Assessment of Novel Techniques for Nuclear Data Evaluation

Petter Helgesson[1 2] , Denise Neudecker[3], Henrik Sjöstrand[1], Michael Grosskopf[4], Donald L. Smith[5], and Roberto Capote[6]

ABSTRACT

The quality of evaluated nuclear data can be impacted by, e.g., the choice of the evaluation algorithm. The objective of this work is to compare the performance of the evaluation techniques GLS, GLS-P, UMC-G, and, UMC-B, by using synthetic data. In particular, the effects of model defects are investigated. For small model defects, UMC-B and GLS-P are found to perform best, while these techniques yield the worst results for a significantly defective model; in particular, they seriously underestimate the uncertainties. If UMC-B is augmented with Gaussian processes, it performs distinctly better for a defective model but is more susceptible to an inadequate experimental covariance estimate.

Keywords

Nuclear Data Evaluation, Evaluation Techniques, Model Defects, Experimental Biases

Introduction

Nuclear data (ND) evaluations provide one recommended data set and associated uncertainties for a specific nuclear physics observable (e.g., cross-sections, spectra) based on a statistical analysis of multiple experimental data sets and theoretical model calculations. Evaluated data are needed, e.g., as input for reactor physics or neutron dosimetry simulations. The evaluation technique Generalized least squares, GLS, is frequently employed for ND evaluations [1]. It is known to fail for large uncertainties (~30% or more), discrepant input data, non-linear dependences between experimental data and model values or parameters, non-normally distributed input data, and observables covering many orders of magnitude [1-3]. The "Unified Monte Carlo" evaluation algorithms "B" [2] (UMC-B) and "G" [4] (UMC-G) were developed to address the issues of non-normally distributed model values and non-linear dependences. UMC-G was already shown to be significantly better suited for evaluations of ratio data [5]. However, UMC-B and UMC-G have not been used for real ND evaluations.

The quality of evaluated data and covariances can be impacted by the chosen evaluation algorithm. The objective of this work is to compare the performance of different evaluation techniques, i.e., GLS, a non-linear least squares algorithm using the Levenberg-Marquardt algorithm (GLS-P) [6], and two versions of Unified Monte Carlo (UMC-G and UMC-B). We have also augmented UMC-B with Gaussian processes [8-10]. We investigate the accurateness of evaluated mean values and the reliability of the associated covariances for these techniques.

It is studied how the results of each technique are affected by experimental errors, model defects, the choice of prior, and the quality of experimental covariance information. To this end, we define a function representing the truth and generate synthetic experimental data based on this function. These examples are representative of a prompt fission neutron spectrum (PFNS) and its typical experimental uncertainties, as well as of the typical model values and associated uncertainties including model defects. PFNS was chosen as it exhibits features that frequently

[1] Department of Physics and Astronomy, Uppsala University, Uppsala, 75120, Sweden;
[2] Nuclear Research and Consultancy Group NRG, Petten, The Netherlands
[3] XCP Division, Los Alamos National Laboratory, Los Alamos, NM, 87545, USA;
[4] Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, V5A 1S6, Canada;
[5] Argonne Associate of Seville, Argonne National Laboratory, Coronado, CA, 92118, USA;
[6] NAPC-Nuclear Data Section, International Atomic Energy Agency, Vienna, A-1400, Austria;

lead GLS to fail; it covers many orders of magnitude, its model values are non-normally distributed [11], its experimental data are discrepant, and have significant uncertainties [12].

Note that we distinguish between the terms error and uncertainty following the convention in GUM [13]. Errors are the estimates' deviation from the underlying truth; they are unknown (in real situations) and modeled with random variables (both for random and systematic errors). Uncertainties describe the dispersion of these random variables, here quantified by standard deviations.

Evaluation Techniques

We consider evaluation techniques which use information from a physical model with a prior distribution for its parameters and merge this information with experimental data. The techniques are either deterministic (GLS, GLS-P) or stochastic (UMC-G, UMC-B) [14]. GLS and UMC-G work in the observable domain, meaning that the observable is fitted directly on a chosen grid; the model is used only to determine a prior distribution for the observable. Here, the prior mean vector and the covariance matrix are found by sampling the model parameters. GLS-P and UMC-B work in the parameter domain, i.e., the *model parameters* are fitted. The resulting distribution of the model parameters can be used directly for uncertainty propagation, or to obtain posterior mean values and covariances for the observable.

