

CHAPTER 8: VARIATIONS OF THE GENERALIZED INVERSE

8.1 Linear Transformations

8.1.1 Analysis of the Generalized Inverse Operator \mathbf{G}_g^{-1}

Recall Equation (2.30)

$$\mathbf{ABC} = \mathbf{D} \quad (2.30)$$

which states that if the matrix \mathbf{D} is given by the product of matrices \mathbf{A} , \mathbf{B} , and \mathbf{C} , then each column of \mathbf{D} is the weighted sum of the columns of \mathbf{A} and each row of \mathbf{D} is the weighted sum of the rows of \mathbf{C} . Applying this to $\mathbf{Gm} = \mathbf{d}$, we saw in Equation (2.15) that the data vector \mathbf{d} is the weighted sum of the columns of \mathbf{G} . Note that both the data vector and the columns of \mathbf{G} are $N \times 1$ vectors in data space.

We can extend this analysis by using singular-value decomposition. Specifically, writing out \mathbf{G} as

$$\mathbf{G} = \mathbf{U}_P \quad \Lambda_P \quad \mathbf{V}_P^T \quad (6.69)$$

$$N \times M \quad N \times P \quad P \times P \quad P \times M$$

Each column of \mathbf{G} is now seen as a weighted sum of the columns of \mathbf{U}_P . Each column of \mathbf{G} is an $N \times 1$ dimensional vector (i.e., in data space), and is the weighted sum of the P eigenvectors $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_P$ in \mathbf{U}_P . Each row of \mathbf{G} is a weighted sum of the rows of \mathbf{V}_P^T , or equivalently, the columns of \mathbf{V}_P . Each row of \mathbf{G} is a $1 \times M$ row vector in model space. It is the weighted sum of the P eigenvectors $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_P$ in \mathbf{V}_P .

A similar analysis may be considered for the generalized inverse operator, where

$$\mathbf{G}_g^{-1} = \mathbf{V}_P \quad \Lambda_P^{-1} \quad \mathbf{U}_P^T \quad (7.8)$$

$$M \times N \quad M \times P \quad P \times P \quad P \times N$$

Each column of \mathbf{G}_g^{-1} is a weighted sum of the columns of \mathbf{V}_P . Each row of \mathbf{G}_g^{-1} is the weighted sum of the rows of \mathbf{U}_P^T , or equivalently, the columns of \mathbf{U}_P .

Let us now consider what happens in the system of equations $\mathbf{Gm} = \mathbf{d}$ when we take one of the eigenvectors in \mathbf{V}_P as \mathbf{m} . Let $\mathbf{m} = \mathbf{v}_i$, the i th eigenvector in \mathbf{V}_P . Then

$$\mathbf{G}\mathbf{v}_i = \mathbf{U}_P \Lambda_P \mathbf{V}_P^T \mathbf{v}_i \quad (8.1)$$

$N \times 1 \quad N \times P \quad P \times P \quad P \times M \quad M \times 1$

We can expand this as

$$\mathbf{G}\mathbf{v}_i = \mathbf{U}_P \Lambda_P \begin{bmatrix} \dots & \mathbf{v}_1^T & \dots \\ & \vdots & \\ \dots & \mathbf{v}_i^T & \dots \\ & \vdots & \\ \dots & \mathbf{v}_P^T & \dots \end{bmatrix} \mathbf{v}_i \quad (8.2)$$

The product of \mathbf{V}_P^T with \mathbf{v}_i is a $P \times 1$ vector with zeros everywhere except for the i th row, which represents the dot product of \mathbf{v}_i with itself.

Continuing, we have

$$\begin{aligned} \mathbf{G}\mathbf{v}_i &= \mathbf{U}_P \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & & \lambda_i & 0 \\ 0 & \dots & 0 & \lambda_P \end{bmatrix} \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \\ &= \begin{bmatrix} u_{11} & \dots & u_{1i} & \dots & u_{1P} \\ u_{21} & & u_{2i} & & u_{2P} \\ \vdots & & \vdots & & \vdots \\ u_{N1} & \dots & u_{Ni} & \dots & u_{NP} \end{bmatrix} \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \lambda_i \\ 0 \\ \vdots \\ 0 \end{bmatrix} \\ &= \lambda_i \begin{bmatrix} u_{1i} \\ u_{2i} \\ \vdots \\ u_{Ni} \end{bmatrix} \quad (8.3) \end{aligned}$$

Or simply,

$$\mathbf{G}\mathbf{v}_i = \lambda_i \mathbf{u}_i \quad (8.4)$$

This is, of course, simply the statement of the shifted eigenvalue problem from Equation (6.16). The point was not, however, to reinvent the shifted eigenvalue problem, but to emphasize the linear algebra, or mapping, between vectors in model and data space.

Note that \mathbf{v}_i , a unit-length vector in model space, is transformed into a vector of length λ_i (since \mathbf{u}_i is also of unit length) in data space. If λ_i is large, then a unit-length change in model space in the \mathbf{v}_i direction will have a large effect on the data. Conversely, if λ_i is small, then a unit length change in model space in the \mathbf{v}_i direction will have little effect on the data.

8.1.2 \mathbf{G}_g^{-1} Operating on a Data Vector \mathbf{d}

Now consider a similar analysis for the generalized inverse operator \mathbf{G}_g^{-1} , which operates on a data vector \mathbf{d} . Suppose that \mathbf{d} is given by one of the eigenvectors in \mathbf{U}_P , say \mathbf{u}_i . Then

$$\mathbf{G}_g^{-1} \mathbf{u}_i = \mathbf{V}_P \Lambda_P^{-1} \mathbf{U}_P^T \mathbf{u}_i \quad (8.5)$$

$$\begin{matrix} M \times 1 & M \times P & P \times P & P \times N & N \times 1 \end{matrix}$$

Following the development above, note that the product of \mathbf{U}_P^T with \mathbf{u}_i is a $P \times 1$ vector with zeros everywhere except the i th row, which represents the dot product of \mathbf{u}_i with itself. Then

$$\mathbf{G}_g^{-1} \mathbf{u}_i = \mathbf{V}_P \begin{bmatrix} \lambda_1^{-1} & 0 & \dots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & & \lambda_i^{-1} & 0 \\ 0 & \dots & 0 & \lambda_P^{-1} \end{bmatrix} \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

Continuing,

$$= \begin{bmatrix} v_{11} & \dots & v_{1i} & \dots & v_{1P} \\ v_{21} & & v_{2i} & & v_{2P} \\ \vdots & & \vdots & & \vdots \\ v_{M1} & \dots & v_{Mi} & \dots & v_{MP} \end{bmatrix} \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \lambda_i^{-1} \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

$$= \lambda_i^{-1} \begin{bmatrix} v_{1i} \\ v_{2i} \\ \vdots \\ v_{Mi} \end{bmatrix} \quad (8.6)$$

Or simply,

$$\mathbf{G}_g^{-1} \mathbf{u}_i = \lambda_i^{-1} \mathbf{v}_i \quad (8.7)$$

This is not a statement of the shifted eigenvalue problem, but has an important implication for the mapping between data and model spaces. Specifically, it implies that a unit-length vector (\mathbf{u}_i) in data space is transformed into a vector of length $1/\lambda_i$ in model space. If λ_i is large, then small changes in \mathbf{d} in the direction of \mathbf{u}_i will have little effect in model space. This is good if, as usual, these small changes in the data vector are associated with noise. If λ_i is small, however, then small changes in \mathbf{d} in the \mathbf{u}_i direction will have a large effect on the model parameter estimates. This reflects a basic instability in inverse problems whenever there are small, nonzero singular values. Noise in the data, in directions parallel to eigenvectors associated with small singular values, will be amplified into very unstable model parameter estimates.

