

# A Concise Method for Storing and Communicating the Data Covariance Matrix

August 2008

Prepared by  
N. M. Larson

## DOCUMENT AVAILABILITY

Reports produced after January 1, 1996, are generally available free via the U.S. Department of Energy (DOE) Information Bridge.

**Web site** <http://www.osti.gov/bridge>

Reports produced before January 1, 1996, may be purchased by members of the public from the following source.

National Technical Information Service  
5285 Port Royal Road  
Springfield, VA 22161  
**Telephone** 703-605-6000 (1-800-553-6847)  
**TDD** 703-487-4639  
**Fax** 703-605-6900  
**E-mail** [info@ntis.gov](mailto:info@ntis.gov)  
**Web site** <http://www.ntis.gov/support/ordernowabout.htm>

Reports are available to DOE employees, DOE contractors, Energy Technology Data Exchange (ETDE) representatives, and International Nuclear Information System (INIS) representatives from the following source.

Office of Scientific and Technical Information  
P.O. Box 62  
Oak Ridge, TN 37831  
**Telephone** 865-576-8401  
**Fax** 865-576-5728  
**E-mail** [reports@osti.gov](mailto:reports@osti.gov)  
**Web site** <http://www.osti.gov/contact.html>

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

Nuclear Science and Technology Division

**A CONCISE METHOD FOR STORING AND COMMUNICATING  
THE DATA COVARIANCE MATRIX**

N. M. Larson

Date Published: October 2008

Prepared by  
OAK RIDGE NATIONAL LABORATORY  
Oak Ridge, Tennessee 37831-6283  
managed by  
UT-BATTELLE, LLC  
for the  
U.S. DEPARTMENT OF ENERGY  
under contract DE-AC05-00OR22725

# CONTENTS

	<b>Page</b>
ACKNOWLEDGMENTS .....	v
ABSTRACT .....	vii
1. INTRODUCTION .....	1
2. COMPONENTS OF THE DATA COVARIANCE MATRIX.....	1
3. IMPLICIT DATA COVARIANCE METHODOLOGY .....	2
4. EXAMPLE.....	4
5. RECOMMENDATIONS FOR DATA RESPOSITORIES .....	6
6. SUMMARY .....	6
REFERENCES .....	7
APPENDIX.....	9

## **ACKNOWLEDGMENTS**

The author is indebted to the many SAMMY users who have encouraged and supported her work on this and other topics, and helped her refine these ideas. Special thanks are given to Giles Noguere for the use of his thesis data to illustrate the efficacy of the techniques advocated in this report. Funding for this work was provided by the U. S. DOE Nuclear Criticality Safety Program.

## ABSTRACT

The covariance matrix associated with experimental data (cross section, transmission, or other) consists of several components. Statistical uncertainties on the measured quantity (counts) provide a diagonal contribution. Off-diagonal components arise from uncertainties on the parameters (such as normalization or background) that figure into the data reduction process; these are denoted systematic or “common” uncertainties, since they are common to many data points.

The full off-diagonal data covariance matrix (DCM) can be extremely large, since the size is the square of the number of data points. Fortunately, it is not necessary to explicitly calculate, store, or invert the DCM. Likewise, it is not necessary to explicitly calculate, store, or use the inverse of the DCM. Instead, it is more efficient to accomplish the same results using only the various component matrices that appear in the definition of the DCM. Those component matrices are either diagonal or small (the number of data points times the number of data-reduction parameters); hence, this “implicit data covariance” method requires far less array storage and far fewer computations while producing more accurate results.

The purpose of this report is to encourage experimentalists to report all information necessary for the creation of the DCM without actually creating the full DCM. Data repositories can then readily store the needed information and communicate it to analysts and evaluators.

