta} \exp\left(-\frac{1}{2}\frac{\kappa_i^2}{\delta^2}\right), \tag{51}$$

$$\rho_{\mathrm{L}}(\boldsymbol{\kappa}\,|\,\delta) = \prod_{i=1}^{11} \frac{1}{\sqrt{2}\,\delta} \exp\left(-\frac{\sqrt{2}\,|\kappa_i|}{\delta}\right). \tag{52}$$

The first pdf is an improper uniform pdf. Using this prior pdf, the posterior pdf $\rho(\boldsymbol{\kappa}\,|\,\boldsymbol{\sigma}_{\mathrm{exp}})$ is exclusively determined by the marginal likelihood $\rho(\boldsymbol{\sigma}_{\mathrm{exp}}\,|\,\boldsymbol{\kappa})$. The second pdf is a product of identical normal distribution and the third pdf a product of identical Laplace pdfs. The parameter δ signifies in both cases the standard deviation of the distribution.

1. Selection of δ

In order to carry out the method with either $\rho_{\mathrm{N}}(\boldsymbol{\kappa}\,|\,\delta)$ or $\rho_{\mathrm{L}}(\boldsymbol{\kappa}\,|\,\delta)$, a suitable δ has to be selected. Linked to these prior pdfs are the following mixture pdfs:

$$\phi_{\mathrm{N}}(\boldsymbol{\kappa}) = \mathcal{I}_{\mathrm{N}} \times \mathcal{N}(\boldsymbol{\sigma}_{\mathrm{exp}}\,|\,S\,\mathbf{y}_0,\ M(\boldsymbol{\kappa})) \times \sum_{j=1}^{30} \rho_{\mathrm{N}}(\boldsymbol{\kappa}\,|\,\delta_j)\,, \tag{53}$$

$$\phi_{\mathrm{L}}(\boldsymbol{\kappa}) = \mathcal{I}_{\mathrm{L}} \times \mathcal{N}(\boldsymbol{\sigma}_{\mathrm{exp}}\,|\,S\,\mathbf{y}_0,\ M(\boldsymbol{\kappa})) \times \sum_{j=1}^{30} \rho_{\mathrm{L}}(\boldsymbol{\kappa}\,|\,\delta_j)\,. \tag{54}$$

The functional form of these pdfs was introduced in eq. (22). The components are characterized by $\delta_j = 0.01 \times j$ and the normalization constants $\mathcal{I}_{\mathrm{N}}, \mathcal{I}_{\mathrm{L}}$ are set to one. Considering fig. 2, the appropriate value of δ is somewhere between 1% and 30% and hence the form of the mixture pdfs justified.

Next, a sample of each mixture has to be obtained by means of the Metropolis-Hastings algorithm. I employed $\psi(\boldsymbol{\kappa}'\,|\,\boldsymbol{\kappa}) = \mathcal{N}(\boldsymbol{\kappa}'\,|\,\boldsymbol{\kappa}, \tau^2\mathbb{1})$ as proposal pdf. After tentative runs of the MH algorithm with different values of τ, the assignment $\tau = 0.045$ turned out to be a good choice yielding acceptance rates around 30% for both mixture pdfs. Unless otherwise stated, this proposal distribution is employed throughout the demonstration. In principle, the choice of τ could be automated, too, e.g. [18]. Investigation in this respect is left as future work.

Finally, I created for each mixture pdf a Monte Carlo chain with one million vectors. The evolution of $\log \phi_{\mathrm{L}}(\boldsymbol{\kappa})$ as a function of the iteration count is illustrated in fig. 3. No obvious drift can be noticed, which gives evidence that the MH chain represents a sample from $\phi_{\mathrm{L}}(\boldsymbol{\kappa})$. The density evolution plot for $\log \phi_{\mathrm{N}}(\boldsymbol{\kappa})$ looked similar without any sign of drifting.

Calculating the marginal likelihood $\rho(\boldsymbol{\sigma}_{\mathrm{exp}}\,|\,\delta_j)$ according to eq. (23) for all δ_j in the mixture pdf $\phi_{\mathrm{L}}(\boldsymbol{\kappa})$, we learn that the maximum appears at $\delta_{\mathrm{L}} = 0.13$. Hence, this value should be used in the procedure. The respective value in the case of $\phi_{\mathrm{N}}(\boldsymbol{\kappa})$ was $\delta_{\mathrm{N}} = 0.11$.

One may be concerned that the peak is rather flat and the relative likelihoods associated with δ values in vicinity are similar, as visualized by the green line in fig. 4. This observation

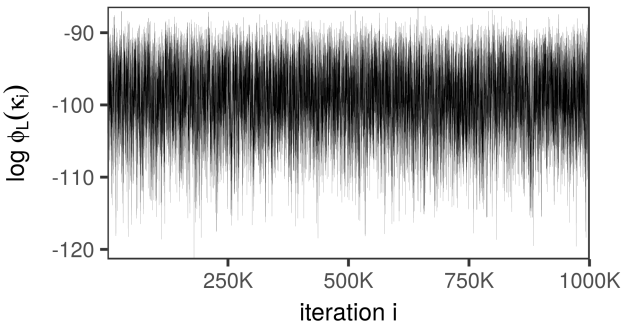

Fig. 3. Evolution of $\log \phi_{\mathrm{L}}(\boldsymbol{\kappa})$, see eq. (54), in the process of MH sampling.