Note also that there is an intrinsic relationship, or coupling, between the eigenvectors \mathbf{v}_i in model space and \mathbf{u}_i in data space. When \mathbf{G} operates on \mathbf{v}_i , it returns \mathbf{u}_i , scaled by the singular value λ_i . Conversely, when \mathbf{G}_g^{-1} operates on \mathbf{u}_i it returns \mathbf{v}_i , scaled by λ_i^{-1} . This represents a very strong coupling between \mathbf{v}_i and \mathbf{u}_i directions, even though the former are in model space and the latter are in data space. Finally, the linkage between these vectors depends very strongly on the size of the nonzero singular value λ_i .

8.1.3 Mapping Between Data and Model Space: An Example

One useful way to graphically represent the mapping back and forth between model and data spaces is with the use of “stick figures.” These are formed by plotting the components of the eigenvectors in model and data space for each model parameter and observation as a “stick,” or line, whose length is given by the size of the component. These can be very helpful in illustrating directions in model space associated with stability and instability, as well as directions in data space where noise will have a large effect on the estimated solution.

For example, recall the previous example, given by

$$\begin{bmatrix} 1.00 & 1.00 \\ 2.00 & 2.01 \end{bmatrix} \begin{bmatrix} m_1 \\ m_2 \end{bmatrix} = \begin{bmatrix} 2.00 \\ 4.10 \end{bmatrix} \quad (7.131)$$

The singular values and associated eigenvectors are given by

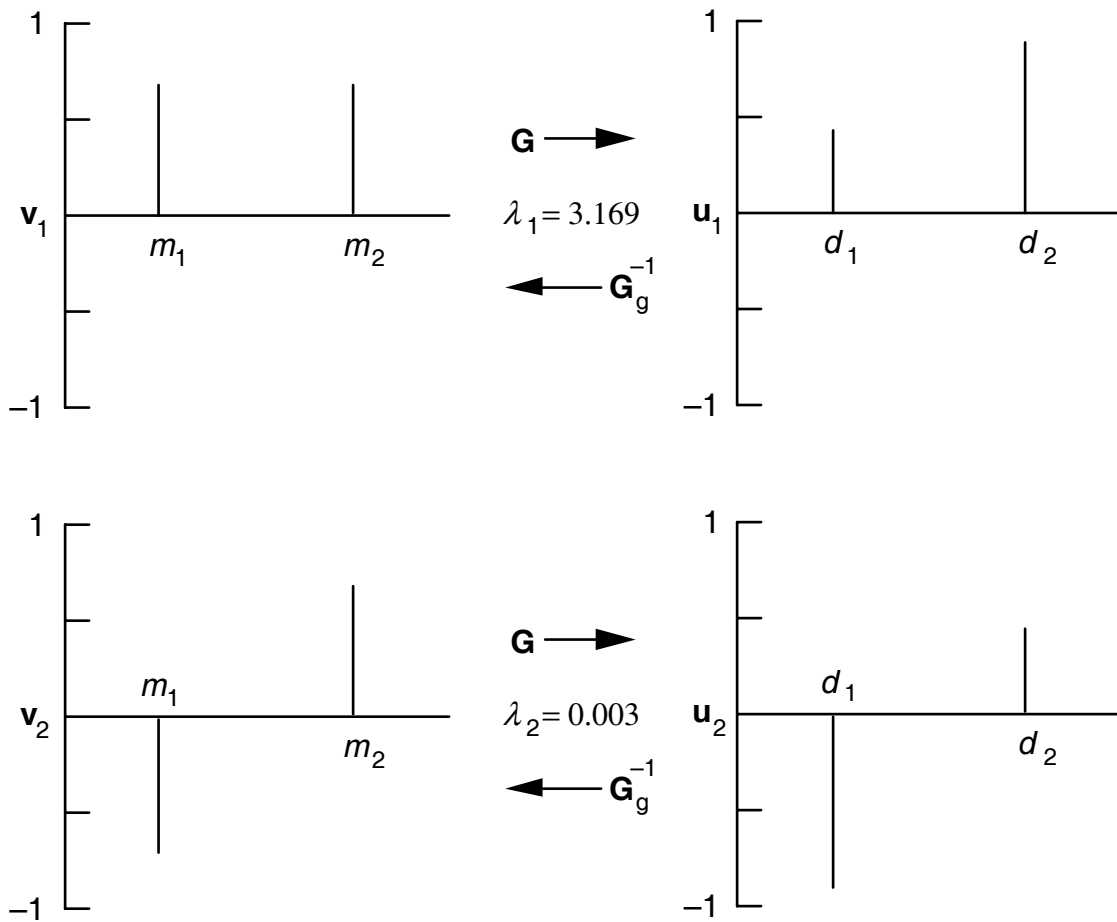
$$\lambda_1 = 3.169 \text{ and } \lambda_2 = 0.00316 \quad (8.8)$$

$$\mathbf{V}_P = \mathbf{V} = \begin{bmatrix} 0.706 & -0.710 \\ 0.709 & 0.704 \end{bmatrix} \quad (8.9)$$

and

$$\mathbf{U}_P = \mathbf{U} = \begin{bmatrix} 0.446 & -0.895 \\ 0.895 & 0.446 \end{bmatrix} \quad (8.10)$$

From this information, we may plot the following figure:



From \mathbf{V}_P , we see that $\mathbf{v}_1 = [0.706, 0.709]^T$. Thus, on the figure for \mathbf{v}_1 , the component along m_1 is +0.706, while the component along m_2 is +0.709. Similarly, $\mathbf{u}_1 = [0.446, 0.895]^T$, and thus the components along d_1 and d_2 are +0.446 and +0.895, respectively. For $\mathbf{v}_2 = [-0.710, 0.704]^T$, the components along m_1 and m_2 are -0.710 and +0.704, respectively. Finally, the components of $\mathbf{u}_2 = [-0.895, 0.446]^T$ along d_1 and d_2 are -0.895 and 0.446, respectively.

These figures illustrate, in a simple way, the mapping back and forth between model and data space. For example, the top figure shows that a unit length change in model space in the $[0.706, 0.709]^T$ direction will be mapped by \mathbf{G} into a change of length 3.169 in data space in the $[0.446, 0.895]^T$ direction. A unit length change in data space along the $[0.446, 0.895]^T$ direction will be mapped by \mathbf{G}^{-1} into a change of length $1/(3.169) = 0.316$ in model space in the $[0.706, 0.709]^T$ direction. This is a stable mapping back and forth, since noise in the data is damped when it is mapped back into model space. The pairing between \mathbf{v}_2 and \mathbf{u}_2 directions is less stable, however, since a unit length change in data space parallel to \mathbf{u}_2 will be mapped back into a change of length $1/(0.00316) = 317$ parallel to \mathbf{v}_2 . The \mathbf{v}_2 direction in model space will be associated with a very large variance. Since \mathbf{v}_2 has significant components along both m_1 and m_2 , they will both individually have large variances, as seen in the unit model covariance matrix for this example, given by

$$[\text{cov}_{\mathbf{u}} \mathbf{m}] = \begin{bmatrix} 50,551 & -50,154 \\ -50,154 & 49,753 \end{bmatrix} \quad (8.11)$$

For a particular inverse problem, these figures can help one understand both the directions in model space that affect the data the least or most and the directions in data space along which noise will affect the estimated solution the least or most.

8.2 Including Prior Information, or the Weighted Generalized Inverse

8.2.1 Mathematical Background

As we have seen, the generalized inverse operator is a very powerful operator, combining the attributes of both least squares and minimum length estimators. Specifically, the generalized inverse minimizes both

$$\mathbf{e}^T \mathbf{e} = [\mathbf{d} - \mathbf{d}^{\text{pre}}]^T [\mathbf{d} - \mathbf{d}^{\text{pre}}] = [\mathbf{d} - \mathbf{G}\mathbf{m}]^T [\mathbf{d} - \mathbf{G}\mathbf{m}] \quad (8.12)$$

and $[\mathbf{m} - \langle \mathbf{m} \rangle]^T [\mathbf{m} - \langle \mathbf{m} \rangle]$, where $\langle \mathbf{m} \rangle$ is the *a priori* estimate of the solution.

