

Treatment of Data Uncertainties

N. M. Larson

Oak Ridge National Laboratory, P.O. Box 2008, Oak Ridge, TN 37831-6171 USA

Abstract. The generation and use of data covariance matrices are discussed within the context of the analysis of neutron-induced cross section data via the R-matrix code SAMMY. Two complementary approaches are described, the first involving mathematical manipulation of Bayes' Equations and the second utilizing computer simulations. A new procedure for propagating uncertainties on unvaried parameters will allow the effect of all relevant experimental uncertainties to be reflected in the analysis results, without placing excessive additional burden on the analyst. Implementation of this procedure within SAMMY is described and illustrated via the simulations.

INTRODUCTION

For analysis of data using either least squares or Bayes' Equations (generalized least squares), it is necessary to provide appropriate and accurate information regarding the uncertainties for those data. If the data are correlated, then covariances must be provided to the analysis code and properly incorporated into the fitting procedure. At present, a number of competing methods exist both for generating such covariance matrices and for making use of them. There is no universal consensus as to which methods are correct, nor is there sufficient understanding regarding the consequences of the choices.

In this paper, a proposed method for generating and using the data covariance matrix (DCM) is described. Development of this method begins with a universally accepted premise and proceeds (via simple matrix algebra techniques) to the following conclusion: A modest change in the current generally-accepted definition of the DCM will consistently yield reasonable results, free from odd behavior sometimes seen with the old definition.

Hand calculations of simple examples and computer simulations of more realistic situations are used to demonstrate the applicability of the new method. Calculations and simulations are also used to illustrate the erroneous results that can be generated using incorrect methods.

REWRITING THE EQUATIONS

We begin with the assumption that Bayes' Equations are appropriate for determining those parameter values that give the best fit of theory to data. [This assumption can, of course, be challenged, as it relies on the dual hypotheses that all quantities obey Gaussian distributions and that the theory is linear with respect to the varied parameters. Neither hypothesis is strictly true. Nevertheless, both are sufficiently close to true that Bayes' Equations are almost correct. For the remainder of this discussion, these complications will be ignored.]

Bayes' Equations can be written in the form

$$\begin{aligned} \mathcal{P}' &= \mathcal{P} + \mathcal{M}' \mathcal{Y} & \mathcal{M}' &= (\mathcal{M}^{-1} + \mathcal{W})^{-1} \\ \mathcal{Y} &= \mathcal{G}^t \mathcal{V}^{-1} (\mathcal{D} - \mathcal{T}) & \mathcal{W} &= \mathcal{G}^t \mathcal{V}^{-1} \mathcal{G} \end{aligned} \quad (1)$$

where \mathcal{P} represents all parameters, \mathcal{M} the full covariance matrix for all parameters, \mathcal{D} the measured data, \mathcal{T} the corresponding theoretical calculation, \mathcal{G} the partial derivative of \mathcal{T} with respect to \mathcal{P} , and \mathcal{V} the DCM. The quantities \mathcal{Y} and \mathcal{W} are defined by the expressions in Eq. (1). Primes represent updated values for \mathcal{P} and \mathcal{M} . Superscript t indicates transpose.

Note that substituting zero in place of \mathcal{M}^{-1} reduces Eq. (1) to the more familiar least-squares equations.

Consider the case of fitting to raw (uncorrelated) data, for example, counts per time-channel as measured in a time-of-flight experiment. While it is seldom practical to calculate directly the quantities

measured in an actual time-of-flight experiment, nevertheless it is possible to formally express Bayes' Equations in this manner. Further, because raw data are uncorrelated, there is little ambiguity or argument regarding the treatment of the diagonal DCM.

Bayes' Equations may be written in terms of two distinct types of parameters: Define P as those parameters that are related to the theory (e.g., the R-matrix parameters) and p as those related to the measurement conditions (the normalizations, backgrounds, and other corrections required in converting from raw to reduced data, collectively denoted the "data-reduction parameters"). The prior covariance matrices M and m (for P and p respectively) are not correlated to each other.

If d denotes the raw data, v the associated diagonal DCM, and t the corresponding theoretical calculation, then the components of Eq. (1) may be written in terms of these quantities as

$$\begin{aligned} \begin{bmatrix} P' \\ p' \end{bmatrix} &= \begin{bmatrix} P \\ p \end{bmatrix} + M'Y \\ M &= \left(\begin{bmatrix} M & 0 \\ 0 & m \end{bmatrix}^{-1} + W \right)^{-1} \\ Y &= \begin{bmatrix} G^t \\ g^t \end{bmatrix} v^{-1} (d-t) \\ W &= \begin{bmatrix} G^t \\ g^t \end{bmatrix} v^{-1} [G \ g]. \end{aligned} \quad (2)$$

Here G represents the partial derivatives of t with respect to the theory parameters and g the partial derivatives of t with respect to the data-related parameters.