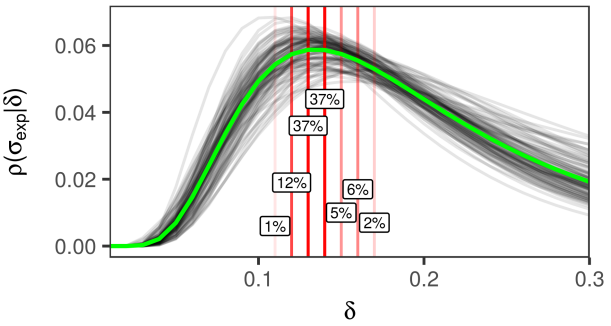

Fig. 4. Fluctuations in the estimate of the parameter δ that maximizes $\rho_{\mathrm{L}}(\boldsymbol{\sigma}_{\mathrm{exp}}\,|\,\delta)$. Red vertical lines and overlaid percentage numbers indicate the proportion of cases that led to the estimate at the respective value of δ.

begs the question of how strong the position of the maximum fluctuates if estimated from a smaller chain. To address this question, I cut the MC chain into chunks consisting of 10^4 vectors and estimated the relative likelihoods and the position of the maximum on the basis of each chunk. The ensemble of black curves in fig. 4 conveys an impression of the variations in the relative likelihoods. The red vertical lines denote which δ values were identified as maxima according to the chunks. The overlaid percentages display the proportion of chunks with the respective location as maximum. In spite of the rather large fluctuations of the relative likelihoods, the maximum of δ was estimated to be either 0.13 or 0.14 in 74% of the cases. In all cases, the maximum was situated between 0.11 and 0.17. At first glance, this rather large spread suggests to always construct long chains—a time-consuming process. For example, the generation of one million vectors took about six hours on a contemporary personal computer. However, one reason for the large spread is the insensitivity of the likelihood $\rho(\boldsymbol{\sigma}_{\mathrm{exp}}\,|\,\boldsymbol{\kappa})$ to the choice of δ. This feature implies that the mean vector $\hat{\mathbf{y}}$ calculated from the posterior pdf according to eq. (27) is also rather insensitive to the exact value of δ. Thus, the fluctuations of a few percent are acceptable and a MC chain with 10^4 vectors seems (at least in the studied example) sufficient. Further evidence for the validity of this statement will be provided in the next sections.

2. Segmenting Experiment Data into Subsets

Consistent subsets of data can be found by maximizing the marginal posterior pdf $\rho(\boldsymbol{\kappa}\,|\,\boldsymbol{\sigma}_{\text{exp}})$. The three prior pdfs introduced in eqs. (50) to (52) lead to the following posterior pdfs:

$$\rho_{\text{U}}(\boldsymbol{\kappa}\,|\,\boldsymbol{\sigma}_{\text{exp}}) = \mathcal{I}_{\text{U}} \times \mathcal{N}(\boldsymbol{\sigma}_{\text{exp}}\,|\,S\,\boldsymbol{y}_0,\ M(\boldsymbol{\kappa})) \times \rho_{\text{U}}(\boldsymbol{\kappa})\,, \quad (55)$$

$$\rho_{\text{N}}(\boldsymbol{\kappa}\,|\,\boldsymbol{\sigma}_{\text{exp}}) = \mathcal{I}_{\text{N}} \times \mathcal{N}(\boldsymbol{\sigma}_{\text{exp}}\,|\,S\,\boldsymbol{y}_0,\ M(\boldsymbol{\kappa})) \times \rho_{\text{N}}(\boldsymbol{\kappa}\,|\,\delta_{\text{N}})\,, \quad (56)$$

$$\rho_{\text{L}}(\boldsymbol{\kappa}\,|\,\boldsymbol{\sigma}_{\text{exp}}) = \mathcal{I}_{\text{L}} \times \mathcal{N}(\boldsymbol{\sigma}_{\text{exp}}\,|\,S\,\boldsymbol{y}_0,\ M(\boldsymbol{\kappa})) \times \rho_{\text{L}}(\boldsymbol{\kappa}\,|\,\delta_{\text{L}})\,. \quad (57)$$

The normalization constants $\mathcal{I}_{\text{U}}, \mathcal{I}_{\text{N}}, \mathcal{I}_{\text{L}}$ are not required for the maximization. We can conveniently set them to one. The values $\delta_{\text{N}} = 0.11$ and $\delta_{\text{L}} = 0.13$ are taken from the previous section.

Because the posterior probability densities usually cover a huge range, which easily leads to a numerical over- or underflow, the numerical maximization was carried out for the logarithms of these pdfs. The L-BFGS-B algorithm [16] as implemented by the *optim* function in the statistical programming language *R* [19] proved to be reliable. This algorithm takes advantage of an analytic expression of the gradient, see the derivation starting from eq. (42). Furthermore, it permits the specification of box constraints. Box constraints can be used to effectively deal with proper uniform priors. Yet, more important for our case, we can constrain parameters to be positive and exclude zero as solution. Taking into account the form of eq. (49) and how it enters the multivariate normal pdf in eq. (37), we recognize the posterior pdfs to be symmetric around $\boldsymbol{\kappa} = \mathbf{0}$. This insight justifies the restriction to positive values. The exclusion of zero is important in the case of $\rho_{\text{L}}(\boldsymbol{\kappa}\,|\,\boldsymbol{\sigma}_{\text{exp}})$ as the gradient exhibits a discontinuity if some $\kappa_i = 0$. The L-BFGS-B algorithm relies on the gradient to be continuous, and thus discontinuities potentially cause problems.

Due to these reasons, I constrained all κ_i to lie between 1×10^{-4} and 5×10^{-1}. The upper bound protects against unreasonable solutions with normalization uncertainties greater than 50%. Concerning the overall setup of the L-BFGS-B algorithm, I limited the maximal number of iterations in the numerical maximization procedure to thousand and specified that the Hessian matrix should be estimated based on the precedent twenty iteration steps. For each maximization attempt, the vector $\boldsymbol{\kappa}$ was initialized with values drawn uniformly from the range between 1×10^{-4} and 5×10^{-1}. Each maximization attempt was repeated ten times to ensure that the global maximum has been indeed found.

The results of the numerical maximization are summarized in table I. The solutions based on ρ_{U} and ρ_{N} are comparable in structure. The same κ_i are set to zero, only the remaining κ_i are pulled closer to zero in the case of ρ_{N}. Albeit the slightly more constrained solution, the χ^2/N value associated with ρ_{N} is only a bit larger and still below one. This observation suggests that ρ_{N} should be preferred over ρ_{U}.