As discussed in Chapter 3, however, it is useful to include as much prior information into an inverse problem as possible. Two forms of prior information were included in weighted least squares and weighted minimum length, and resulted in new minimization criteria given by

$$\mathbf{e}^T \mathbf{W}_e \mathbf{e} = \mathbf{e}^T [\text{cov } \mathbf{d}]^{-1} \mathbf{e} = [\mathbf{d} - \mathbf{G}\mathbf{m}]^T [\text{cov } \mathbf{d}]^{-1} [\mathbf{d} - \mathbf{G}\mathbf{m}] \quad (8.13)$$

and

$$\mathbf{m}^T \mathbf{W}_m \mathbf{m} = [\mathbf{m} - \langle \mathbf{m} \rangle]^T [\text{cov } \mathbf{m}]^{-1} [\mathbf{m} - \langle \mathbf{m} \rangle] \quad (8.14)$$

where $[\text{cov } \mathbf{d}]$ and $[\text{cov } \mathbf{m}]$ are *a priori* data and model covariance matrices, respectively. It is possible to include this information in a generalized inverse analysis as well.

The basic procedure is as follows. First, transform the problem into a coordinate system where the new data and model parameters each have uncorrelated errors and unit variance. The transformations are based on the information contained in the *a priori* data and model parameter covariance matrices. Then, perform a generalized inverse analysis in the transformed coordinate system. This is the appropriate inverse operator because both of the covariance matrices are identity matrices. Finally, transform everything back to the original coordinates to obtain the final solution.

One may assume that the data covariance matrix $[\text{cov } \mathbf{d}]$ is a positive definite Hermitian matrix. This is equivalent to assuming that all variances are positive, and none of the correlation coefficients are exactly equal to plus or minus one. Then the data covariance matrix can be decomposed as

$$[\text{cov } \mathbf{d}] = \begin{matrix} \mathbf{B} & \Lambda_d & \mathbf{B}^T \\ N \times N & N \times N & N \times N \end{matrix} \quad (8.15)$$

where Λ_d is a diagonal matrix containing the eigenvalues of $[\text{cov } \mathbf{d}]$ and \mathbf{B} is an orthonormal matrix containing the associated eigenvectors. \mathbf{B} is orthonormal because $[\text{cov } \mathbf{d}]$ is Hermitian, and all of the eigenvalues are positive because $[\text{cov } \mathbf{d}]$ is positive definite.

The inverse data covariance matrix is easily found as

$$[\text{cov } \mathbf{d}]^{-1} = \begin{matrix} \mathbf{B} & \Lambda_d^{-1} & \mathbf{B}^T \\ N \times N & N \times N & N \times N \end{matrix} \quad (8.16)$$

where we have taken advantage of the fact that \mathbf{B} is an orthonormal matrix. It is convenient to write the right-hand side of (8.16) as

$$\begin{matrix} \mathbf{B} & \Lambda_d^{-1} & \mathbf{B}^T \\ N \times N & N \times N & N \times N \end{matrix} = \begin{matrix} \mathbf{D}^T & \mathbf{D} \\ N \times N & N \times N \end{matrix} \quad (8.17)$$

where

$$\mathbf{D} = \Lambda_d^{-1/2} \mathbf{B}^T \quad (8.18)$$

Thus,

$$[\text{cov } \mathbf{d}]^{-1} = \begin{matrix} \mathbf{D}^T & \mathbf{D} \\ N \times N & N \times N \end{matrix} \quad (8.19)$$

The reason for writing the data covariance matrix in terms of \mathbf{D} will be clear when we introduce the transformed data vector. The covariance matrix itself can be expressed in terms of \mathbf{D} as

$$\begin{aligned} [\text{cov } \mathbf{d}] &= \{[\text{cov } \mathbf{d}]^{-1}\}^{-1} \\ &= [\mathbf{D}^T \mathbf{D}]^{-1} \\ &= \mathbf{D}^{-1} [\mathbf{D}^T]^{-1} \end{aligned} \quad (8.20)$$

Similarly, the positive definite Hermitian model covariance matrix may be decomposed as

$$[\text{cov } \mathbf{m}] = \begin{matrix} \mathbf{M} & \Lambda_m & \mathbf{M}^T \\ M \times M & M \times M & M \times M \end{matrix} \quad (8.21)$$

where Λ_m is a diagonal matrix containing the eigenvalues of $[\text{cov } \mathbf{m}]$ and \mathbf{M} is an orthonormal matrix containing the associated eigenvectors.

The inverse model covariance matrix is thus given by

$$[\text{cov } \mathbf{m}]^{-1} = \begin{matrix} \mathbf{M} & \Lambda_m^{-1} & \mathbf{M}^T \\ M \times M & M \times M & M \times M \end{matrix} \quad (8.22)$$

where, as before, we have taken advantage of the fact that \mathbf{M} is an orthonormal matrix. The right-hand side of (8.22) can be written as

$$\begin{matrix} \mathbf{M} & \Lambda_m^{-1} & \mathbf{M}^T \\ M \times M & M \times M & M \times M \end{matrix} = \begin{matrix} \mathbf{S}^T & \mathbf{S} \\ M \times M & M \times M \end{matrix} \quad (8.23)$$

where

$$\mathbf{S} = \Lambda_m^{-1/2} \mathbf{M}^T \quad (8.24)$$

Thus,

$$[\text{cov } \mathbf{m}]^{-1} = \begin{matrix} \mathbf{S}^T & \mathbf{S} \\ M \times M & M \times M \end{matrix} \quad (8.25)$$

As before, it is possible to write the covariance matrix in terms of \mathbf{S} as

$$\begin{aligned} [\text{cov } \mathbf{m}] &= \{[\text{cov } \mathbf{m}]^{-1}\}^{-1} \\ &= [\mathbf{S}^T \mathbf{S}]^{-1} \\ &= \mathbf{S}^{-1} [\mathbf{S}^T]^{-1} \end{aligned} \quad (8.26)$$

8.2.2 Coordinate System Transformation of Data and Model Parameter Vectors

The utility of \mathbf{D} and \mathbf{S} can now be seen as we introduce transformed data and model parameter vectors. First, we introduce a transformed data vector \mathbf{d}' as

$$\mathbf{d}' = \Lambda_d^{-1/2} \mathbf{B}^T \mathbf{d} \quad (8.27)$$

or

$$\mathbf{d}' = \mathbf{D} \mathbf{d} \quad (8.28)$$

The transformed model parameter \mathbf{m}' is given by

$$\mathbf{m}' = \Lambda_m^{-1/2} \mathbf{M}^T \mathbf{m} \quad (8.29)$$

or

$$\mathbf{m}' = \mathbf{S} \mathbf{m} \quad (8.30)$$

The forward operator \mathbf{G} must also be transformed into \mathbf{G}' , the new coordinates. The transformation can be found by recognizing that

$$\mathbf{G}' \mathbf{m}' = \mathbf{d}' \quad (8.31)$$

$$\mathbf{G}' \mathbf{S} \mathbf{m} = \mathbf{D} \mathbf{d} \quad (8.32)$$

or

$$\mathbf{D}^{-1} \mathbf{G}' \mathbf{S} \mathbf{m} = \mathbf{d} = \mathbf{G} \mathbf{m} \quad (8.33)$$

That is

$$\mathbf{D}^{-1} \mathbf{G}' \mathbf{S} = \mathbf{G} \quad (8.34)$$

Finally, by pre- and postmultiplying by \mathbf{D} and \mathbf{S}^{-1} , respectively, we obtain \mathbf{G}' as

$$\mathbf{G}' = \mathbf{D} \mathbf{G} \mathbf{S}^{-1} \quad (8.35)$$

The transformations back from the primed coordinates to the original coordinates are given by

$$\mathbf{d} = \mathbf{B} \Lambda_d^{1/2} \mathbf{d}' \quad (8.36)$$

or

$$\mathbf{d} = \mathbf{D}^{-1} \mathbf{d}' \quad (8.37)$$

$$\mathbf{m} = \mathbf{M}\Lambda_m^{1/2} \mathbf{m}' \quad (8.38)$$

or

$$\mathbf{m} = \mathbf{S}^{-1}\mathbf{m}' \quad (8.39)$$

and

$$\mathbf{G} = \mathbf{B}\Lambda_d^{1/2}\mathbf{G}'\Lambda_m^{1/2}\mathbf{M}^T \quad (8.40)$$

or

$$\mathbf{G} = \mathbf{D}^{-1}\mathbf{G}'\mathbf{S} \quad (8.41)$$

In the new coordinate system, the generalized inverse will minimize

$$\mathbf{e}'^T\mathbf{e}' = [\mathbf{d}' - \mathbf{d}'_{\text{pre}}]^T[\mathbf{d}' - \mathbf{d}'_{\text{pre}}] = [\mathbf{d}' - \mathbf{G}'\mathbf{m}']^T[\mathbf{d}' - \mathbf{G}'\mathbf{m}'] \quad (8.42)$$

and $[\mathbf{m}']^T\mathbf{m}'$.