GLS is a standard regression technique allowing for correlated data and a prior distribution. The method assumes multivariate normal distributions for both the prior and for the experimental data. In GLS-P [5], the model parameters are fitted using a linearization of the model found through the Levenberg-Marquardt algorithm. This method assumes multivariate normal distributions for the parameters and for the experimental data. No assumption is necessary for the distribution of the observables.

For UMC-G and UMC-B, we use the Metropolis algorithm [7] to yield a sample from the posterior distribution for the observable and the parameters, respectively. The evaluated mean values and covariances are obtained from the sample mean and covariance. It is necessary to choose a probability density function (PDF) for the experimental data in both cases, and for the observable for UMC-G and for the prior parameters for UMC-B. Normal distributions are used in this work. No assumption is necessary for the distribution of the observables and for the posterior of the model parameters for UMC-B.

For the Metropolis algorithm, it is necessary to choose a proposal distribution (from which proposed steps in the algorithm are sampled). In this work, it is a Gaussian with a covariance matrix proportional to the prior covariance matrix (as suggested in [15]), with the proportionality constant adjusted using numerical root finding until the acceptance rate is between 0.22 and 0.24 (because, under certain conditions, 0.23 is an optimal acceptance rate [16]). The adjustment is done twice to avoid the dependence of the starting guess. Also, the Metropolis algorithm results in samples which are auto-correlated (i.e., not independent), which impacts the uncertainty of the resulting estimates. This is addressed by estimating the autocorrelation (after the adjustment above), and only recording samples for which the autocorrelation is insignificant.

Creating Synthetic Data

We assume knowing the truth, i.e., the true functional form of the sought physics observable (Eq. (2)), to study how well the different evaluation techniques reproduce this truth. In reality, the truth is unknown to the evaluator, who works with experimental data, a physical model (Eq. (1)), and a prior distribution for the model parameters.

We study with N = 1200 different realizations of reality, including the truth, the prior knowledge, and the experimental data. The truth is sampled such that the model assumed by the evaluator is affected by a model defect with a magnitude that varies "continuously" between the realities. For each reality, the experimental errors are sampled, corresponding to N independent replications of the same experiment. The prior distribution of the model parameters is sampled such that the deviation from the truth is consistent with the prior uncertainties.

TRUTH AND MODEL

For all evaluation techniques and in all "realities", the model,

$$f_{\mathrm{M}}(x;\ \mathbf{c}) = \left(c_1\sqrt{x}\ + \frac{c_2}{\sqrt{x}}\right) \exp\left(-\frac{x}{c_3}\right), (1)$$

is used, where $\mathbf{c}$ = (c_1, c_2, c_3). The model gives shapes similar to PFNS at thermal incident neutron energies. The variable x has 11 grid values $x \in \{.01, .02, .05, .1, .2, .5, 1, 2, 5, 10,\ 15\}$, which can be interpreted as outgoing neutron energies in MeV.

The truth follows the functional form

$$f_{\mathrm{T}}(x;\ \mathbf{a}) = \left(a_1\sqrt{x}\ +\ a_2 x\right)\exp\left(-\frac{x}{a_3}\right), (2)$$

with $\mathbf{a}$ = (a_1=1, a_2, a_3=2); a_2 is sampled (as detailed below) with an expected (expectation) value of 0. Comparing Eqs. (1) and (2), there is a model defect as long as $a_2 \neq 0$. Otherwise, the model can reproduce the truth by setting $c_1 = a_1$, $c_2 = 0$ and $c_3 = a_3$. The truth is ensured to be positive on the considered evaluation grid if $a_2 > \frac{a_1}{\max(\sqrt{x})}\ \approx -0.258$. We define

$$a_2 = -\frac{a_1}{\max(\sqrt{x})} + Z, (3)$$

where Z is an exponentially distributed random variable with an expected value $\langle Z\rangle = \frac{a_1}{\max(\sqrt{x})}$.

These definitions yield $<a_2>$ = 0 and that the truth is ensured to be positive on the chosen grid. Also, a qualitative inspection shows that the model defect varies within a reasonable range.