Experimentalists transform the raw data d into reduced data \tilde{d} by a series of operations involving the data-reduction parameters p . This transformation, which we shall call T , also takes the theory t into \tilde{t} and (applied twice) v into \tilde{v} (which is not the covariance matrix for the reduced data, but instead represents only the diagonal "statistical" portion thereof). Similarly \tilde{G} and \tilde{g} indicate partial derivatives of \tilde{t} with respect to P and p respectively. The quantity $TT^{-1} = 1$ may be inserted as needed into Eq. (2), with the goal that Bayes' Equations be expressed entirely in terms of reduced data rather than raw data. After many pages of algebra (available from

the author on request), the transformed equations reduce to the form

$$\begin{aligned} P' &= P + M'Y \quad M' = (M^{-1} + W)^{-1} \\ Y &= \tilde{G}^t \tilde{v}^{-1} (\tilde{d} - \tilde{t}) \quad W = \tilde{G}^t \tilde{v}^{-1} \tilde{G} \end{aligned} \quad (3)$$

where V (the entire, off-diagonal, covariance matrix for the reduced data) is given by

$$V = \tilde{v} + \tilde{g} m \tilde{g}^t. \quad (4)$$

Equations (3 and 4) represent only those portions of the transformed equations that apply to the theory parameters P . (Similar equations are found for the data-reduction parameters p ; equations for the covariance matrix elements connecting P and p are also found. These, however, will not be discussed further here.)

Use of the equations in Eq. (3) produces results for P' and M' (updated parameter and covariance matrix) exactly equivalent to those that would be produced if one could fit directly to the raw data. This assertion has been verified by studies of simple cases and by computer simulations, as described later in this report.

Examination of Eqs. (3 and 4) shows that those equations are identical to the equations in general use for analyzing reduced (correlated) data, with one notable exception: The definition of \tilde{g} in Eq. (4) is different. The usual definition involves the derivative of the reduced data, not of the theory, with respect to the parameters p . This is a subtle distinction, often unnoticeable with high-quality data. However, when data discrepancies exist, this small difference can lead to seemingly paradoxical results. One well-known example is Peelle's Pertinent Puzzle.

Peelle's Pertinent Puzzle

In 1987, Peelle [1] postulated a simple situation wherein the usual approach led to seemingly paradoxical results. Two "measurements" were made of the same quantity, and those measurements were correlated. A subsequent averaging of those two measurements resulted in a value that did not lie between the two measured values – a clearly unreasonable result.

This puzzle has been examined and properly interpreted by several authors (see, for example, [2,3]). Nevertheless, it is worth revisiting because only recently have the correct techniques been introduced

into practical applications; hence those techniques are not widely understood. (Implementation in the SAMMY code [4, 5] is discussed in the next section of this report; see [6] for a discussion of use of this technique in the GMA code.)

Peelle postulated two data points D_1 and D_2 with values 1.5 and 1.0. Both had statistical uncertainties of 10 % and normalization uncertainty of 0.2. Hence, the [original] DCM was

$$\begin{aligned} V &= \begin{bmatrix} 0.15^2 + 1.5^2 \cdot 0.2^2 & 0.15 \times 1.0 \times 0.2^2 \\ 0.15 \times 1.0 \times 0.2^2 & 1.0^2 + 1.0^2 \cdot 0.2^2 \end{bmatrix} \\ &= \begin{bmatrix} 0.1125 & 0.6 \\ 0.6 & 0.05 \end{bmatrix} \end{aligned} \quad (5)$$

Applying Bayes' Equations (3) with one parameter and two data points, assuming $M^{-1} = 0$ and $G = 1$, gives the solution $P' = 15/17 \approx 0.88$ and $\Delta P' \approx 0.22$, an unacceptable result. However, if one uses the appropriate version of Eq. (4),

$$v = \begin{bmatrix} 0.15^2 + P^2 \times 0.2^2 & P^2 \times 0.2^2 \\ P^2 \times 0.2^2 & 1.0^2 + P^2 \times 0.2^2 \end{bmatrix}, \quad (6)$$

rather than Eq. (5) for the DCM, then the solution becomes $P' = 15/13 \approx 1.15$ with $\Delta P' \approx 0.25$, a far more reasonable result. This is also the identical result that would be obtained if the normalization were included as a fitting parameter (which is equivalent to "fitting the raw data").

IMPLEMENTATION IN SAMMY

The form of Bayes' Equations found in Eq. (3) has been implemented in the multilevel multi-channel R-matrix code SAMMY [4, 5] and is available for use with any parameter for which SAMMY is able to calculate partial derivatives. That is, any parameter previously permitted to be varied (treated as a search parameter) may now be used in the calculation of the data covariance matrix. Parameters used in this fashion are designated "propagated uncertainty parameters" (PUPs).

The PUP option is useful when the analyst has reason to believe that the input value of the parameter is the "best" and therefore should not be modified by the analysis of the current data set; nevertheless, there is uncertainty associated with the parameter value. Designating this parameter as a PUP allows its

uncertainty to be propagated through the analysis process so that it can be reflected in the final results.

The procedure for using this option in SAMMY is to replace the flag "0" (meaning "hold this value fixed") or "1" (meaning "this parameter is to be varied") with "3" (meaning "this parameter is a PUP"). In the output file SAMMY.LPT, varied parameters are designated by ordinal numbers in (rounded) parentheses; PUP'd parameters are designated by ordinal numbers in <pointed> parentheses.