Replacing \mathbf{d}' , \mathbf{m}' and \mathbf{G}' in (8.42) with Equations (8.27)–(8.35), we have

$$\begin{aligned} [\mathbf{d}' - \mathbf{G}'\mathbf{m}']^T[\mathbf{d}' - \mathbf{G}'\mathbf{m}'] &= [\mathbf{D}\mathbf{d} - \mathbf{D}\mathbf{G}\mathbf{S}^{-1}\mathbf{S}\mathbf{m}]^T[\mathbf{D}\mathbf{d} - \mathbf{D}\mathbf{G}\mathbf{S}^{-1}\mathbf{S}\mathbf{m}] \\ &= [\mathbf{D}\mathbf{d} - \mathbf{D}\mathbf{G}\mathbf{m}]^T[\mathbf{D}\mathbf{d} - \mathbf{D}\mathbf{G}\mathbf{m}] \\ &= \{\mathbf{D}[\mathbf{d} - \mathbf{G}\mathbf{m}]\}^T\{\mathbf{D}[\mathbf{d} - \mathbf{G}\mathbf{m}]\} \\ &= [\mathbf{d} - \mathbf{G}\mathbf{m}]^T\mathbf{D}^T\mathbf{D}[\mathbf{d} - \mathbf{G}\mathbf{m}] \\ &= [\mathbf{d} - \mathbf{G}\mathbf{m}]^T[\text{cov } \mathbf{d}]^{-1}[\mathbf{d} - \mathbf{G}\mathbf{m}] \end{aligned} \quad (8.43)$$

where we have used (8.19) to replace $\mathbf{D}^T\mathbf{D}$ with $[\text{cov } \mathbf{d}]^{-1}$.

Equation (8.43) shows that the unweighted misfit in the primed coordinate system is precisely the weighted misfit to be minimized in the original coordinates. Thus, the least squares solution in the primed coordinate system is equivalent to weighted least squares in the original coordinates.

Furthermore, using (8.29) for \mathbf{m}' , we have

$$\begin{aligned} \mathbf{m}'^T\mathbf{m}' &= [\mathbf{S}\mathbf{m}]^T\mathbf{S}\mathbf{m} \\ &= \mathbf{m}^T\mathbf{S}^T\mathbf{S}\mathbf{m} \end{aligned}$$

$$= \mathbf{m}^T [\text{cov } \mathbf{m}]^{-1} \mathbf{m} \quad (8.44)$$

where we have used (8.25) to replace $\mathbf{S}^T \mathbf{S}$ with $[\text{cov } \mathbf{m}]^{-1}$.

Equation (8.44) shows that the unweighted minimum length solution in the new coordinate system is equivalent to the weighted minimum length solution in the original coordinate system. Thus minimum length in the new coordinate system is equivalent to weighted minimum length in the original coordinates.

8.2.3 The Maximum Likelihood Inverse Operator, Resolution, and Model Covariance

The generalized inverse operator in the primed coordinates can be transformed into an operator in the original coordinates. We will show later that this is, in fact, the maximum likelihood operator in the case where all distributions are Gaussian. Let this inverse operator be \mathbf{G}_{MX}^{-1} , and be given by

$$\begin{aligned} \mathbf{G}_{MX}^{-1} &= [\mathbf{D}^{-1} \mathbf{G}' \mathbf{S}]_g^{-1} \\ &= \mathbf{S}^{-1} [\mathbf{G}']_g^{-1} \mathbf{D} \end{aligned} \quad (8.45)$$

The solution in the original coordinates, \mathbf{m}_{MX} , can be expressed either as

$$\mathbf{m}_{MX} = \mathbf{G}_{MX}^{-1} \mathbf{d} \quad (8.46)$$

or as

$$\begin{aligned} \mathbf{m}_{MX} &= \mathbf{S}^{-1} \mathbf{m}_g' \\ &= \mathbf{S}^{-1} [\mathbf{G}']_g^{-1} \mathbf{d}' \end{aligned} \quad (8.47)$$

Now that the operator has been expressed in the original coordinates, it is possible to calculate the resolution matrices and an *a posteriori* model covariance matrix.

The model resolution matrix \mathbf{R} is given by

$$\begin{aligned} \mathbf{R} &= \mathbf{G}_{MX}^{-1} \mathbf{G} \\ &= \{\mathbf{S}^{-1} [\mathbf{G}']_g^{-1} \mathbf{D}\} \{\mathbf{D}^{-1} \mathbf{G}' \mathbf{S}\} \\ &= \mathbf{S}^{-1} [\mathbf{G}']_g^{-1} \mathbf{G}' \mathbf{S} \\ &= \mathbf{S}^{-1} \mathbf{R}' \mathbf{S} \end{aligned} \quad (8.48)$$

where \mathbf{R}' is the model resolution matrix in the transformed coordinate system.

Similarly, the data resolution matrix \mathbf{N} is given by

$$\begin{aligned}
 \mathbf{N} &= \mathbf{G}\mathbf{G}_{\text{MX}}^{-1} \\
 &= \{\mathbf{D}^{-1}\mathbf{G}'\mathbf{S}\}\{\mathbf{S}^{-1}[\mathbf{G}'^{-1}_{\text{g}}\mathbf{D}]\} \\
 &= \mathbf{D}^{-1}\mathbf{G}'[\mathbf{G}'^{-1}_{\text{g}}\mathbf{D}] \\
 &= \mathbf{D}^{-1}\mathbf{N}'\mathbf{D}
 \end{aligned} \tag{8.49}$$

The *a posteriori* model covariance matrix $[\text{cov } \mathbf{m}]_P$ is given by

$$[\text{cov } \mathbf{m}]_P = \mathbf{G}_{\text{MX}}^{-1} [\text{cov } \mathbf{d}][\mathbf{G}_{\text{MX}}^{-1}]^T \tag{8.50}$$

Replacing $[\text{cov } \mathbf{d}]$ in (8.50) with (8.20) gives

$$\begin{aligned}
 [\text{cov } \mathbf{m}]_P &= \mathbf{G}_{\text{MX}}^{-1} \mathbf{D}^{-1}[\mathbf{D}^T]^{-1}[\mathbf{G}_{\text{MX}}^{-1}]^T \\
 &= \{\mathbf{S}^{-1}[\mathbf{G}'^{-1}_{\text{g}}\mathbf{D}]\mathbf{D}^{-1}[\mathbf{D}^T]^{-1}\{\mathbf{S}^{-1}[\mathbf{G}'^{-1}_{\text{g}}\mathbf{D}]\}^T \\
 &= \mathbf{S}^{-1}[\mathbf{G}'^{-1}_{\text{g}}\mathbf{D}\mathbf{D}^{-1}[\mathbf{D}^T]^{-1}\mathbf{D}^T\{[\mathbf{G}'^{-1}_{\text{g}}]\}^T[\mathbf{S}^{-1}]^T \\
 &= \mathbf{S}^{-1}[\mathbf{G}'^{-1}_{\text{g}}\{[\mathbf{G}'^{-1}_{\text{g}}]\}^T[\mathbf{S}^{-1}]^T \\
 &= \mathbf{S}^{-1}[\text{cov}_u \mathbf{m}'][\mathbf{S}^{-1}]^T
 \end{aligned} \tag{8.51}$$

That is, an *a posteriori* estimate of model parameter uncertainties can be obtained by transforming the unit model covariance matrix from the primed coordinates back to the original coordinates.