EXPERIMENTAL DATA AND EXPERIMENTAL COVARIANCE MATRICES

For six experimental values, $x_j \in \{.02, .1, 1, 2, 5, 10\}$, the experimental data points, $f_{\mathrm{E}}\left(x_j;\ \mathbf{a}, \mathbf{d}\right)$, are sampled around the truth $f_{\mathrm{T}}\left(x_j;\ \mathbf{a}\right)$, using

$$f_{\mathrm{E}}\left(x_j;\ \mathbf{a}, \mathbf{d}\right) = f_{\mathrm{T}}\left(x_j;\ \mathbf{a}\right) + d_1 {x_j}^{-(d_2+1/4)} \exp\left(-\frac{x_j}{a_3}\right) +\ d_3 + \varepsilon(x_j), (4)$$

with the vector $\mathbf{d}$ = (d_1, d_2, d_3). The term $d_1 x^{-(d2\ +\ 1/4)}\exp(-x/a_3)$ simulates a systematic error due the an erroneous correction of the multiple scattering in a PFNS experiment, d_3 is from an inadequately corrected constant background, and $\varepsilon(x_j)$ is a random error. The elements of $\mathbf{d}$ are sampled independently from a Gaussian with expected values $<\mathbf{d}>$ = (0, 0, 0) and standard deviations $\Delta\mathbf{d}$ = (0.01, 0.5, 0.03). The $\varepsilon(x_j)$ are sampled independently for each j from Gaussians with uncertainties of 0.15, 0.07, 0.02, 0.01, 0.10, 0.70 relative to the truth in each reality.

Two qualitatively different cases of experimental covariances are studied here: good and poor covariances. In the "good" case, the experimental covariances are consistent with how the experimental data are generated, i.e., $\Delta\mathbf{d}$ = (0.01, 0.5, 0.03). In the "poor" case, the experimental data covariances are produced assuming that the systematic uncertainties are Δ**d** = (0, 0, 0.03), i.e., as if the evaluator is unaware of the uncertainties due to d_1 and d_2. For a PFNS experiment, this corresponds to incorrectly estimating the effect of multiple scattering [12].

PRIOR INFORMATION

If the evaluator chooses to include a prior distribution into the evaluation, the evaluator effectively assumes that the corresponding PDF describes the probability density of the truth. We assume that the evaluator is correct in this assumption, given a non-defective model. Therefore, the expected values of the prior parameters in the different realities are sampled around the parameters assumed for the truth without model defect, following independent Gaussians with expected values <**c**> = (a_1, <a_2>, a_3) = (1, 0, 2) and the standard deviations Δ**c** = (0.3, 0.01, 0.6). I.e., Δ**c** is used to sample **c** around <**c**> and is also the prior uncertainty in each reality. Note that with a model defect there exists no parameter set giving the truth.

Results and Discussion

STUDIED QUANTITIES

Evaluated mean values and covariance estimates of our chosen physical quantity are obtained for each reality given its associated experimental data and prior. To summarize this information, a few quantities are considered for each grid point, namely:

* Evaluation bias: Mean value of the relative difference to the truth, i.e.,

$$\bar{d} = \frac{1}{N}\sum_{i=1}^{N}\frac{f_{\mathrm{ev},i}-f_{\mathrm{true},i}}{f_{\mathrm{true},i}}, \quad (5)$$

where $f_{\mathrm{ev},i}$ and $f_{\mathrm{true},i}$ are the evaluation's central value and the truth, respectively, for the *i*th reality at each x_j. The evaluation bias should be close to zero for a good evaluation technique; it estimates the bias of the evaluation including the variability of the truth, prior and experiments. The evaluation bias should not be confused with the systematic error in a single reality.

* Observed standard deviation: Sample standard deviation for the relative difference to the truth,

$$\sigma_{obs} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(d_i - \bar{d})^2}, \quad (6)$$

where $d_i = \frac{f_{\mathrm{ev},i}-f_{\mathrm{true},i}}{f_{\mathrm{true},i}}$. This quantity includes both systematic and random errors.

* Estimated standard deviation: The root mean square (RMS) of the evaluations' estimates of the standard deviation relative to the truth. We would like each of these estimates to be close to the observed standard deviation, and therefore also their RMS.

* χ^2_{true}/N: The mean value of the χ^2 comparing the evaluation to the truth (pointwise), i.e.,

$$\frac{\chi^2_{\mathrm{true}}}{N} = \frac{1}{N}\sum_{i=1}^{N}\frac{(f_{\mathrm{ev},i}-f_{\mathrm{true},i})^2}{\sigma^2_{\mathrm{ev},i}}, \quad (7)$$

where $\sigma^2_{\mathrm{ev},i}$ is the variance estimate of the evaluation for the *i*th reality. This should be close to one for a good evaluation methodology [5]. This measure summarizes the magnitude of the total error of the individual evaluations and how good the uncertainty estimates are.