## 1. INTRODUCTION

Creation, storage, and utilization of uncertainty information for large experimental data sets are daunting tasks. If the number of data points for a particular measurement is  $N$ , the size of the associated experimental data covariance matrix (DCM) is  $N^2$ . Since  $N$  can be hundreds of thousands, it is often not feasible to store such large arrays. Fortunately, there is an alternative; only the various components of the DCM are actually needed for data analyses, and the size of the arrays needed to hold the components is far smaller than  $N^2$ . Storage space and communication difficulties are therefore minimized. This topic is discussed in Section 2 of this report.

A second advantage of reporting only the components is that the implicit data covariance (IDC) methodology can then be used for fitting those data. The IDC method is faster, more economical, and often more accurate than using the explicit DCM. Details about the IDC method are given in Section 3. Most of the description is borrowed directly from the SAMMY users' manual [1], Section IV.D.3, with additional information from [2]. An example illustrating the advantages of IDC is presented in Section 4 of this report.

Recommendations concerning repository storage of the DCM are found in Section 5. A short summary is given in Section 6.

This paper does not address details of the creation of a particular DCM; for a discussion on that topic, the reader is referred to [3, 4]. Instead, this paper addresses the question of how the numbers required to generate the DCM are to be conveyed to the analysts and evaluators who need that information.

## 2. COMPONENTS OF THE DATA COVARIANCE MATRIX

The DCM can be written as a sum of two terms. The first term is a diagonal portion that contains statistical uncertainties associated with the measurement itself; the  $i^{\text{th}}$  diagonal element  $v_{ii}$  is the square of the statistical uncertainty on the  $i^{\text{th}}$  data point  $D_i$ . The second term characterizes the uncertainties related to parameters involved in the data-reduction process; these are called systematic or common errors, since they can apply systematically to many or all data points.

For example, suppose the reduced data  $D$  are related to the raw data  $R$  by the simple formula

$$D = aR + b + cE \quad , \quad (1)$$

in which  $a$ ,  $b$ , and  $c$  are data-reduction parameters ( $a$  is normalization and  $b + cE$  is background) and  $E$  is energy. The data covariance matrix may be found by first taking small increments of this expression for data point  $D_i$ ,

$$\delta D_i = a \delta R_i + \delta a R_i + \delta b + \delta c E_i \quad , \quad (2)$$

in which we have assumed that the energy is well known. Next, this expression is multiplied by the corresponding expression for data point  $D_j$  and expectation values are taken, giving

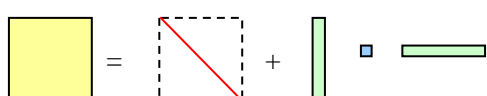
$$V_{ij} = \langle \delta D_i \delta D_j \rangle = \left\langle \left( a \delta R_i + \delta a R_i + \delta b + \delta c E_i \right) \left( a \delta R_j + \delta a R_j + \delta b + \delta c E_j \right) \right\rangle \quad , \quad (3)$$

which may be simplified to the form

$$\begin{aligned}
 V_{ij} &= a^2 \langle \delta R_i \delta R_j \rangle + R_i \langle \delta^2 a \rangle R_j + \langle \delta^2 b \rangle + \langle \delta b \delta c \rangle E_j + E_i \langle \delta c \delta b \rangle + E_i \langle \delta^2 c \rangle E_j \\
 &= a^2 \Delta^2 R_i \delta_{ij} + R_i R_j \Delta^2 a + \Delta^2 b + \langle \delta b \delta c \rangle E_j + E_i \langle \delta c \delta b \rangle + E_i E_j \Delta^2 c .
 \end{aligned}
 \tag{4}$$

Here, the first term represents the statistical uncertainty and the remaining terms represent the systematic uncertainties.

In more general terms, the formula for the DCM can be written in matrix notation as

$$V = v + g m g^t . \tag{5}$$


In this expression, the  $ik$  element of matrix  $g$  is essentially the partial derivative of the  $i^{\text{th}}$  reduced data point  $D_i$  with respect to  $k^{\text{th}}$  data-reduction parameter  $P_k$ . The square matrix  $m$  represents the covariance matrix for the data-reduction parameters; the  $k^{\text{th}}$  diagonal element  $m_{kk}$  is the square of the measured uncertainty on the  $k^{\text{th}}$  parameter  $P_k$ . (For example, if the  $k^{\text{th}}$  parameter is normalization whose nominal value is  $1 \pm 0.02$ , then  $g_{ik} = D_i$  and  $m_{kk} = 0.0004$ .)