It is important to realize that the transformations introduced by \mathbf{D} and \mathbf{S} in (8.27)–(8.41) are not, in general, orthonormal. Thus,

$$\mathbf{d}' = \mathbf{D}\mathbf{d} \tag{8.28}$$

implies that the length of the transformed data vector \mathbf{d}' is, in general, not equal to the length of the original data vector \mathbf{d} . The function of \mathbf{D} is to transform the data space into one in which the data errors are uncorrelated and all observations have unit variance. If the original data errors are uncorrelated, the data covariance matrix will be diagonal and \mathbf{B} , from

$$\begin{aligned}
 [\text{cov } \mathbf{d}]^{-1} &= \mathbf{B} \quad \Lambda_d^{-1} \quad \mathbf{B}^T \\
 N \times N \quad &N \times N \quad N \times N \quad N \times N
 \end{aligned} \tag{8.16}$$

will be an identity matrix. Then \mathbf{D} , given by

$$\mathbf{D} = \Lambda_d^{-1/2} \mathbf{B}^T \quad (8.18)$$

will be a diagonal matrix given by $\Lambda_d^{-1/2}$. The transformed data \mathbf{d}' are then given by

$$\mathbf{d}' = \Lambda_d^{-1/2} \mathbf{d} \quad (8.52)$$

or

$$d'_i = d_i / \sigma_{di} \quad i = 1, N \quad (8.53)$$

where σ_{di} is the data standard deviation for the i th observation. If the original data errors are uncorrelated, then each transformed observation is given by the original observation, divided by its standard deviation. The transformation in this case can be thought of as leaving the direction of each axis in data space unchanged, but stretching or compressing each axis, depending on the standard deviation. To see this, consider a vector in data space representing the d_1 axis. That is,

$$\mathbf{d} = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad (8.54)$$

This data vector is transformed into

$$\mathbf{d}' = \begin{bmatrix} 1 / \sigma_{d1} \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad (8.55)$$

That is, the direction of the axis is unchanged, but the magnitude is changed by $1/\sigma_{d1}$. If the data errors are correlated, then the axes in data space are rotated (by \mathbf{B}^T), and then stretched or compressed.

Very similar arguments can be made about the role of \mathbf{S} in model space. That is, if the *a priori* model covariance matrix is diagonal, then the directions of the transformed axes in model space are the same as in the original coordinates (i.e., m_1, m_2, \dots, m_M), but the lengths are stretched or compressed by the appropriate model parameter standard deviations. If the errors are correlated, then the axes in model space are rotated (by \mathbf{M}^T) before they are stretched or compressed.

8.2.4 Effect on Model- and Data-Space Eigenvectors

This stretching and compressing of directions in data and model space affects the eigenvectors as well. Let $\hat{\mathbf{V}}$ be the set of vectors transformed back into the original coordinates from \mathbf{V}' , the set of model eigenvectors in the primed coordinates. Thus,

$$\hat{\mathbf{V}} = \mathbf{S}^{-1}\mathbf{V}' \quad (8.56)$$

For example, suppose that $[\text{cov } \mathbf{m}]$ is diagonal, then

$$\hat{\mathbf{V}} = \Lambda_m^{1/2}\mathbf{V}' \quad (8.57)$$

For $\hat{\mathbf{v}}_i$, the i th vector in $\hat{\mathbf{V}}$, this implies

$$\begin{bmatrix} \hat{\mathbf{v}}_1 \\ \hat{\mathbf{v}}_2 \\ \vdots \\ \hat{\mathbf{v}}_M \end{bmatrix}_i = \begin{bmatrix} \sigma_{m1} \mathbf{v}_1 \\ \sigma_{m2} \mathbf{v}_2 \\ \vdots \\ \sigma_{mM} \mathbf{v}_M \end{bmatrix}_i \quad (8.58)$$

Clearly, for a general diagonal $[\text{cov } \mathbf{m}]$, $\hat{\mathbf{v}}_i$ will no longer have unit length. This is true whether or not $[\text{cov } \mathbf{m}]$ is diagonal. Thus, in general, the vectors in $\hat{\mathbf{V}}$ are not unit length vectors. They can, of course, be normalized to unit length. Perhaps more importantly, however, the directions of the $\hat{\mathbf{v}}_i$ have been changed, and the vectors in $\hat{\mathbf{V}}$ are no longer perpendicular to each other. Thus, the vectors in $\hat{\mathbf{V}}$ cannot be thought of as orthonormal eigenvectors, even if they have been normalized to unit length.

These vectors still play an important role in the inverse analysis, however. Recall that the solution \mathbf{m}_{MX} is given by

$$\mathbf{m}_{\text{MX}} = \mathbf{G}_{\text{MX}}^{-1} \mathbf{d} \quad (8.46)$$

or as

$$\begin{aligned} \mathbf{m}_{\text{MX}} &= \mathbf{S}^{-1} \mathbf{m}_g' \\ &= \mathbf{S}^{-1} [\mathbf{G}'_g]^{-1} \mathbf{d}' \end{aligned} \quad (8.47)$$

We can expand (8.47) as

$$\begin{aligned} \mathbf{m}_{\text{MX}} &= \mathbf{S}^{-1} \mathbf{V}'_p [\Lambda'_p]^{-1} [\mathbf{U}'_p]^T \mathbf{D} \mathbf{d} \\ &= \hat{\mathbf{V}}_p [\Lambda'_p]^{-1} [\mathbf{U}'_p]^T \mathbf{D} \mathbf{d} \end{aligned} \quad (8.59)$$

Recall that the solution \mathbf{m}_{MX} can be thought of as a linear combination of the columns of the first matrix in a product of several matrices [see Equations (2.23)–(2.30)]. This implies that the solution \mathbf{m}_{MX} consists of a linear combination of the columns of $\hat{\mathbf{V}}_P$. The solution is still a linear combination of the vectors in $\hat{\mathbf{V}}_P$, even if they have been normalized to unit length. Thus, $\hat{\mathbf{V}}_P$ still plays a fundamental role in the inverse analysis.

It is important to realize that [cov \mathbf{m}] will only affect the solution if $P < M$. If $P = M$, then $\mathbf{V}'_P = \mathbf{V}'$, and \mathbf{V}'_P spans all of model space. $\hat{\mathbf{V}}_P$ will also span all of solution space. In this case, all of model space can be expressed as a linear combination of the vectors in $\hat{\mathbf{V}}_P$, even though they are not an orthonormal set of vectors. Thus, the same solution will be reached, regardless of the values in [cov \mathbf{m}]. If $P < M$, however, the mapping of vectors from the primed coordinates back to the original space can affect the part of solution space that is spanned by $\hat{\mathbf{V}}_P$. We will return to this point later with a specific example.

Very similar arguments can be made for the data eigenvectors as well. Let $\hat{\mathbf{U}}$ be the set of vectors obtained by transforming the data eigenvectors \mathbf{U}' in the primed coordinates back into the original coordinates. Then

$$\hat{\mathbf{U}} = \mathbf{D}^{-1}\mathbf{U}' \quad (8.60)$$

In general, the vectors in $\hat{\mathbf{U}}$ will not be either of unit length or perpendicular to each other.

The predicted data $\hat{\mathbf{d}}$ are given by

$$\begin{aligned} \hat{\mathbf{d}} &= \mathbf{G} \mathbf{m}_{MX} \\ &= \mathbf{D}^{-1}\mathbf{G}'\mathbf{S}\mathbf{m}_{MX} \\ &= \mathbf{D}^{-1}\mathbf{U}'_P\Lambda'_P\mathbf{V}'_P\mathbf{S}\mathbf{m}_{MX} \\ &= \hat{\mathbf{U}}_P\Lambda'_P\mathbf{V}'_P\mathbf{S}\mathbf{m}_{MX} \end{aligned} \quad (8.61)$$

Thus, the predicted data are a linear combination of the columns of $\hat{\mathbf{U}}_P$.