The quoted uncertainties for all these quantities (obtained along the lines of [5]) are due to the finite number (N = 1200) of simulated realities, and we refer to these uncertainties as sampling uncertainties to distinguish them from the uncertainty of an experiment or an evaluation.

Correlations between different evaluated uncertainties will impact the results of integral quantities and their uncertainties computed from the evaluated ND. Therefore, on top of studying the values of the evaluations on the grid, the quantities above are studied for two integral quantities: the sum of the evaluation at all grid points and the sum of the evaluation relative to the truth at all grid points. They are referred to as the *constant* and *relative* “benchmarks”, respectively.

GOOD COVARIANCE ESTIMATE

Both in this section and the next, GLS and UMC-G agree very well with each other, and so do GLS-P and UMC-B. Therefore, they are often lumped together and discussed in terms of “GLS/UMC-G” and “GLS-P/UMC-B”, and only results for the UMC methods are plotted.

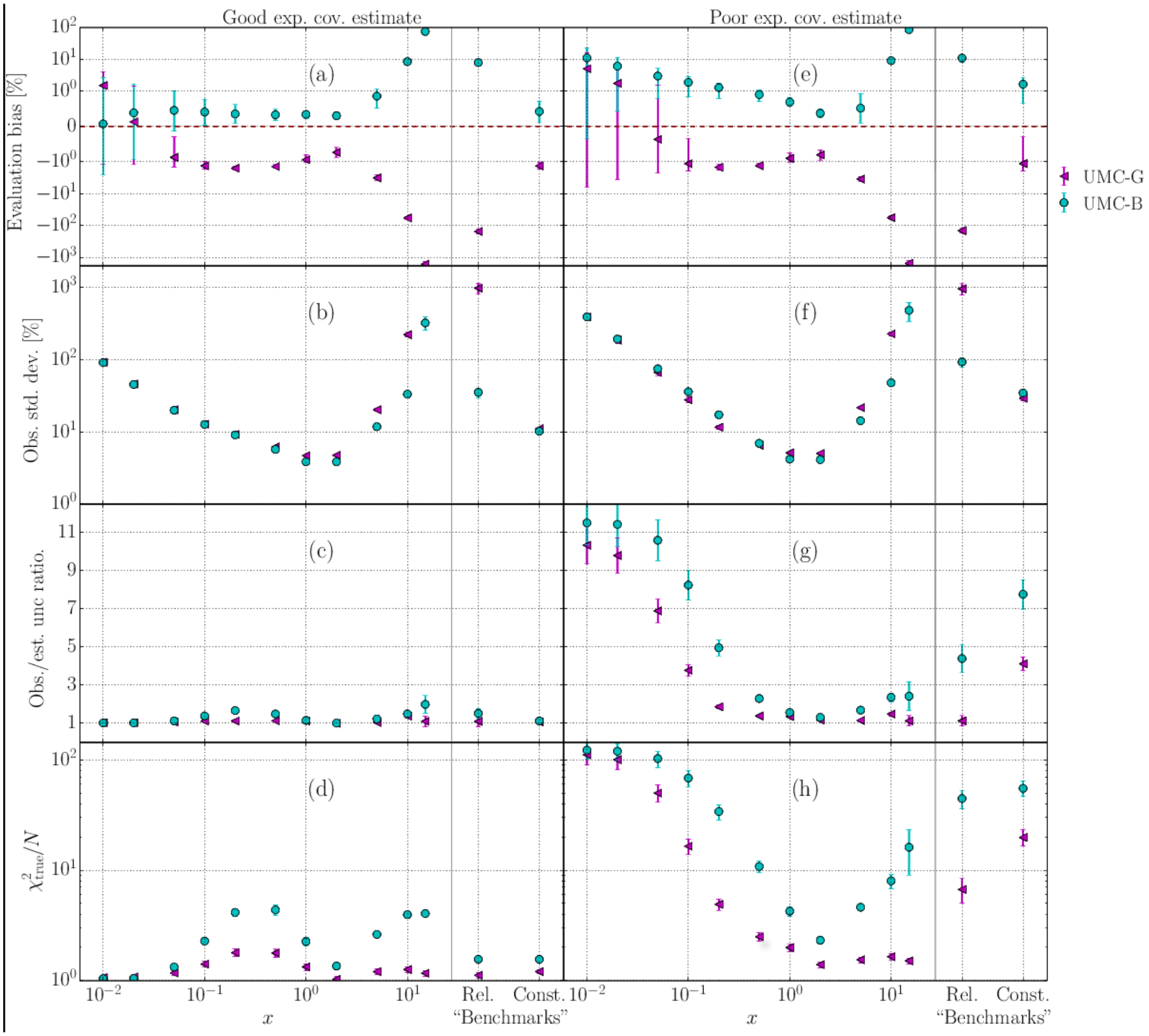

FIG. 1 Summary of the results for both for a good (a-d) and a poor (e-h) covariance estimate. The results for GLS and GLS-P are omitted from the figure, because they are insignificantly different from those of UMC-G and UMC-B, respectively. In (a) and (e), the scale is linear between -1 % and 1 %, and logarithmic otherwise. One sigma sampling uncertainty bars are shown.

Figure 1 (a-d) shows the results when a good experimental covariance matrix is provided as input to the different evaluation techniques. For GLS-P/UMC-B, the evaluation bias is barely significant except for the last two grid points and for the relative benchmark which depends strongly on the final grid points. The bias is generally larger for GLS/UMC-G, but it is on the order of one percent for most grid points, which is small compared to the uncertainties. For the last grid