The boxes following Eq. (5) are intended to represent the size of the matrices and are best viewed logarithmically. The dimensions of  $V$  may be quite large ( $\sim$  tens or hundreds of thousands), and the dimensions of  $m$  are generally quite small ( $\sim$  tens). Solid boxes represent full (non-diagonal) matrices; a dashed box indicates a diagonal matrix. Although  $m$  is often diagonal, it need not be (as in the example above); hence,  $m$  is shown as non-diagonal.

For situations in which the data-reduction formulae are relatively simple, as in the example above, it is sufficient for the experimentalist to report those formulae with values and uncertainties for each data-reduction parameter ( $a$ ,  $b$ , and  $c$  in the example). In the event that  $a$ ,  $b$ , and  $c$  were measured or fitted together, the covariance matrix associated with these data-reduction parameters must be given.

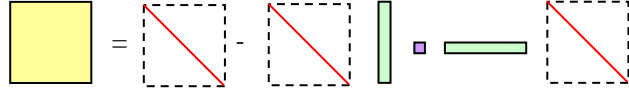
For the more general situation in which the data-reduction process involves complicated operations, the various components ( $v$ ,  $g$ , and  $m$ ) of Eq. (5) should be reported. It is not necessary for the full DCM  $V$  to be generated and reported; indeed, it would be counterproductive to do so, as important information can be lost (as will be seen in the example of Section 4 below).

### 3. IMPLICIT DATA COVARIANCE METHODOLOGY

The major advantage of storing and communicating the DCM in terms of its components is obvious; far fewer numbers are needed. The advantage to the user of the DCM requires a bit more explanation: having the DCM in terms of its components allows the analyst to use the IDC methodology rather than using the DCM directly. In this section, the IDC methodology is described. In the next section, the advantages are illustrated with an example using real data.

Because  $V$  has the form shown above, the inverse may be calculated symbolically as

$$\begin{aligned}
 V^{-1} &= (v + g m g^t)^{-1} \\
 &= v^{-1} - v^{-1} g (m^{-1} + g^t v^{-1} g)^{-1} g^t v^{-1} \\
 &= v^{-1} - v^{-1} g Z^{-1} g^t v^{-1} ,
 \end{aligned} \tag{6}$$



in which  $Z$  is given by

$$\begin{aligned}
 Z &= m^{-1} + g^t v^{-1} g . \\
 \square &= \square + \text{---} \begin{array}{|c|} \hline \diagdown \\ \hline \end{array} \begin{array}{|c|} \hline \\ \hline \end{array}
 \end{aligned} \tag{7}$$

A derivation of this form for the inverse is given in the appendix, along with verification by direct calculation of  $V V^{-1} = I$ .

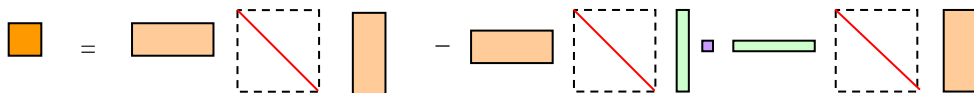
Equations (6) and (7) provide an easy way to generate the inverse of the DCM. However, further simplification is possible. It is not  $V^{-1}$  that is needed for use in a fitting procedure but rather the product of  $V^{-1}$  with other quantities. If values for a set of resonance parameters (or other parameters) are to be obtained by fitting, Bayes' equations (generalized least-squares equations) may be invoked:

$$\begin{aligned}
 M' &= (M^{-1} + W)^{-1} & P' &= P + M' Y \\
 W &= G^t V^{-1} G & Y &= G^t V^{-1} (D - T) .
 \end{aligned} \tag{8}$$

Here array  $D$  represents the experimental data and  $T$  the theoretical values,  $V$  is again the DCM,  $P$  represents the initial parameter values and  $M$  the parameter covariance matrix, and  $G$  is the partial derivatives of  $T$  with respect to  $P$ . Primes indicate updated values found by the fitting procedure. Matrices  $W$  and  $Y$  are defined by the second line of Eq. (8). Note that the only occurrences of  $V$  are in the expressions for  $W$  and  $Y$ .