It is important to realize that the transformations introduced by [cov \mathbf{d}] will only affect the solution if $P < N$. If $P = N$, then $\mathbf{U}'_P = \mathbf{U}'$, and \mathbf{U}'_P spans all of data space. The matrix $\hat{\mathbf{U}}_P$ will also span all of data space. In this case, all of data space can be expressed as a linear combination of the vectors in $\hat{\mathbf{U}}_P$, even though they are not an orthonormal set of vectors. Thus, the same solution will be reached, regardless of the values in [cov \mathbf{d}]. If $P < N$, however, the mapping of vectors from the primed coordinates back to the original space can affect the part of solution space that is spanned by $\hat{\mathbf{U}}_P$. We are now in a position to consider a specific example.

8.2.5 An Example

Consider the following specific example of the form $\mathbf{G}\mathbf{m} = \mathbf{d}$, where \mathbf{G} and \mathbf{d} are given by

$$\mathbf{G} = \begin{bmatrix} 1.00 & 1.00 \\ 2.00 & 2.00 \end{bmatrix} \quad (8.62)$$

$$\mathbf{d} = \begin{bmatrix} 4.00 \\ 5.00 \end{bmatrix} \quad (8.63)$$

If we assume for the moment that the *a priori* data and model parameter covariance matrices are identity matrices and perform a generalized inverse analysis, we obtain

$$P = 1 < M = N = 2 \quad (8.64)$$

$$\lambda_1 = 3.162 \quad (8.65)$$

$$\mathbf{V} = \begin{bmatrix} 0.707 & 0.707 \\ 0.707 & -0.707 \end{bmatrix} \quad (8.66)$$

$$\mathbf{U} = \begin{bmatrix} 0.447 & 0.894 \\ 0.894 & -0.447 \end{bmatrix} \quad (8.67)$$

$$\mathbf{R} = \begin{bmatrix} 0.500 & 0.500 \\ 0.500 & 0.500 \end{bmatrix} \quad (8.68)$$

$$\mathbf{N} = \begin{bmatrix} 0.200 & 0.400 \\ 0.400 & 0.800 \end{bmatrix} \quad (8.69)$$

$$\mathbf{m}_g = \begin{bmatrix} 1.400 \\ 1.400 \end{bmatrix} \quad (8.70)$$

$$\hat{\mathbf{d}} = \begin{bmatrix} 2.800 \\ 5.600 \end{bmatrix} \quad (8.71)$$

$$\mathbf{e}^T \mathbf{e} = \mathbf{e}^T [\text{cov } \mathbf{d}]^{-1} \mathbf{e} = 1.800 \quad (8.72)$$

The two rows (or columns) of \mathbf{G} are linearly dependent, and thus the number of nonzero singular values is one. Thus, the first column of \mathbf{V} (or \mathbf{U}) gives \mathbf{V}_P (or \mathbf{U}_P), while the second column gives \mathbf{V}_0 (or \mathbf{U}_0). The generalized inverse solution \mathbf{m}_g must lie in \mathbf{V}_P space, and is thus parallel to the $[0.707, 0.707]^T$ direction in model space. Similarly, the predicted data $\hat{\mathbf{d}}$ must lie

in \mathbf{U}_P space, and is thus parallel to the $[0.447, 0.894]^T$ direction in data space. The model resolution matrix \mathbf{R} indicates that only the sum, equally weighted, of the model parameters m_1 and m_2 is resolved. Similarly, the data resolution matrix \mathbf{N} indicates that only the sum of d_1 and d_2 , with more weight on d_2 , is resolved, or important, in constraining the solution.

Now let us assume that the *a priori* data and model parameter covariance matrices are not equal to a constant times the identity matrix. Suppose

$$[\text{cov } \mathbf{d}] = \begin{bmatrix} 4.362 & -2.052 \\ -2.052 & 15.638 \end{bmatrix} \quad (8.73)$$

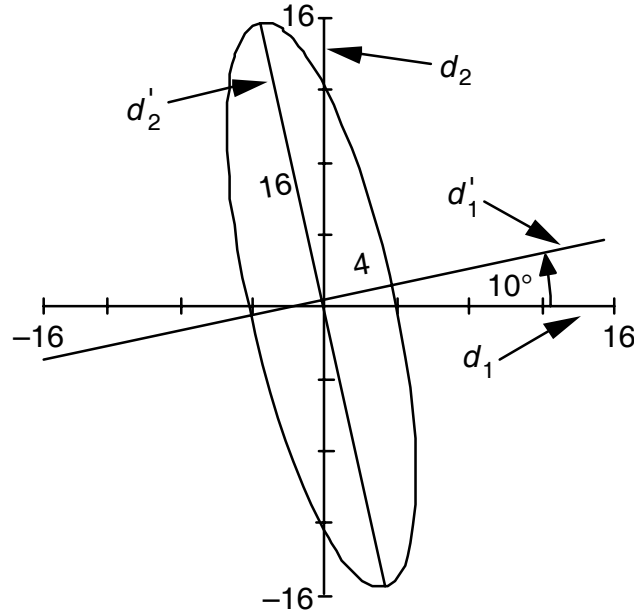
and

$$[\text{cov } \mathbf{m}] = \begin{bmatrix} 23.128 & 5.142 \\ 5.142 & 10.872 \end{bmatrix} \quad (8.74)$$

The data covariance matrix $[\text{cov } \mathbf{d}]$ can be decomposed as

$$\begin{aligned} [\text{cov } \mathbf{d}] &= \mathbf{B} \Lambda_d \mathbf{B}^T \\ &= \begin{bmatrix} 0.985 & -0.174 \\ 0.174 & 0.985 \end{bmatrix} \begin{bmatrix} 4.000 & 0.000 \\ 0.000 & 16.000 \end{bmatrix} \begin{bmatrix} 0.985 & 0.174 \\ -0.174 & 0.985 \end{bmatrix} \\ &= \begin{bmatrix} 4.362 & -2.052 \\ -2.052 & 15.638 \end{bmatrix} \end{aligned} \quad (8.75)$$

Recall that \mathbf{B} contains the eigenvectors of the symmetric matrix $[\text{cov } \mathbf{d}]$. Furthermore, these eigenvectors represent the directions of the major and minor axes of an ellipse. Thus, for the present case, the first vector in \mathbf{B} , $[0.985, 0.174]^T$, is the direction in data space of the minor axis of an ellipse having a half-length of 4. Similarly, the second vector in \mathbf{B} , $[-0.174, 0.985]^T$, is the direction in data space of the major axis of an ellipse having length 16. The eigenvectors in \mathbf{B}^T represent a 10° counterclockwise rotation of data space, as shown on the next page:



The negative off-diagonal entries in $[\text{cov } \mathbf{d}]$ indicate a negative correlation of errors between d_1 and d_2 . Compare the figure above with figure (c) just after Equation (2.43).