USING THE DCM

It is possible to generate V directly from Eq. (4) and then invert it for use in Eq. (3). However, that method is both costly (in terms of computer time and memory) and inefficient. Instead, the matrix V can be inverted by matrix manipulation of its components,

$$V^{-1} = v^{-1} - v^{-1} g Z^{-1} g^t v^{-1} \quad (7)$$

in which tildes have been dropped for simplicity, and Z is defined as

$$Z = m^{-1} + g^t v^{-1} g \quad (8)$$

Even V^{-1} need never be stored. Instead, Eq. (7) can be inserted directly into the final two equations of Eq. (3), giving

$$\begin{aligned} Y &= G^t v^{-1} (d - t) \\ &\quad - G^t v^{-1} g Z^{-1} g^t v^{-1} (d - t) \end{aligned} \quad (9)$$

and

$$\begin{aligned} W &= G^t v^{-1} G \\ &\quad - G^t v^{-1} g Z^{-1} g^t v^{-1} G \end{aligned} \quad (10)$$

Although the equations look more complex in this form, and indeed they are more difficult to program, the substantial savings in computer time and memory make the effort well worth while. Detailed examples illustrating these savings are available from the author.

In SAMMY, both PUPs and user-supplied DCM are treated in this fashion, which is denoted the implicit data covariance (IDC) method.

While it remains possible to provide an explicit DCM for SAMMY runs, use of explicit DCMs is strongly discouraged. In addition to requiring orders of magnitude more computer time and memory, explicit DCMs are prone to accuracy problems caused by the inability to transmit sufficient significant digits. The IDC method does not suffer from that shortcoming.

COMPUTATIONAL STUDIES

During the implementation process for PUPs in SAMMY, the author made a series of tests to be certain that both the implementation and the theory were correct. For these tests, the parameter being studied (which we will denote as X) was treated in six different methods:

1. Create the DCM V from Eq. 4, using X as a data-reduction parameter. Use SAMMY's explicit DCM option to find values and covariance matrix for resonance parameters.
2. Generate the pieces v , g , and m of the DCM of Eq. (4), again using X as a data-reduction parameter. Analyze via SAMMY's user-supplied IDC option.
3. Include X as one of the parameters to be fitted during the analysis process.
4. Treat X as a PUP.
5. If X is a normalization or background parameter, use SAMMY's original IDC option.
6. Ignore the uncertainty on X .

For each parameter tested, Methods 1 and 2 gave nearly identical results (so long as care was taken to include sufficient significant digits for the explicit DCM). Methods 3 and 4 gave exactly identical results for the first iteration. (Iterations are required to overcome the effects of non-linearity; after the first iteration, Methods 3 and 4 will necessarily differ, since the value of X will be modified in Method 3 but not in 4.) When Method 5 was possible, Methods 4 and 5 gave exactly identical results.

The first five methods give nearly identical results for the first iteration. This would not, of course, be true in the case of severely discrepant but correlated data, for which Methods 1 and 2 would not be appropriate. Method 6, in general, gave different results from all the other methods, and also produced smaller uncertainties for most parameters.

(In the previous paragraphs, the word "results" refers to values and covariance matrix for all parameters other than X .)

SUMMARY

In order to analyze reduced (and therefore correlated) data and obtain results equivalent to those obtained from fitting raw data, the data covariance matrix (DCM) should be generated using theoretical rather than experimental values.

Using the implicit data covariance (IDC) method for storing and inverting the DCM will decrease the run time and increase the accuracy of the calculation.

Both of these techniques are implemented in SAMMY for propagated-uncertainty parameters. Extensive testing has demonstrated that equivalent results are obtained using any valid method.

ACKNOWLEDGMENTS

The author wishes to thank members of the International Atomic Energy Agency Coordinated Research Project on Light Element Standards for stimulating discussions on topics addressed in this paper. Goran Arbanas's careful proofreading of the author's algebra is also gratefully acknowledged. This work was sponsored by the US Department of Energy Nuclear Criticality Safety Program and DOE/EM-22 under Contract No. DE-AC05-00OR22725 with UT-Battelle, LLC.

REFERENCES

1. Peelle, R. W., "Peelle's Pertinent Puzzle," informal memorandum dated October 13, 1987, Oak Ridge National Laboratory, Oak Ridge, TN, USA, 1987.
2. Zhao, Z. and Perey, P., *The Covariance Matrix of Derived Quantities and Their Combination*, ORNL/TM-12106, Oak Ridge National Laboratory, 1992.
3. Chiba, S. and Smith, D. L., "Some Comments on Peelle's Pertinent Puzzle", AERI-M 94-068, 1994.
4. Larson, N. M., *Updated Users' Guide for SAMMY*, ORNL/TM-9179/R6, 2003.
5. Larson, N. M., *Introduction to the Theory and Analysis of Resolved (and Unresolved) Neutron Resonances via SAMMY*, ORNL/M-6576, 1998.
6. Pronyaev, V. G. et al., "Status of the International Neutron Cross Section Standards File", ND2004.