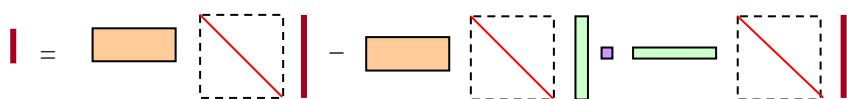
Using the expression for  $V^{-1}$  from Eq. (6),  $W$  takes the form

$$\begin{aligned}
 W &= G^t V^{-1} G \\
 &= G^t v^{-1} G - G^t v^{-1} g Z^{-1} g^t v^{-1} G .
 \end{aligned} \tag{9}$$



In this expression, a third dimension has been added, this being the number of varied parameters (~ hundreds or thousands). Typically, this is intermediate between the number of data points and the number of data-reduction parameters.

The expression for  $Y$  is

$$\begin{aligned}
 Y &= G^t V^{-1} (D-T) \\
 &= G^t v^{-1} (D-T) - G^t v^{-1} g Z^{-1} g^t v^{-1} (D-T) .
 \end{aligned} \tag{10}$$


Because  $D - T$  and  $Y$  are simple vectors, the new dimension introduced in this expression is unity.

On first inspection, the expressions in the second lines of Eqs. (9) and (10) appear to be more complicated than the original expressions in the first lines. Indeed, they are somewhat more complicated to program. However, there are significant advantages to using the second expressions: The only large matrix,  $v^{-1}$ , is diagonal and therefore trivial to compute. The other two matrices that must be inverted,  $m$  and  $Z$ , are both very small. Thus computation time is reduced because no large dense matrix is ever inverted. The required computer memory is reduced because no large matrix is ever stored. Finally, numerical accuracy and stability are improved because there are fewer opportunities to encounter round-off problems.

#### 4. EXAMPLE

Computational verification of the claims made in the preceding paragraph is easily obtained by comparing the two methods in a data-analysis example using real data. The example shown here is based on  $^{129}\text{I}$  transmission data from the thesis of Noguere [5], and has been presented at several conferences [2, 6, 7] and at the SAMMY workshops [8]; it is also test case tr140 for the SAMMY code [1]. In this example, only 1245 experimental data points are used; the full data set had 32660 data points. Uncertainties associated with 9 data-reduction parameters were used to determine the DCM. There were a total of 655 resonances in the full energy range, though only 4 are within the range used in this example. Resonance energy, capture width, and neutron width were varied for each of the three large resonances shown in Fig. 1, for a total of 9 parameters.

These data were analyzed using four different treatments of the DCM, as shown in Table 1. The first two treatments included only data uncertainties with no off-diagonal correlations. In treatment (a), only statistical uncertainties were used; in treatment (b), systematic uncertainties were included but only on the diagonal. Neither of those treatments is correct (as discussed in [2, 8] and elsewhere). Treatment (b) has previously often been used for data analysis, due to the perceived difficulty of implementing correct treatments. Treatment (c) uses the explicit DCM; treatment (d) uses the IDC method advocated here. Treatments (c) and (d) give essentially the same results, provided sufficient significant digits are communicated for (c).