The inverse data covariance $[\text{cov } \mathbf{d}]^{-1}$ can also be written as

$$\begin{aligned} [\text{cov } \mathbf{d}]^{-1} &= \mathbf{B}\Lambda_d^{-1}\mathbf{B}^T \\ &= \mathbf{D}^T\mathbf{D} \end{aligned} \quad (8.76)$$

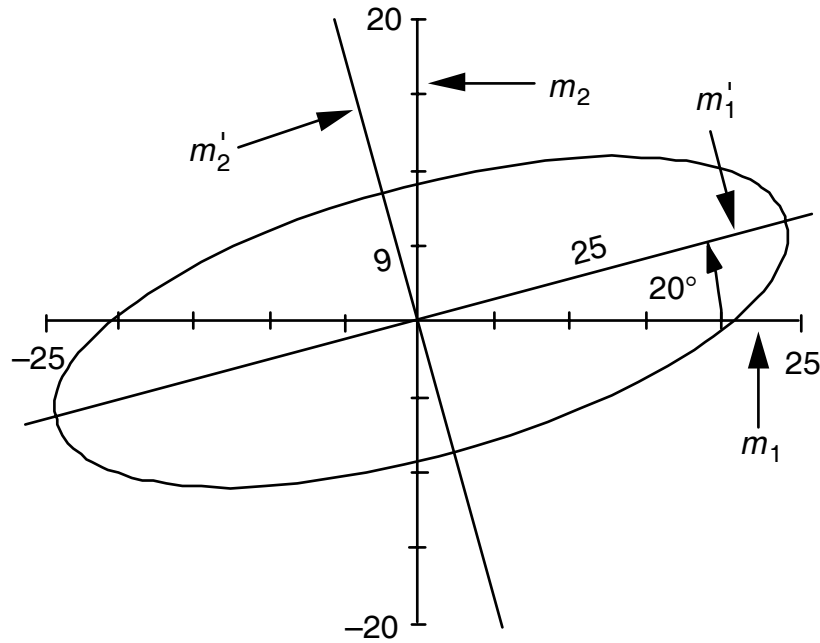
where \mathbf{D} is given by

$$\begin{aligned} \mathbf{D} &= \Lambda_d^{-1/2}\mathbf{B}^T \\ &= \begin{bmatrix} 0.492 & 0.087 \\ -0.043 & 0.246 \end{bmatrix} \end{aligned} \quad (8.77)$$

Similarly, the model covariance matrix $[\text{cov } \mathbf{m}]$ can be decomposed as

$$\begin{aligned} [\text{cov } \mathbf{m}] &= \mathbf{M}\Lambda_m\mathbf{M}^T \\ &= \begin{bmatrix} 0.940 & -0.342 \\ 0.342 & 0.940 \end{bmatrix} \begin{bmatrix} 25.000 & 0.000 \\ 0.000 & 9.000 \end{bmatrix} \begin{bmatrix} 0.940 & 0.342 \\ -0.342 & 0.940 \end{bmatrix} \end{aligned} \quad (8.78)$$

The matrix \mathbf{M}^T represents a 20° counterclockwise rotation of the m_1 and m_2 axes in model space. In the new coordinate system, the *a priori* model parameter errors are uncorrelated and have variances of 25 and 9, respectively. The major and minor axes of the error ellipse are along the $[0.940, 0.342]^T$ and $[-0.342, 0.940]^T$ directions, respectively. The geometry of the problem in model space is shown below:



The inverse model parameter covariance matrix $[\text{cov } \mathbf{m}]^{-1}$ can also be written as

$$\begin{aligned} [\text{cov } \mathbf{m}]^{-1} &= \mathbf{M}\Lambda_m^{-1}\mathbf{M}^T \\ &= \mathbf{S}^T\mathbf{S} \end{aligned} \tag{8.79}$$

where \mathbf{S} is given by

$$\begin{aligned} \mathbf{S} &= \Lambda_m^{-1/2}\mathbf{M}^T \\ &= \begin{bmatrix} 0.188 & 0.068 \\ -0.114 & 0.313 \end{bmatrix} \end{aligned} \tag{8.80}$$

With the information in \mathbf{D} and \mathbf{S} , it is now possible to transform \mathbf{G} , \mathbf{d} , and \mathbf{m} into \mathbf{G}' , \mathbf{d}' , and \mathbf{m}' in the new coordinate system:

$$\begin{aligned} \mathbf{G}' &= \mathbf{D}\mathbf{G}\mathbf{S}^{-1} \\ &= \begin{bmatrix} 0.492 & 0.087 \\ -0.043 & 0.246 \end{bmatrix} \begin{bmatrix} 1.000 & 1.000 \\ 2.000 & 2.000 \end{bmatrix} \begin{bmatrix} 4.698 & -1.026 \\ 1.710 & 2.819 \end{bmatrix} \\ &= \begin{bmatrix} 4.26844 & 1.19424 \\ 2.87739 & 0.80505 \end{bmatrix} \end{aligned} \tag{8.81}$$

and

$$\begin{aligned}
 \mathbf{d}' &= \mathbf{D}\mathbf{d} \\
 &= \begin{bmatrix} 0.492 & 0.087 \\ -0.043 & 0.246 \end{bmatrix} \begin{bmatrix} 4.000 \\ 5.000 \end{bmatrix} \\
 &= \begin{bmatrix} 2.40374 \\ 1.05736 \end{bmatrix}
 \end{aligned} \tag{8.82}$$

In the new coordinate system, the data and model parameter covariance matrices are identity matrices. Thus, a generalized inverse analysis gives

$$P = 1 < M = N = 2 \tag{8.83}$$

$$\lambda_1 = 5.345 \tag{8.84}$$

$$\mathbf{V}' = \begin{bmatrix} 0.963 & -0.269 \\ 0.269 & 0.963 \end{bmatrix} \tag{8.85}$$

$$\mathbf{U}' = \begin{bmatrix} 0.829 & -0.559 \\ 0.559 & 0.829 \end{bmatrix} \tag{8.86}$$

$$\mathbf{G}'_g^{-1} = \begin{bmatrix} 0.149 & 0.101 \\ 0.042 & 0.028 \end{bmatrix} \tag{8.87}$$

$$\mathbf{R}' = \begin{bmatrix} 0.927 & 0.259 \\ 0.259 & 0.073 \end{bmatrix} \tag{8.88}$$

$$\mathbf{N}' = \begin{bmatrix} 0.688 & 0.463 \\ 0.463 & 0.312 \end{bmatrix} \tag{8.89}$$

$$\mathbf{m}'_g = \begin{bmatrix} 0.466 \\ 0.130 \end{bmatrix} \tag{8.90}$$

$$\hat{\mathbf{d}}' = \begin{bmatrix} 2.143 \\ 1.145 \end{bmatrix} \tag{8.91}$$

$$[\mathbf{e}']^T \mathbf{e}' = [\mathbf{e}']^T [\text{cov } \mathbf{d}']^{-1} \mathbf{e}' = 0.218 \tag{8.92}$$

The results may be transformed back to the original coordinates, using Equations (8.37), (8.39), (8.44), (8.48), (8.49), (8.56), and (8.60) as

$$\mathbf{G}_{MX}^{-1} = \begin{bmatrix} 0.305 & 0.167 \\ 0.173 & 0.094 \end{bmatrix} \quad (8.93)$$

$$\lambda_1 = 5.345 \quad (8.94)$$

$$\begin{aligned} \mathbf{m}_{MX} &= \mathbf{S}^{-1} \mathbf{m}'_g \\ &= \begin{bmatrix} 2.054 \\ 1.163 \end{bmatrix} \end{aligned} \quad (8.95)$$

$$\begin{aligned} \hat{\mathbf{d}} &= \mathbf{D}^{-1} \hat{\mathbf{d}}' \\ &= \begin{bmatrix} 3.217 \\ 6.434 \end{bmatrix} \end{aligned} \quad (8.96)$$

$$\mathbf{e}^T \mathbf{e} = 2.670 \quad (8.97)$$

$$\begin{aligned} \mathbf{e}^T [\text{cov } \mathbf{d}]^{-1} \mathbf{e} &= \begin{bmatrix} 0.783 & -1.434 \end{bmatrix} \begin{bmatrix} 0.244 & 0.032 \\ 0.032 & 0.068 \end{bmatrix} \begin{bmatrix} 0.783 \\ -1.434 \end{bmatrix} \\ &= 0.218 \end{aligned} \quad (8.98)$$

$$\hat{\mathbf{V}} = \begin{bmatrix} 0.870 & -0.707 \\ 0.493 & 0.707 \end{bmatrix} \quad (8.99)$$

$$\hat{\mathbf{U}} = \begin{bmatrix} 0.447 & -0.479 \\ 0.894 & 0.878 \end{bmatrix} \quad (8.100)$$

$$\mathbf{R} = \begin{bmatrix} 0.639 & -0.639 \\ 0.362 & 0.362 \end{bmatrix} \quad (8.101)$$

$$\mathbf{N} = \begin{bmatrix} 0.478 & 0.261 \\ 0.956 & 0.522 \end{bmatrix} \quad (8.102)$$

Note that $\mathbf{e}^T \mathbf{e} = 2.670$ for the weighted case is larger than the misfit $\mathbf{e}^T \mathbf{e} = 1.800$ for the unweighted case. This is to be expected because the unweighted case should produce the smallest misfit. The weighted case provides an answer that gives more weight to better-known data, but it produces a larger total misfit.