points ($x \geq 5$), the bias becomes very large for GLS/UMC-G, which also impacts the relative benchmark. Note that these evaluation results at the last grid point are negative in many realities, because the relative uncertainty is large and these methodologies do not guarantee positive results. Such results are unphysical and would not have been accepted by the evaluator.

Considering the evaluation bias, GLS-P/UMC-B are slightly better than GLS/UMC-G. Also, the observed standard deviation is less for GLS-P/UMC-B than for GLS/UMC-G at greater *x*. However, GLS/UMC-G are much better at *estimating* the uncertainties. This leads to χ^2_{true} values which are substantially greater for GLS-P/UMC-B than for GLS/UMC-G.

The poor performance of GLS/UMC-G for large *x* can be explained by noting that the distribution of the observable is highly skewed here, while these methods assume normal distributions.

One may suspect that the generally poor uncertainty estimates of GLS-P/UMC-B result from working in the parameter domain, making them more sensitive to model defects. Indeed, when plotting $\frac{f_{\mathrm{ev},i}-f_{\mathrm{true},i}}{\sigma_{\mathrm{ev},i}}$, i.e., the deviation to the truth normalized by the estimated standard deviation, vs. a_2 (the parameter giving the model defect), there is a clear correlation, see Fig. 2(a). This correlation indicates that $\chi^2 > 1$ can be largely explained by a_2. There is a similar effect for GLS/UMC-G, but it is weaker, as illustrated by Figure 2 (b). Thus, GLS-P/UMC-B are more sensitive to the model defect.

We have also studied what happens if $a_2 = 0$ in all realities, i.e., if there are no model defects. In this case, GLS-P/UMC-B have an even smaller evaluation bias and also succeed in estimating the uncertainties such that they agree with the observed standard deviations within the sampling uncertainty. GLS/UMC-G are slightly worse in general and for large *x* in particular.