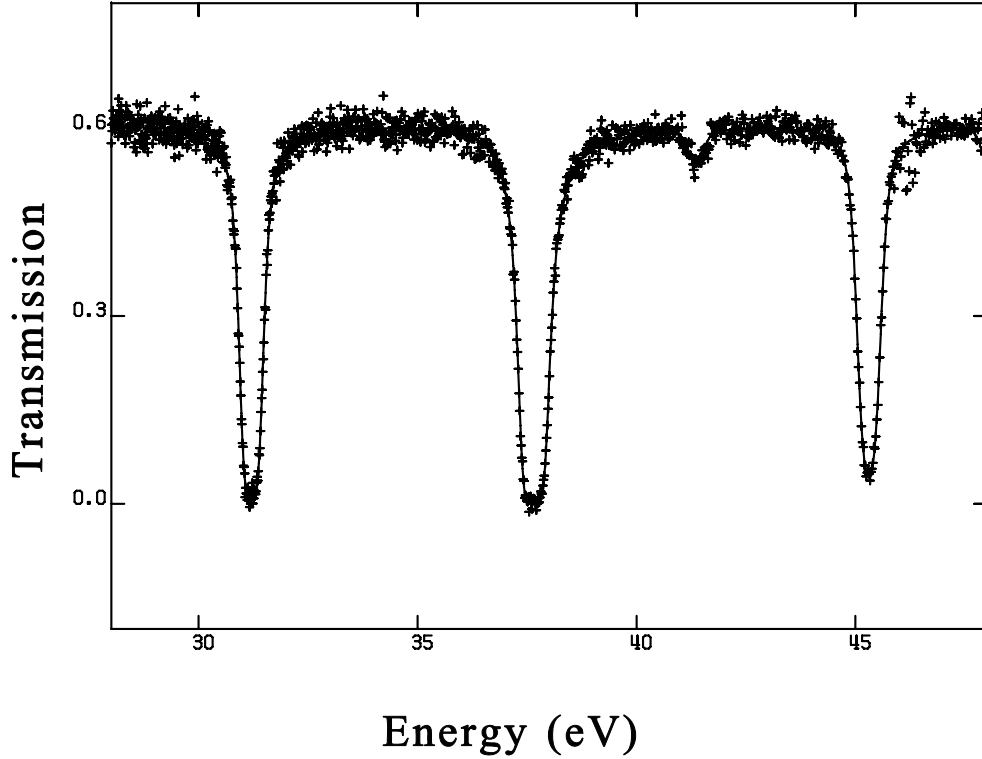


Fig. 1.  $^{129}\text{I}$  transmission data from the thesis of Noguere. Crosses are the measured data; solid curve is the SAMMY fit.

Table 1. Comparison of computer resources required for various treatments of the data covariance matrix

	Description of data covariance treatment for this run	Cpu time for least-squares fitting		Array size (K)
		routines (sec)	Total cpu time (sec)	
a	only statistical errors	0.03	14	254
b	statistical plus systematic, only on diagonal	0.03	14	254
c	explicit data covariance matrix	16.46	59	1800
d	IDC matrix	0.06	14	267

From the table, it is seen that both array size and computation time for solving Bayes' equations are significantly increased when the explicit DCM is used. Compared to the standard (but incorrect) treatment (b), cpu time increased by a factor of  $\sim 500$  and array size by a factor of 6. Using the IDC method, however, cpu time for the Bayes' solver increased by a factor of 2, a negligible amount when cpu time for the entire SAMMY run is considered. Array size increased only slightly.

These differences would be even more pronounced if the size of the data set were increased from the small number (1245) of data points used in this example to the full 32660 data points used in the Noguere analysis.

Considering storage alone, with the full data set, the DCM would require  $(32660)^2 = 1,066,675,600$  values, or a mere  $(32660 \times 32661) / 2 = 533,354,130$  values if symmetry is taken into account. Using the IDC method, only  $32660 + 9 \times 32660 + (9 \times 8) / 2 = 326,636$  values are required.

Use of the IDC also improves accuracy of the results. When these computations were first performed, very large differences were found between treatments (c) and (d). Eventually, these differences were discovered to be due to the small number of significant digits used to communicate the original DCM file. When the file was recreated using double precision and reporting all significant digits, the two treatments gave nearly identical results. The fact that results are not exactly identical is probable evidence of round-off error from the many computations required for the explicit treatment.