Comparing ρ_{N} and ρ_{L}, we can make the interesting observation that more parameters κ_i are set to zero in the case of ρ_{L} despite the standard deviation δ_{L} being larger than δ_{N}. Figure 1 visualized the reason for this behavior. The non-zero parameters of the two solutions are very similar. Further, both

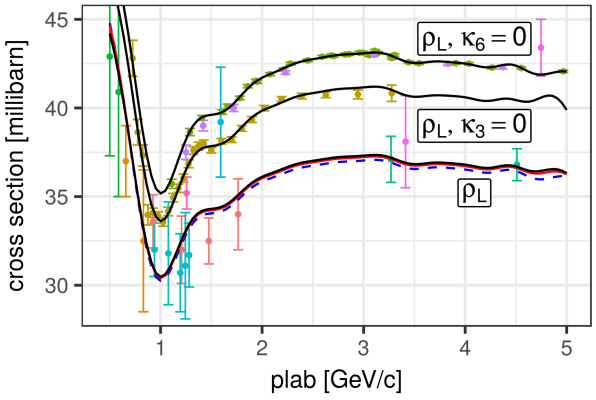

Fig. 5. The black line labelled with ρ_{L} shows the posterior maximum associated with the prior in eq. (52) with $\delta = 0.13$. The blue dashed line and the red solid line in vicinity are the maxima of ρ_{U} and ρ_{N} with $\delta = 0.11$. The upper two black lines are the maxima of ρ_{L} with $\delta = 0.13$ under the constraint that $\kappa_3 = 0$ and $\kappa_6 = 0$, respectively.

solutions are consistent because their χ^2/N values are close to one. Ockham's principle states that among the many explanations compatible with a certain observation, the explanation with the least assumptions should be chosen. According to this principle, the pdf ρ_{L} should be preferred over ρ_{N} because it favors sparse solutions.

Figure 4 showed the fluctuations in the determination of δ_{L}. To study the sensitivity of the maximum $\boldsymbol{\kappa}$ to the choice of δ_{L}, I performed the maximization also for ρ_{L} with $\delta_L = 0.17$. The result displayed in table I is hardly different from the result based on ρ_{L} with $\delta_L = 0.13$. Consequently, as already conjectured in the previous section, the fluctuations of δ do not significantly alter the results.

The discussion so far provided arguments in favor of $\rho_{\text{L}}(\boldsymbol{\kappa})$ as prior pdf. Please note, however, that all prior specifications led to acceptable values of χ^2/N. If the χ^2 value really comes from a χ^2-distribution, then its standard deviation is $\sigma = \sqrt{2N}$. Because there are 84 experiment data points in total, we get $\sigma/N = 0.15$. The associated 95% interval $[0.7, 1.3]$ includes all observed χ^2/N values. This fact indicates that the method is effectively able to *correct* the uncertainty assumptions of the experiments. Contrary to that, the GLS fit of the uncorrected data is associated with the too large value $\chi^2/N = 16.13$.

The *segmentation of data sets* can be performed by removing data sets whose normalization constants exceed a certain threshold. We can assume that one (or several) data sets are correct and fix their κ_i at zero during the optimization. Table I shows this constrained optimization for ρ_{L} with $\delta = 0.13$ and either $\kappa_3 = 0$ or $\kappa_6 = 0$ fixed. The predictions of the GLS method, see eq. (26), using the obtained vectors $\boldsymbol{\kappa}$ are depicted in fig. 5. In contrast to the GLS fit of the uncorrected data shown in fig. 2, these fits agree well with the data sets that were *a priori* assumed to be correct.

Finally, one mode of failure must be mentioned. If a normalization uncertainty for each data set is not sufficient to achieve consistency among data sets, we may also get unfavorable results with the new approach. For example, just

	δ	ℓ	κ_1	κ_2	κ_3	κ_4	κ_5	κ_6	κ_7	κ_8	κ_9	κ_{10}	κ_{11}	χ^2/N
GLS	—	—	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	16.13
ρ_U	—	—	**.00**	.04	.10	**.00**	**.00**	.14	.13	.06	.10	**.00**	.13	0.73
ρ_N	.11	—	**.00**	.03	.07	**.00**	**.00**	.10	.09	.04	.05	**.00**	.09	0.79
ρ_L	.13	1.0×10^{0}	**.00**	**.00**	.07	**.00**	**.00**	.09	.09	.03	**.00**	**.00**	.08	0.82
ρ_L	.13	1.8×10^{-3}	**.00**	.09	**.00**	.07	.07	.03	.03	.01	**.00**	.07	**.00**	0.98
ρ_L	.13	1.2×10^{-5}	**.00**	.12	.03	.10	.10	**.00**	**.00**	.06	**.00**	.10	**.00**	1.07
ρ_L	.17	—	**.00**	**.00**	.07	**.00**	**.00**	.10	.09	.04	**.00**	**.00**	.09	0.81

TABLE I. Posterior maxima $\boldsymbol{\kappa}$ based on the prior distributions specified in eqs. (50) to (52). For the pdfs ρ_N and ρ_L, results based on different values of δ are presented. Square brackets denote that the respective κ_i was fixed at zero. The index i refers to the experiment data set, see fig. 2. The value χ^2/N is the result of eq. (19) divided by the number of data points. Relative likelihoods ℓ are stated for the case ρ_L with $\delta = 0.13$.

as in the standard GLS fit of the uncorrected data, the prediction could run in between the data sets and exhibit too small uncertainties.

Nevertheless, even in this scenario, we may still use the method in combination with normalization uncertainties as exploration technique. Even though a normalization uncertainty would not be enough to make the data sets consistent, very likely the method would still identify inconsistent data sets by introducing large normalization uncertainties.

Another option is to use a more flexible parametrization of the covariance matrix $C(\boldsymbol{\kappa})$, which is able to model more elaborate uncertainty assumptions. This approach will be discussed and demonstrated in section 4..