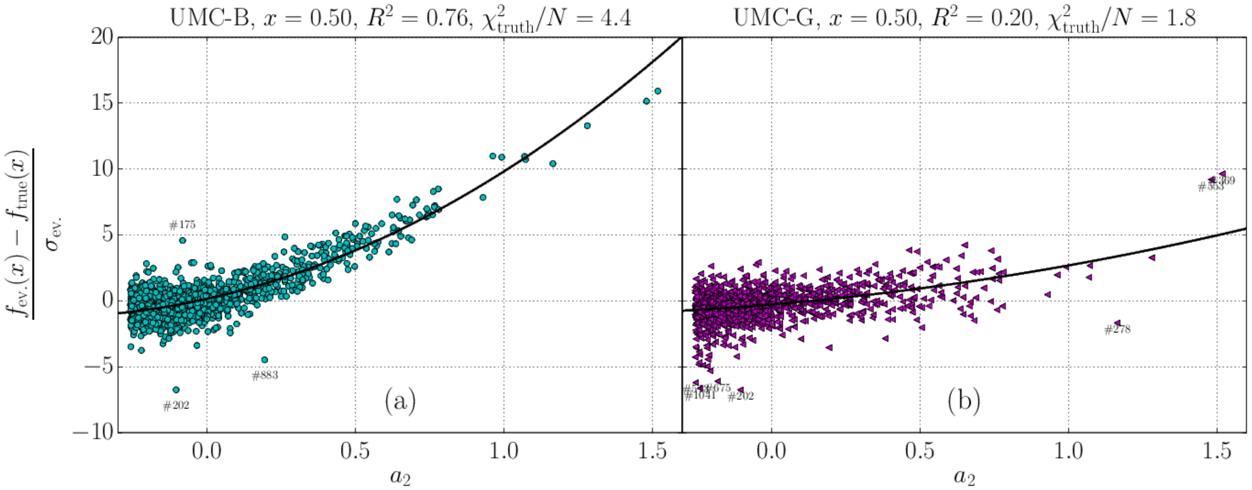

FIG. 2 The evaluations' deviation from truth at x = 0.5 normalized by the evaluated uncertainty vs. the model defect parameter a_2. Second degree polynomials are fitted to each case, and R^2 is the coefficient of determination for each fit.

POOR COVARIANCE ESTIMATE

Figure 1 (e-h) shows the results when experimental covariance estimate are estimated poorly. The poor covariance estimate is worst for low energies. Consequently, for greater x, the evaluations give quite similar results as in the case with the good covariance estimate. The observed standard deviations are somewhat greater, while the estimates are somewhat smaller. Hence, the χ^2_{true} values increase.

Even for small x, the evaluation bias is quite similar to what was obtained with a good covariance estimate. However, the observed standard deviations are substantially greater, plausibly because there is too much confidence in the experiments. Nevertheless, the *estimated* standard deviations are less than with a good covariance estimate (as expected); the uncertainty is severely underestimated by all methods, especially by GLS-P/UMC-B. Therefore, the χ^2_{true} values are all quite large. Again, they are worst for GLS-P/UMC-B. If one studies how the deviation from the truth depends on a_2 (cf., Figure 2), the GLS-P/UMC-B results with the highest χ^2_{true} can, to a large extent, be attributed to the poor ability to handle model defects.

It should be mentioned that 19 realities are left out from the summary, either because GLS-P failed to converge, or because UMC-B estimates the minimum uncertainty to be less than 1 %, indicating a poor convergence.

If a goodness-of-fit test is used to reject evaluations which disagree with the experimental data, cases may be rejected which suffer from serious model defects, poor priors, or poor experimental data covariances. For example, if we assume 6 degrees of freedom (6 points + 3 prior parameters - 3 fitted parameters) for the generalized χ^2 comparing the fit to the experiments, and reject those that have one-sided p-values of less than 5 %, the χ^2_{true} values for UMC-B for the constant "benchmark" decrease from 1.56(8) and 1200(600) to 1.33(6) and 2.18(9) for the good and poor covariance estimates, respectively.

Gaussian processes to handle model defects

Gaussian processes (GPs) are a standard approach to account for uncertainty related to potential model defects [8-10]. The GP is defined by a mean function, $\mu(x)$, and covariance function, $k(x_i, x_j)$. Here, the GP defines a prior on the model defect function, $\varepsilon_{\text{defect}}(x)$. For this prior we use $\mu(x) = 0$, to reflect that we expect the experimental data to be centered on $f_{\text{M}}(x)$, and a covariance function which is parameterized by a correlation scale, ρ, and marginal variance, σ^2, namely, $k(x_i, x_j) = \sigma^2 \mathrm{e}^{-\rho(x_i - x_j)^2}$. The covariance captures the observed discrepancy between the model and the experimental data.

Adding the GP to account for model uncertainty in UMC-B requires sampling of ρ and σ^2 together with the model parameters. We place an exponential prior with mean 1/4 for ρ and with mean 1/10 for σ^2, indicating prior expectation for a smooth discrepancy and a small magnitude of the discrepancy compared to the variation in the function. The covariance in the UMC-B likelihood is then

$$\Sigma = \Sigma_{\text{exp}} + \Sigma_{\text{defect}},$$

where Σ_{exp} is the experimental covariance and Σ_{defect} is the covariance of the GP with element *i,j* of Σ_{defect} being $k(x_i, x_j)$. For each sample point, the parameters are drawn and we compute f_{M}. Then, the posterior of $\varepsilon_{\text{defect}}(x)$ is determined using Σ as well as the discrepancy between f_{M} and the experiments. $\varepsilon_{\text{defect}}(x)$ is drawn from its posterior distribution and added to f_{M}.