## 5. RECOMMENDATIONS FOR DATA RESPOSITORIES

It is clear that storage and communication of the components of the DCM are easier and more economical than storage and communication of the full DCM. Additionally, in order to exploit the many advantages of the IDC methodology, analysis codes require the components rather than the explicit DCM. Repositories of experimental data sets should therefore contain only the components  $v$ ,  $g$ , and  $m$  [from the definition of  $V$  in Eq. (5)].

Further, the more accurate representation of the DCM is in terms of the theoretical rather than experimental cross sections [1, 2]. When practical, it is therefore advisable that formulae defining the relationship of reduced to raw data be provided, so that analysis codes may use the more accurate representation of the DCM. Values and uncertainties must be specified for each data-reduction parameter. In the event that any of the parameters were measured or fitted together, the covariance matrix associated with the data-reduction parameters must also be given.

## 6. SUMMARY

The DCM should be stored in data repositories in terms of its components, with formulae describing the data-reduction process included wherever feasible. In addition to minimizing storage requirements and communication difficulties, this option maximizes the information available for analysts who wish to make use of the uncertainty information.

## REFERENCES

1. N. M. Larson, *Updated Users' Guide for SAMMY: Multilevel R-Matrix Fits to Neutron Data Using Bayes' Equations*, ORNL/TM-9179/R8, Oak Ridge National Laboratory, Oak Ridge, TN, USA (2008). Also ENDF-364/R2. Available at [http://www.ornl.gov/sci/nuclear\\_science\\_technology/nuclear\\_data/sammy/](http://www.ornl.gov/sci/nuclear_science_technology/nuclear_data/sammy/).
2. N. M. Larson, "Use of Covariance Matrices in SAMMY," *Workshop on Nuclear Data Evaluation for Reactor Application (WONDER 2006)* held at the Chateau de Cadarache, October 9–11, 2006.
3. D. C. Larson, N. M. Larson, J. A. Harvey, N. W. Hill, and C. H. Johnson, *Application of New Techniques to ORELA Neutron Transmission Measurements and their Uncertainty Analysis: the Case of Natural Nickel from 2 keV to 20 MeV*, ORNL/TM-8203, Oak Ridge National Laboratory, Oak Ridge, TN, USA (1983). Also ENDF-333. Available at [http://www.ornl.gov/sci/nuclear\\_science\\_technology/nuclear\\_data/sammy/](http://www.ornl.gov/sci/nuclear_science_technology/nuclear_data/sammy/).
4. N. M. Larson, *User's Guide to ALEX: Uncertainty Propagation from Raw Data to Final Results for ORELA Transmissions Measurements*, ORNL/TM-8676, Oak Ridge National Laboratory, Oak Ridge, TN, USA (1984). Also ENDF-332. Available at [http://www.ornl.gov/sci/nuclear\\_science\\_technology/nuclear\\_data/sammy/](http://www.ornl.gov/sci/nuclear_science_technology/nuclear_data/sammy/).
5. Gilles Noguere, "Measurements and Analysis of the  $^{127}\text{I}$  and  $^{129}\text{I}$  Neutron Capture and Total Cross Sections," PhD thesis, Report CEA-R-6071, ISSN 0429-3460, CEA Cadarache, France, 2005.
6. N. M. Larson, "Treatment of Data Uncertainties," *ND2004 (International Conference on Nuclear Data for Science and Technology)*, Sept. 26–Oct. 1, 2004, edited by R. C. Haight et al., AIP Conf. Proc. 769, 2005, pages 374–377.
7. N. M. Larson, "On the Efficient Treatment of Data Covariance Matrices," *71st Annual Meeting of the Southeastern Section of the APS*, Oak Ridge, TN, November 11–13, 2004.
8. Workshop notes are available at the web site [http://www.ornl.gov/sci/nuclear\\_science\\_technology/nuclear\\_data/sammy/](http://www.ornl.gov/sci/nuclear_science_technology/nuclear_data/sammy/).