3. Overall Prediction and Covariance matrix

The additional layer of uncertainty about the normalization uncertainties κ_i also increases the uncertainty in the final estimates of the model parameters. The posterior pdf $\rho(\boldsymbol{\kappa}\,|\,\boldsymbol{\sigma}_{\text{exp}})$ provides the probability density for any choice of $\boldsymbol{\kappa}$. Each of these choices produces a different result in the GLS method. Therefore, the overall prediction is calculated as a weighted mean of all these results, see eq. (27). Analogously, also the overall covariance matrix is—loosely speaking—the weighted mean of the covariance matrices conditioned on the different choices of $\boldsymbol{\kappa}$, see eq. (31).

It turned out that the uniform prior ρ_U did not lead to reasonable solutions. The MC chain to draw samples from $\rho_U(\boldsymbol{\kappa}\,|\,\boldsymbol{\sigma}_{\text{exp}})$ given in eq. (56) did not reach its stationary distribution even after one million iterations. The employed proposal pdf $\psi(\boldsymbol{\kappa}'\,|\,\boldsymbol{\kappa}) = \mathcal{N}(\boldsymbol{\kappa}'\,|\,\boldsymbol{\kappa}, \tau^2\mathbb{1})$ with $\tau = 0.3$ lead to acceptance rate of 30% and more. The evolution of the chain is illustrated in fig. 6. I performed several runs of the MH algorithm, but the observed behavior persisted. The parameters κ_i acquired values in the order of hundred and sometimes even thousand. Normalization uncertainties of 10000% are clearly absurd in our scenario. Taking into account that the maximum of $\rho_U(\boldsymbol{\kappa}\,|\,\boldsymbol{\sigma}_{\text{exp}})$ shown in table I was reasonable, I conclude that the data alone do not sufficiently constrain the normalization uncertainties. Technically, the determinant in eq. (37) responsible for the decline of the probability density does not grow fast enough with increasing κ_i. Inspecting plots of the probability density as a function of the parameters κ_i confirmed this hypothesis.

The chains to draw samples from $\rho_N(\boldsymbol{\kappa}\,|\,\boldsymbol{\sigma}_{\text{exp}})$ with $\delta_N = 0.11$ and $\rho_L(\boldsymbol{\kappa}\,|\,\boldsymbol{\sigma}_{\text{exp}})$ with $\delta_L = 0.13$ behaved well. Because their predictions and associated uncertainties were visually indistinguishable, I only discuss the case of ρ_L. The overall prediction is illustrated in fig. 7. Interestingly, it coincides with the prediction associated with the posterior maximum, compare with fig. 5. Only the 1σ uncertainty band of the overall prediction is larger than that one of the posterior maximum prediction, which is the expected behavior due to averaging over the covariance matrices.

Two marginal posterior distributions $\rho(\kappa_i\,|\,\boldsymbol{\sigma}_{\text{exp}}), i \in \{1, 3\}$ are shown in fig. 8. These distributions are considerably right-skewed. The distribution for κ_3 rises sharply on the left and declines moderately on the right. The sharp rise is due to the term proportional to the negated χ^2-value in the second row of eq. (37). This term saturates for large enough vectors $\boldsymbol{\kappa}$. From this point onwards, the multivariate Laplace prior $\rho_L(\boldsymbol{\kappa})$ is mainly responsible for the decline. Without the Laplace prior, the decline would be driven only by the determinant in the first row of eq. (37). As was illustrated in fig. 6, the rate of decline in the latter case is too low. Please note that all these observations are specific for the assumption of a normalization uncertainty.

It may disturb to see the overall prediction running below the majority of the data points, but one should keep in mind that the method makes an assessment at the level of

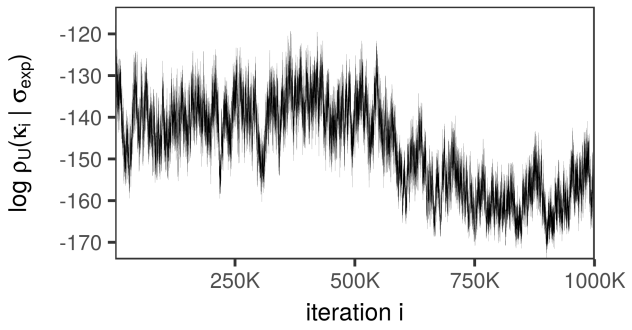

Fig. 6. Evolution of $\phi_U(\boldsymbol{\kappa}\,|\,\boldsymbol{\sigma}_{\text{exp}})$, see eq. (56), in the process of MH sampling.

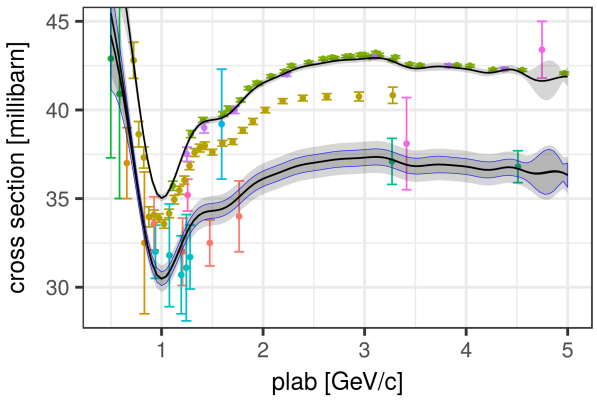

Fig. 7. Lower curve shows the overall prediction and associated 1σ-confidence band. The light-blue edge indicates the extent of uncertainty linked to the posterior maximum prediction. The upper curve shows the overall prediction under the constraint $\kappa_6 = 0.005$.

complete data sets. A data set with more points does not get more weight than one with less points, hence we may call the method 'democratic'. This feature is also reasonable from an evaluation point of view. A data set is a unit which usually comprises measurements from the same experiment. If one data point is affected by an unrecognized systematic error, very likely the other data points are too. This democratic feature is also backed up by the numbers. The solution corresponding to the overall prediction in the forth row of table I is indeed associated with the least number of non-zero normalization uncertainties.