Figure 3 shows a comparison of predicting the truth using UMC-B and UMC-B-GP for some selected realities. In Fig. 3a, the case of no model defects and a good experimental covariance, UMC-B and UMC-B-GP capture the truth, with UMC-B-GP inflating the magnitude of the uncertainty. In the case of a defective model, (Fig 3b.), UMC-B no longer captures the truth, while UMC-B-GP balances the experimental data and model values to better approximate the truth, with wider uncertainty intervals reflecting the uncertainty from the model defect. However, as expected, if the experimental data has systematic biases from the true curve and the experimental covariance fails to account for this, the GP discrepancy term will adjust the prediction to the (erroneous) experimental data, even if the physical model can reproduce the true curve (Fig 3 c).

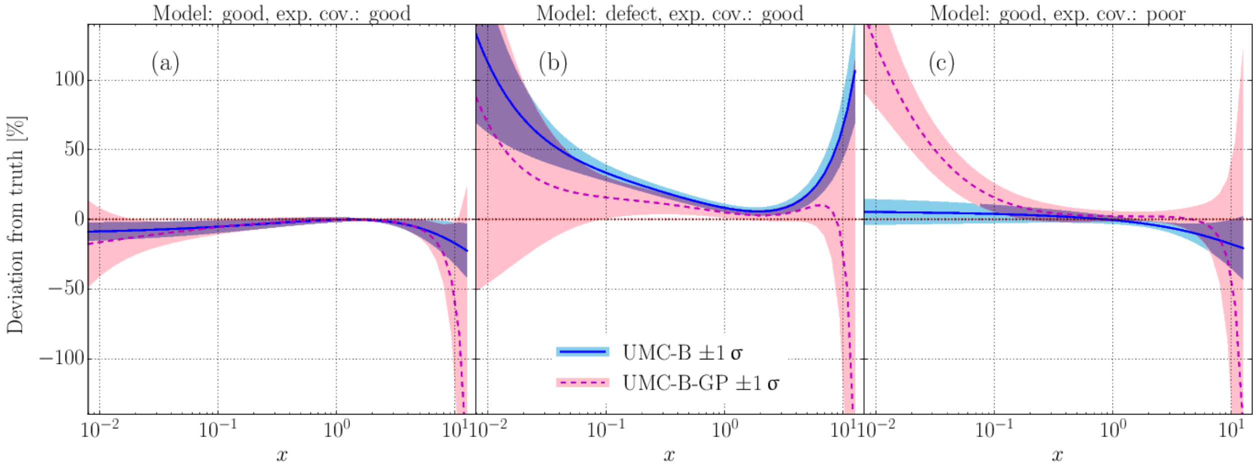

FIG 3 A comparison of UMC-B and UMC-B-GP for a selection of realities. One sigma uncertainty bars are shown.

Conclusions

This work compares how well various ND evaluation techniques (GLS, UMC-G, GLS-P, UMC-B) reproduce a predefined truth, given experimental data with random and systematic errors, and a physical model with defects as well as a prior distribution for the model parameters. For the studied situation, the results differ mainly between observable domain methods (GLS/UMC-G) and parameter domain methods (GLS-P/UMC-B). Comparing the two groups, GLS-P/UMC-B are on average somewhat closer to the truth, and have less dispersed results, especially where the normality assumption for the observable used by GLS/UMC-G is erratic. On the other hand, GLS/UMC-G yield better uncertainty *estimates*.

The problems for GLS-P/UMC-B are explained by the worse handling of model defects. If the model is non-defective, GLS-P/UMC-B performs better than GLS/UMC-G, because of their ability to make better use of the model. We have also shown an example indicating that the model defects may be addressable using Gaussian processes as previously suggested in [15].

It was seen in [4] that UMC-G outperforms GLS for ratio data (this has also been confirmed during the course of this work). UMC-B uses fewer assumptions than GLS-P which also could make a difference in other evaluation situations. However, the stochastic evaluation methods need more function calls. The approach outlined in this work can be used to decide which evaluation technique should be chosen on a case-by-case basis according to the type and quality of experimental data, the confidence in the model, and its computation time.

ACKNOWLEDGMENTS

The authors thank D. Vaughan and M.C. White for insightful discussions related to this work. We thank the Institute for Nuclear Theory at the University of Washington for its hospitality and the U.S. Department of Energy for partial support of the INT Program INT-16-2a which furthered this work by offering opportunities for collaboration (Report Number: INT-PUB-17-013). This work was partially carried out under the auspices of the NNSA of the U.S. Department of Energy at LANL under Contract No. DE-AC52-06NA25396. Funding from the Swedish Nuclear Center is gratefully acknowledged.