## APPENDIX

In Eq. (5), the DCM was given in the form

$$V = v + g m g^t . \quad (11)$$

Simple matrix manipulation techniques and algebra are used to obtain the inverse of this expression. First, add and subtract the same quantity, giving

$$V^{-1} = (g m g^t + v)^{-1} = v^{-1} - v^{-1} + (g m g^t + v)^{-1} . \quad (12)$$

Next, insert the identity  $I = (g m g^t + v)(g m g^t + v)^{-1}$  on the right-hand side of the second term and  $I = v v^{-1}$  on the left-hand side of the third term.

$$V^{-1} = v^{-1} - v^{-1} (g m g^t + v) (g m g^t + v)^{-1} + v^{-1} v (g m g^t + v)^{-1} . \quad (13)$$

Rearranging terms then gives

$$\begin{aligned} V^{-1} &= v^{-1} - v^{-1} (g m g^t + v - v) (g m g^t + v)^{-1} \\ &= v^{-1} - v^{-1} g m g^t (g m g^t + v)^{-1} . \end{aligned} \quad (14)$$

Next, introduce the identity  $I = (g^t v^{-1} g + m^{-1})^{-1} (g^t v^{-1} g + m^{-1})$  into the middle of the second term, giving

$$\begin{aligned} V^{-1} &= v^{-1} - v^{-1} g (g^t v^{-1} g + m^{-1})^{-1} (g^t v^{-1} g + m^{-1}) m g^t (g m g^t + v)^{-1} \\ &= v^{-1} - v^{-1} g (g^t v^{-1} g + m^{-1})^{-1} [(g^t v^{-1} g + m^{-1}) m g^t] (g m g^t + v)^{-1} . \end{aligned} \quad (15)$$

The expression in square brackets in Eq. (15) can be written as

$$(g^t v^{-1} g + m^{-1}) m g^t = g^t v^{-1} g m g^t + m^{-1} m g^t = g^t v^{-1} g m g^t + g^t v v^{-1} , \quad (16)$$

in which we have used the identities  $m^{-1} m = I$  and  $v v^{-1} = I$ . Inserting Eq. (16) into the appropriate location in Eq. (15) gives

$$V^{-1} = v^{-1} - v^{-1} g (g^t v^{-1} g + m^{-1})^{-1} (g^t v^{-1} g m g^t + g^t v^{-1} v) (g m g^t + v)^{-1} . \quad (17)$$

Rearranging once more yields

$$V^{-1} = v^{-1} - v^{-1} g (g^t v^{-1} g + m^{-1})^{-1} g^t v^{-1} (g m g^t + v) (g m g^t + v)^{-1} . \quad (18)$$

Because  $(g m g^t + v)(g m g^t + v)^{-1} = I$ , this expression becomes

$$V^{-1} = v^{-1} - v^{-1} g (g^t v^{-1} g + m^{-1})^{-1} g^t v^{-1} , \quad (19)$$

as given in Eq. (6).

Finally, this expression for the inverse can be verified by direct calculation, giving

$$\begin{aligned}
VV^{-1} &= (v + g m g^t)(v^{-1} - v^{-1} g Z^{-1} g^t v^{-1}) \\
&= v v^{-1} + g m g^t v^{-1} - v v^{-1} g Z^{-1} g^t v^{-1} - g m g^t v^{-1} g Z^{-1} g^t v^{-1} \\
&= I + g m (I - m^{-1} Z^{-1} - g^t v^{-1} g Z^{-1}) g^t v^{-1} \\
&= I + g m (I - (m^{-1} - g^t v^{-1} g) Z^{-1}) g^t v^{-1} \\
&= I + g m (I - Z Z^{-1}) g^t v^{-1} = I + g m (I - I) g^t v^{-1} = I .
\end{aligned} \tag{20}$$