Nevertheless, if we believe one of the data sets containing more points to be adequate, because it is more recent, comes with a detailed error analysis or measurements were performed with superior technology, we can account for this prior knowledge. As an example, I fixed the normalization uncertainty of Bugg to $\kappa_6 = 0.005$ and only allowed variations of the remaining parameters in the MH algorithm. The resulting prediction is also depicted in fig. 7.

4. Beyond a Normalization Uncertainty

Introducing a normalization uncertainty for each data set may not always be enough to correct inconsistencies. A χ^2/N value calculated according to eq. (19) significantly larger than one indicates such cases. We may use then a more general parametrization of the adjusted covariance matrix C. We recall the assumption of vanishing correlations between data sets, so the matrix C is block-diagonal with the blocks C_i. An example of a more general parametrization of the blocks is given by

$$C_{ijk}(\kappa_i, \lambda_i) = B_i + \kappa_i^2 \exp\left(-\frac{1}{2\lambda_i^2}(E_{ij} - E_{ik})^2\right)\sigma_{ij}\sigma_{ik}\,. \quad (58)$$

This form is motivated by the squared exponential function commonly used in Gaussian process regression, e.g. [20]. The quantities E_{ij} and E_{ik} denote the momentum of the j^{th} and k^{th} data point of the experiment data set associated with C_i.

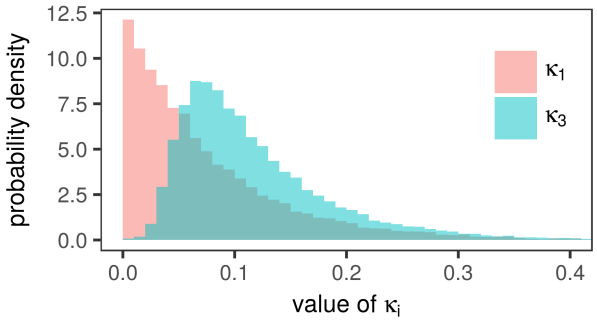

Fig. 8. Estimates of the marginal posterior distributions $\rho_{\text{L}}(\kappa_1 \,|\, \boldsymbol{\sigma}_{\text{exp}})$ and $\rho_{\text{L}}(\kappa_3 \,|\, \boldsymbol{\sigma}_{\text{exp}})$ obtained from a MH chain.

The notation for the measured cross sections σ_{ij} and σ_{ik} is analogous.

The variable κ_i denotes the relative standard deviation of the additional uncertainty component at all energies. The length-scale λ_i determines how quickly the *a priori* unknown error in the measured data points is allowed to change as a function of momentum. In the limit $\lambda_i \to \infty$, this parametrization is equivalent to the assumption of a normalization uncertainty, compare with eq. (49). Since the presented experiment data span a momentum range from 0.5 to 5 GeV/c, values beyond $\lambda_i = 20$ already resemble in good approximation a normalization uncertainty. The other extreme case $\lambda_i \to 0$ implements the assumption of white noise. Intermediate values are well suited to capture unrecognized momentum-dependent uncertainties, such as those related to detector efficiency.

Locating the posterior maximum profits from the availability of an analytic expression for the gradient. The computation of the gradient was discussed starting from eq. (42). It involved the partial derivatives of the adjusted covariance matrix C. For the parametrization in eq. (58), they are given by

$$\frac{\partial C_i(\kappa_i, \lambda_i)}{\partial \kappa_i} = 2\kappa_i \exp\left(-\frac{1}{2\lambda_i^2}(E_{ij} - E_{ik})^2\right)\sigma_{ij}\sigma_{ik} \quad (59)$$

$$\frac{\partial C_i(\kappa_i, \lambda_i)}{\partial \lambda_i} = \frac{\kappa_i^2}{\lambda_i^3}(E_{ij} - E_{ik})^2 \exp\left(-\frac{1}{2\lambda_i^2}(E_{ij} - E_{ik})^2\right)\sigma_{ij}\sigma_{ik} \quad (60)$$

The parameters κ_i here have the same meaning as the equally named parameters linked to the magnitude of the normalization uncertainty in the previous sections. Therefore, we can also impose the multivariate Laplace prior $\rho_{\text{L}}(\boldsymbol{\kappa} \,|\, \delta)$ in eq. (52) on them. Even though the automatic selection of δ could be done as for the normalization uncertainty, I just adopted the value $\delta_L = 0.13$ for the sake of simplicity.

Some testing indicated that the marginal likelihood is rather sensitive to a length-scale λ_i if the points of the respective data set $\mathcal{D}_i$ are dispersed over a broad momentum range. However, the marginal likelihood becomes insensitive if the length-scale is much larger than the momentum spread of the points. Owing to these two observations, I opted for a multivariate Laplace prior $\rho_{\text{L}}(\boldsymbol{\lambda} \,|\, \delta)$ with a large standard deviation $\delta = 100$. This choice ensures that the experiment data

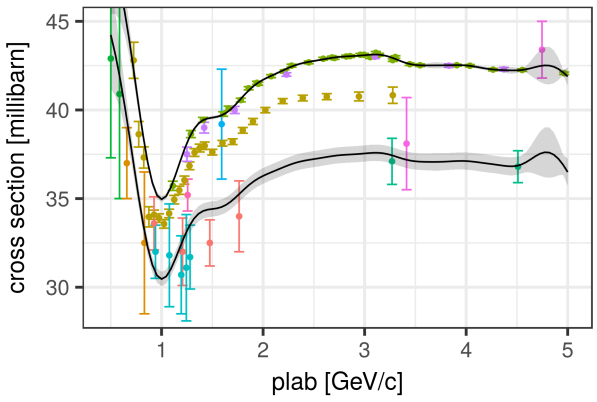

Fig. 9. Lower curve shows the maximum posterior prediction and 1σ-confidence band associated with the GP uncertainty in eq. (58) The upper curve shows the maximum posterior prediction under the constraint $\lambda_6 = 0.5$.

can dictate the length-scale if it matters. And whenever the length-scale becomes too large, the prior $\rho_L(\boldsymbol{\lambda}\,|\,\delta)$ regularizes the solution.