REFERENCES

[1] Herman, M. et al., “Covariance Data in the Fast Neutron Region,” OECD Nuclear Energy Agency Report NEA/NSC/WPEC/DOC(2010)427, 2011.

[2] Capote, R., Smith, D.L., Trkov, A. and Meghzifene, M., “A New Formulation of the Unified Monte Carlo Approach (UMC-B) and Cross-section Evaluation for the Dosimetry Reaction ^{55}Mn(n,γ)^{56}Mn,” *J. ASTM International*, Vol. 9, No. 4, JAI 104119, 2012.

[3] Neudecker, D. et al., “Evaluation of the ^{239}Pu prompt fission neutron spectrum induced by neutrons of 500 keV and associated covariances”, Nuclear Instruments and Methods in Physics Research A 791, 80-92 (2015)

[4] Smith, D.L., “A Unified Monte Carlo Approach to Fast Neutron Cross Section Data Evaluation,” *AccApp'07*, Pocatello, Idaho, July 29–August 2, 2007.

[5] Capote, R., and Smith, D.L., “An Investigation of the Performance of the Unified Monte Carlo Method of Neutron Cross Section Data Evaluation,” *Nucl. Data Sheets*, Vol. 109, 2008,pp. 2768–2773, 2008, https://doi.org/10.1016/j.nds.2008.11.007.

[6] Helgesson, P., and Sjöstrand, H., Fitting an imperfect model with or without prior, e.g., to distinguish nuclear reaction products, submitted to Review of Scientific Instruments.

[7] N. Metropolis et al., “Equation of State Calculations by Fast Computing Machines” Journal

of Chemical Physics 21, 1087-1091, 1953

[8] Kennedy, M. C., O'Hagan, A. “Bayesian calibration of computer models”. *J. Royal Stat. Soc.* B63(3), 425-464 (2001).

[9] Higdon, D., Kennedy, M., Cavendish, J. C., Cafeo, J. A.,Ryne, R. D. ”Combining field data and computer simulations for calibration and prediction”. *SIAM Journal on Scient. Comp.* 26(2), 448-466 (2004).

[10] Rasmussen, C. E., Williams, C. K. I. “Gaussian Processes for Machine Learning”. *The MIT Press*, 2006, ISBN 0-262-18253-X.

[11] Smith, D.L., Neudecker, D., and Capote Noy, R., “Testing the Goodness of Gaussian and Lognormal Emulators via Their Statistically Converged Probability Distribution Moments,” Report INDC(NDS)-0729 (IAEA, Vienna, Austria, 2017).

[12] Neudecker, D. et al., “The Need for Precise and Well-documented Experimental Data on Prompt Fission Neutron Spectra from Neutron-induced Fission of ^{239}Pu,” *Nucl. Data Sheets*, Vol. 131, 2016, pp. 289–318, https://doi.org/10.1016/j.nds.2015.12.005.

[13] Joint Committee for Guides in Metrology JCGM, “Evaluation of Measurement Data – Guide to the Expression of Uncertainty in Measurement”, JCGM 100:2008 (2008)

[14] Capote, R., Smith, D.L., and Trkov, A., “Nuclear Data Evaluation Methodology Including Estimates of Covariances”, *EPJ Web of Conferences*, Vol. 8, 04001 (2010).

[15] Schnabel, G., Ph.D. Thesis, TU Wien, Austria, 2015.

[16] Roberts, G.O. et al., “Weak Convergence and Optimal Scaling of Random Walk Metropolis Algorithms”

List of Figure Captions – (*Note:* Refer to the Author Instructions for details of formats and size.)

FIG. 1 Summary of the results for both for a good (a-c) and a poor (d-f) covariance estimate. In (a) and (d), the scale is linear between -1 % and 1 %, and logarithmic otherwise. One sigma uncertainty bars are shown.

FIG. 2 The evaluations’ deviation from truth at $x = 0.5$ normalized by the evaluated uncertainty vs. the model defect parameter a_2. Second degree polynomials are fitted to each case, and R^2 is the coefficient of determination for each fit.

FIG. 3 A comparison of UMC-B and UMC-B + GP for a selection of realities. One sigma uncertainty bars are shown.