The maximization was again performed with the L-BFGS-B algorithm [16] with the same setup as described in section 2.. The parameters in $\boldsymbol{\kappa}$ were constrained to be between 10^{-4} and 5×10^{-1}. The parameters in $\boldsymbol{\lambda}$ were restricted to the interval between 10^{-1} and 20. I allowed all parameters in $\boldsymbol{\kappa}$ and $\boldsymbol{\lambda}$ to change. Initial values for the maximization were chosen uniformly between the parameter boundaries. I performed ten maximization attempts to ensure that a global maximum has been found.

The found vector $\boldsymbol{\kappa}$ was at the percent level identical to the solution in the case of normalization uncertainties, see the forth line of table I. Also the associated prediction illustrated in fig. 9 resembles that one in fig. 7. For the data sets where κ_i was driven towards zero, also the length-scale was driven towards the lower limit due to the influence of the prior. Besides one exception, all other data sets with non-zero κ_i obtained large length-scales greater than ten. Consequently, their uncertainty parametrization resembles a normalization uncertainty. This result is in agreement with Ockham's principle because a normalization uncertainty is already sufficient to reach consistency, see the χ^2/N values in table I, and a much simpler hypothesis than an energy-dependent uncertainty.

The exceptional case is the data set with $\kappa_{11} = 0.08$ and $\lambda_{11} = 0.89$. It contains the two pink points with large error bars on the right side of fig. 9. Visual inspection suggests that a normalization uncertainty may not be enough to make them consistent with the overall prediction, and likely some energy-dependent error source has to be considered. The method *automatically* inferred this hypothesis by the introduction of a short length-scale.

As a final example, I applied the method another time with the constraint that $\lambda_6 = 0.5$. The upper curve in fig. 9 depicts the result. Interestingly, the fixation of the length-scale λ_6 led to $\kappa_6 \approx 0$. The method determines that such a short-length scale is an overly complex hypothesis and therefore completely eliminates the additional uncertainty from the respective data set. This leads in turn to the introduction of short length-scales for some of the other data sets, see table II.

i	reference	#	κ_i	λ_i
1	SHAPIRO,PR138B,823-65	2	.000	.100
2	CARVALHO,PR96,398-54	2	.120	1.087
3	DEVLIN,PRD8,136-73	26	.038	5.762
4	CHEN,PR103,211-56	4	.113	.572
5	DZHELE,DOKY110,539-56	5	.113	.429
6	BUGG,PR146,980-66	32	.000	.500
7	ABDIVAL,NPB99,445-75	7	.006	1.783
8	KAZARINOV,JNP1,271-65	1	.060	.100
9	LAW,NP9,600-59	1	.000	.100
10	DIDDENS,PRL9,32-62	2	.100	8.902
11	PANTUEV,JNP1,134-65	2	.000	.100

TABLE II. Posterior maximum under the constraint $\lambda_6 = 0.5$ using the GP uncertainty in eq. (58). The column labeled # displays the number of points within each data set.

This last section should have made clear that the method is not bound to the assumption of a normalization uncertainty. Considering the flexibility to choose the uncertainty assumptions of the experiment data, we may regard the method better as a *framework* for inference. For instance, it would be possibly to include both a normalization uncertainty and the energy-dependent uncertainty introduced in eq. (58) to gain even more flexibility to adapt the uncertainty assumption of the experiments.

IV. SUMMARY AND OUTLOOK

A Bayesian method to fit models and evaluate nuclear data has been presented. The method accounts for inconsistencies between experiment data sets by modifying their uncertainty assumptions. The capability to correct the data has been demonstrated with an additional normalization uncertainty for each data set, and also with a more general energy-dependent uncertainty. Due to the freedom to flexibly choose the additional uncertainty structure, we may more appropriately call the method a framework.

Related to the correction of experiment data is the possibility to segment them into consistent subsets. Data sets whose additional uncertainty is beyond acceptable limits can be removed, so that the remaining data sets are coherent with each other. In that respect, the multivariate Laplace prior proved to be superior over the multivariate normal prior and the uniform prior because it favors sparse solutions.

The method also allows to compute an overall prediction and the associated covariance matrix by averaging over different interpretations. The associated uncertainties are enlarged compared to the standard GLS method, which counteracts the common problem of too small uncertainties.

It has been shown that potentially costly operations, such as the inversion of an $N \times N$ matrix with N being the total number of data points, can be performed efficiently by exploiting some matrix identities. Owing to this acceleration, the method is foreseen to work with a large corpus of data sets, and hence may be used in nuclear databases to detect inconsistencies in

an automated fashion.

Future work includes the application of the method using other series expansions and physics models to better understand how the method reacts to different choices. Because not only experiment data can be inconsistent but also models can be inaccurate, the question of model deficiencies, e.g. [21, 22, 23], has to be addressed, too. To be precise, how assumptions about model deficiencies and about inconsistencies in experiment data can be best taken into account in one procedure.

Another line of research is the generalization of the method. Besides normalization uncertainties and the energy-dependent uncertainty introduced in this paper, many other parametrizations are conceivable. Therefore, tests with other parametrization should be performed.

Finally, with the increasing complexity of uncertainty assumptions and the increasing size of databases, possibilities for further optimization of the method likely need investigation. This would primarily concern locating the posterior maxima and efficiently sampling from the posterior distribution. Regarding the latter issue, adaptive sampling algorithms, such as [18], show promise.

ACKNOWLEDGMENTS

This work was performed within the work package WP11 of the CHANDA project (605203) financed by the European Commission.

APPENDIX

Metropolis-Hastings algorithm with symmetric proposal

Suppose that we want to acquire a sample from the pdf $\phi(\boldsymbol{\kappa})$. If it is not possible to directly draw from this pdf, the MH algorithm [14] may be used. The MH algorithm constructs a chain $\boldsymbol{\kappa}_1, \boldsymbol{\kappa}_2, \cdots, \boldsymbol{\kappa}_K$ by drawing a vector $\boldsymbol{\kappa}'$ from a proposal pdf $\psi(\boldsymbol{\kappa}' \,|\, \boldsymbol{\kappa}_i)$ on the basis of the current vector $\boldsymbol{\kappa}_i$. In the case of a symmetric proposal distribution, i.e. $\psi(\boldsymbol{\kappa}' \,|\, \boldsymbol{\kappa}_i) = \psi(\boldsymbol{\kappa}_i \,|\, \boldsymbol{\kappa}')$, the proposed vector $\boldsymbol{\kappa}'$ is accepted with probability $\min(1, \phi(\boldsymbol{\kappa}')/\phi(\boldsymbol{\kappa}_i))$ as the next vector $\boldsymbol{\kappa}_{i+1}$ of the chain. Otherwise it is rejected and $\boldsymbol{\kappa}_i$ is taken as the next vector. The sample represented by the chain has the pdf $\phi(\boldsymbol{\kappa})$ as stationary distribution.

Derivative of an inverse matrix

The matrix $M(\boldsymbol{\kappa})$ is a function of $\boldsymbol{\kappa}$ and so is its inverse $\tilde{M}(\boldsymbol{\kappa})$. The relation between these two matrices in terms of their components is given by

$$\sum_j M_{ij}(\boldsymbol{\kappa}) \tilde{M}_{jk}(\boldsymbol{\kappa}) = \delta_{ij} \,, \tag{B.1}$$

with δ_{ij} being one if $i = j$ and zero otherwise. Taking the partial derivative with respect to an element κ_l of $\boldsymbol{\kappa}$ gives

$$\sum_j \left(\frac{\partial M_{ij}(\boldsymbol{\kappa})}{\partial \kappa_l} \tilde{M}_{jk}(\boldsymbol{\kappa}) + M_{ij}(\boldsymbol{\kappa}) \frac{\partial \tilde{M}_{jk}(\boldsymbol{\kappa})}{\partial \kappa_l} \right) = 0 \,. \tag{B.2}$$

This relation can be expressed in terms of matrix products,

$$M(\boldsymbol{\kappa}) \frac{\partial \tilde{M}(\boldsymbol{\kappa})}{\partial \kappa_l} = - \frac{\partial M(\boldsymbol{\kappa})}{\partial \kappa_l} \tilde{M}(\boldsymbol{\kappa}) \,. \tag{B.3}$$

Multiplying by $\tilde{M}(\boldsymbol{\kappa})$ from the left yields

$$\frac{\partial \tilde{M}(\boldsymbol{\kappa})}{\partial \kappa_l} = - \tilde{M}(\boldsymbol{\kappa}) \frac{\partial M(\boldsymbol{\kappa})}{\partial \kappa_l} \tilde{M}(\boldsymbol{\kappa}) \,. \tag{B.4}$$

Because $M(\boldsymbol{\kappa}) = S A_0 S^T + C(\boldsymbol{\kappa})$ in this paper, we finally get

$$\frac{\partial \tilde{M}(\boldsymbol{\kappa})}{\partial \kappa_l} = - \tilde{M}(\boldsymbol{\kappa}) \frac{\partial C(\boldsymbol{\kappa})}{\partial \kappa_l} \tilde{M}(\boldsymbol{\kappa}) \,. \tag{B.5}$$

Derivative of a logarithmized determinant

Using the chain rule, the derivate of $\ln \det M(\boldsymbol{\kappa})$ can be written as

$$\frac{\partial \ln \det M(\boldsymbol{\kappa})}{\partial \kappa_l} = \frac{1}{\det M(\boldsymbol{\kappa})} \frac{\partial \det M(\boldsymbol{\kappa})}{\partial \kappa_l} \,. \tag{B.6}$$

Jacobi's formula [24, p. 305] provides us with the derivative of the determinant,

$$\frac{\partial \det M(\boldsymbol{\kappa})}{\partial \kappa_l} = \mathrm{Tr} \left(\mathrm{adj}(M(\boldsymbol{\kappa})) \frac{\partial M(\boldsymbol{\kappa})}{\partial \kappa_l} \right) . \tag{B.7}$$

The adjugate matrix appearing in this expression is defined by [24, p. 192]

$$M \,\mathrm{adj}(M) = \det(M)\, \mathbb{1} \;\Rightarrow\; \mathrm{adj}(M) = \det(M) M^{-1} \tag{B.8}$$

Inserting eq. (B.8) into eq. (B.7) and the resulting expression into eq. (B.6) yields

$$\frac{\partial \ln \det M(\boldsymbol{\kappa})}{\partial \kappa_l} = \mathrm{Tr} \left((M(\boldsymbol{\kappa}))^{-1} \frac{\partial M(\boldsymbol{\kappa})}{\partial \kappa_l} \right) . \tag{B.9}$$

Because $M(\boldsymbol{\kappa}) = S A_0 S^T + C(\boldsymbol{\kappa})$ in this paper, we arrive at

$$\frac{\partial \ln \det M(\boldsymbol{\kappa})}{\partial \kappa_l} = \mathrm{Tr} \left((M(\boldsymbol{\kappa}))^{-1} \frac{\partial C(\boldsymbol{\kappa})}{\partial \kappa_l} \right) . \tag{B.10}$$

Matrix determinant lemma

For the derivation of the matrix determinant lemma in the version used in this paper, note that

$$\begin{aligned} \det\left(A + UV^T\right) &= \det\left(A\left(\mathbb{1} + A^{-1}UV^T\right)\right) \\ &= \det(A) \det\left(\mathbb{1} + A^{-1}UV^T\right) . \end{aligned} \tag{B.11}$$

The application of Sylvester's determinant identity [24, p. 416], i.e. $\det(\mathbb{1} + AB) = \det(\mathbb{1} + BA)$, yields

$$\det\left(A + UV^T\right) = \det(A) \det\left(\mathbb{1} + V^T A^{-1} U\right) . \tag{B.12}$$

Now replace U by the matrix product UW and extract W to obtain

$$\det\left(A + UWV^T\right) = \det(A) \det\left(\left(W^{-1} + V^T A^{-1} U\right) W\right) . \tag{B.13}$$

Making use of $\det(AB) = \det(A)\det(B)$, we get

$$\det\left(A + UWV^T\right) = \det(A) \det(W) \det\left(W^{-1} + V^T A^{-1} U\right) . \tag{B.14}$$

Woobury identity

The Woodbury identity [24, p. 424] states that

$$\left(A + UWV^T\right)^{-1} = A^{-1} - A^{-1}U\left(W^{-1} + V^TA^{-1}U\right)^{-1}V^TA^{-1}. \tag{B.15}$$

This identity can be verified by multiplying both sides with $(A + UWV^T)$. Applying this identity to $M(\boldsymbol{\kappa}) = SA_0S^T + C(\boldsymbol{\kappa})$ gives

$$\left(SA_0S^T + C\right)^{-1} = C^{-1} - C^{-1}S\left(A_0^{-1} + S^TC^{-1}S\right)^{-1}S^TC^{-1} \tag{B.16}$$

REFERENCES

1. R. CAPOTE, D. SMITH, and A. TRKOV, "Nuclear data evaluation methodology including estimates of covariances," *EPJ Web of Conferences*, **8**, 04001 (2010).
2. R. N. PÉREZ, J. E. AMARO, and E. R. ARRIOLA, "Coarse-grained potential analysis of neutron-proton and proton-proton scattering below the pion production threshold," *Physical Review C*, **88**, *6* (Dec. 2013).
3. R. N. PÉREZ, J. E. AMARO, and E. R. ARRIOLA, "The low-energy structure of the nucleon–nucleon interaction: statistical versus systematic uncertainties," *Journal of Physics G: Nuclear and Particle Physics*, **43**, *11*, 114001 (Nov. 2016).
4. S. VARET, N. VAYATIS, and P. DOSSANTOS-UZARRALDE, "A Statistical Approach for Experimental Cross-Section Covariance Estimation Via Shrinkage," *Nuclear Science and Engineering*, **179**, *4* (Apr. 2015).
5. S. VARET, P. DOSSANTOS-UZARRALDE, N. VAYATIS, and E. BAUGE, "A Method Using Pseudo-measurements and Shrinkage for the Estimation of Cross Section Covariances," *Nuclear Data Sheets*, **118**, 357–359 (Apr. 2014).
6. S. VARET, P. DOSSANTOS-UZARRALDE, N. VAYATIS, A. GARLAUD, and E. BAUGE, "Kriging approach for the experimental cross-section covariances estimation," *EPJ Web of Conferences*, **42**, 07003 (2013).
7. M. HERMAN, M. PIGNI, P. OBLOZINSKY, S. MUGHABGHAB, C. MATTOON, R. CAPOTE, YOUNG-SIK CHO, and A. TRKOV, "Development of covariance capabilities in EMPIRE code," Tech. Rep. BNL-81624-2008-CP, Brookhaven National Laboratory (Jun. 2008).
8. E. T. JAYNES, *Probability Theory: The Logic of Science*, Cambridge University Press, Cambridge, UK ; New York, NY (Jun. 2003).
9. C. M. BISHOP, *Pattern recognition and machine learning*, Information science and statistics, Springer, New York (2006).
10. G. SCHNABEL and H. LEEB, "A modified Generalized Least Squares method for large scale nuclear data evaluation," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, **841**, 87–96 (Jan. 2017).
11. T. M. COVER and J. A. THOMAS, *Elements of information theory*, Wiley series in telecommunications, Wiley, New York (1991).
12. R. TIBSHIRANI, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society*, **58**, *1*, 267–288 (1996).
13. R. Y. RUBINSTEIN, *Simulation and the Monte Carlo Method*, Wiley Series in Probability and Statistics, John Wiley & Sons, Inc., Hoboken, NJ, USA (Apr. 1981).
14. W. K. HASTINGS, "Monte Carlo Sampling Methods Using Markov Chains and Their Applications," *Biometrika*, **57**, 97 (1970).
15. C. G. BROYDEN, "The Convergence of a Class of Double-rank Minimization Algorithms 1. General Considerations," *IMA Journal of Applied Mathematics*, **6**, *1*, 76–90 (1970).
16. R. H. BYRD, P. LU, J. NOCEDAL, and C. ZHU, "A Limited Memory Algorithm for Bound Constrained Optimization," *SIAM Journal on Scientific Computing*, **16**, *5*, 1190–1208 (Sep. 1995).
17. H. LANDOLT, K.-H. HELLWEGE, and O. MADELUNG, editors, *Total Cross-Sections for Reactions of High Energy Particles*, vol. 12, Springer, Berlin, neue serie ed. (1988).
18. H. HAARIO, E. SAKSMAN, and J. TAMMINEN, "An adaptive Metropolis algorithm," *Bernoulli*, pp. 223–242 (2001).
19. R DEVELOPMENT CORE TEAM, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria (2008).
20. C. E. RASMUSSEN and C. K. I. WILLIAMS, *Gaussian Processes for Machine Learning*, MIT Press, Cambridge, Mass. (2006).
21. D. NEUDECKER, R. CAPOTE, and H. LEEB, "Impact of model defect and experimental uncertainties on evaluated output," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, **723**, 163–172 (Sep. 2013).
22. D. ROCHMAN, A. J. KONING, E. BAUGE, and A. J. M. PLOMPEN, "From Flatness to Steepness: Updating TALYS Covariances with Experimental Information," *Annals of Nuclear Energy*, **73**, 7–16 (Nov. 2014).
23. G. SCHNABEL and H. LEEB, "Differential Cross Sections and the Impact of Model Defects in Nuclear Data Evaluation," *EPJ Web of Conferences*, **111**, 09001 (2016).
24. D. A. HARVILLE, *Matrix Algebra From a Statistician's Perspective*, Springer New York, New York, NY (1997).