The $\hat{\mathbf{U}}$ and $\hat{\mathbf{V}}$ matrices were obtained by transforming each eigenvector in the primed coordinate system into a vector in the original coordinates, and then scaling to unit length. Note that the vectors in $\hat{\mathbf{U}}$ (and $\hat{\mathbf{V}}$) are not perpendicular to each other. Note also that the solution \mathbf{m}_{MX} is parallel to the $[0.870, 0.493]^T$ direction in model space, also given by the first column of

$\hat{\mathbf{V}}$. The predicted data $\hat{\mathbf{d}}$ is parallel to the $[0.447, 0.894]^T$ direction in data space, also given by the first column in $\hat{\mathbf{U}}$.

The resolution matrices were obtained from the primed coordinate resolution matrices after Equations (8.48)–(8.49). Note that they are no longer symmetric matrices, but that the trace has remained equal to one. The model resolution matrix \mathbf{R} still indicates that only a sum of the two model parameters m_1 and m_2 is resolved, but now we see that the estimate of m_1 is better resolved than that of m_2 . This may not seem intuitively obvious, since the *a priori* variance of m_2 is less than that of m_1 , and thus m_2 is “better known.” Because m_2 is better known, the inverse operator will leave m_2 closer to its prior estimate. Thus, m_1 will be allowed to vary further from its prior estimate. It is in this sense that the resolution of m_1 is greater than that of m_2 . The data resolution matrix \mathbf{N} still indicates that only the sum of d_1 and d_2 is resolved, or important, in constraining the solution. Now, however, the importance of the first observation has been increased significantly from the unweighted case, reflecting the smaller variance for d_1 compared to d_2 .

8.3 Damped Least Squares and the Stochastic Inverse

8.3.1 Introduction

As we have seen, the presence of small singular values causes significant stability problems with the generalized inverse. One approach is simply to set small singular values to zero, and relegate the associated eigenvectors to the zero spaces. This improves stability, with an inevitable decrease in resolution. Ideally, the cut-off value for small singular values should be based on how noisy the data are. In practice, however, the decision is almost always arbitrary.

We will now introduce a damping term, the function of which is to improve the stability of inverse problems with small singular values. First, however, we will consider another inverse operator, the *stochastic inverse*.

8.3.2 The Stochastic Inverse

Consider a forward problem given by

$$\mathbf{Gm} + \mathbf{n} = \mathbf{d} \quad (8.103)$$

where \mathbf{n} is an $N \times 1$ noise vector. It is similar to

$$\mathbf{Gm} = \mathbf{d} \quad (1.13)$$

except that we explicitly separate out the contribution of noise to the total data vector \mathbf{d} . This has some important implications, however.

We assume that both \mathbf{m} and \mathbf{n} are stochastic (i.e., random variables, as described in Chapter 2, that are characterized by their statistical properties) processes, with mean (or expected) values of zero. This is natural for noise, but implies that the mean value must be subtracted from all model parameters. Furthermore, we assume that we have estimates for the model parameter and noise covariance matrices, [cov \mathbf{m}] and [cov \mathbf{n}], respectively.

The stochastic inverse is defined by minimizing the average, or statistical, discrepancy between \mathbf{m} and $\mathbf{G}_s^{-1}\mathbf{d}$, where \mathbf{G}_s^{-1} is the stochastic inverse. Let $\mathbf{G}_s^{-1} = \mathbf{L}$, and determine \mathbf{L} by minimizing

$$m_i - \sum_{j=1}^N L_{ij} d_j \quad (8.104)$$

for each i . Consider repeated experiments in which \mathbf{m} and \mathbf{n} are generated. Let these values, on the k th experiment, be \mathbf{m}_k and \mathbf{n}_k , respectively. If there are a total of q experiments, then we seek \mathbf{L} which minimizes

$$\frac{1}{q} \sum_{k=1}^q \left(m_i^k - \sum_{j=1}^N L_{ij} d_j^k \right)^2 \quad (8.105)$$

The minimum of Equation (8.105) is found by differentiating with respect to L_{il} and setting it equal to zero:

$$\frac{\partial}{\partial L_{il}} \left[\frac{1}{q} \sum_{k=1}^q \left(m_i^k - \sum_{j=1}^N L_{ij} d_j^k \right)^2 \right] = 0 \quad (8.106)$$

or

$$\frac{2}{q} \sum_{k=1}^q \left(m_i^k - \sum_{j=1}^N L_{ij} d_j^k \right) (-d_l^k) = 0 \quad (8.107)$$

This implies

$$\frac{1}{q} \sum_{k=1}^q m_i^k d_l^k = \frac{1}{q} \sum_{k=1}^q \left(\sum_{j=1}^N L_{ij} d_j^k \right) d_l^k \quad (8.108)$$

The left-hand side of Equation (8.108), when taken over i and l , is simply the covariance matrix between the model parameters and the data, or

$$[\text{cov } \mathbf{m}\mathbf{d}] = \langle \mathbf{m}\mathbf{d}^T \rangle \quad (8.109)$$

The right-hand side, again taken over i and l and recognizing that \mathbf{L} will not vary from experiment to experiment, gives [see Equation (2.63)]

$$\mathbf{L}[\text{cov } \mathbf{d}] = \mathbf{L}\langle \mathbf{d}\mathbf{d}^T \rangle \quad (8.110)$$

where $[\text{cov } \mathbf{d}]$ is the data covariance matrix. Note that $[\text{cov } \mathbf{d}]$ is not the same matrix used elsewhere in these notes. As used here, $[\text{cov } \mathbf{d}]$ is a derived quantity, based on $[\text{cov } \mathbf{m}]$ and $[\text{cov } \mathbf{n}]$. With Equations (8.109) and (8.110), we can write Equation (8.108), taken over i and l , as

$$[\text{cov } \mathbf{m}\mathbf{d}] = \mathbf{L}[\text{cov } \mathbf{d}] \quad (8.111)$$

or

$$\mathbf{L} = [\text{cov } \mathbf{m}\mathbf{d}][\text{cov } \mathbf{d}]^{-1} \quad (8.112)$$

We now need to rewrite $[\text{cov } \mathbf{d}]$ and $[\text{cov } \mathbf{m}\mathbf{d}]$ in terms of $[\text{cov } \mathbf{m}]$, $[\text{cov } \mathbf{n}]$, and \mathbf{G} . This is done as follows:

$$\begin{aligned} [\text{cov } \mathbf{d}] &= \langle \mathbf{d}\mathbf{d}^T \rangle \\ &= \langle [\mathbf{G}\mathbf{m} + \mathbf{n}][\mathbf{G}\mathbf{m} + \mathbf{n}]^T \rangle \\ &= \mathbf{G}\langle \mathbf{m}\mathbf{n}^T \rangle + \mathbf{G}\langle \mathbf{m}\mathbf{m}^T \rangle\mathbf{G}^T + \langle \mathbf{n}\mathbf{m}^T \rangle\mathbf{G}^T + \langle \mathbf{n}\mathbf{n}^T \rangle \end{aligned} \quad (8.113)$$

If we assume that model parameter and noise errors are uncorrelated, that is, that $\langle \mathbf{m}\mathbf{n}^T \rangle = 0 = \langle \mathbf{n}\mathbf{m}^T \rangle$, then Equation (8.113) reduces to

$$\begin{aligned} [\text{cov } \mathbf{d}] &= \mathbf{G}\langle \mathbf{m}\mathbf{m}^T \rangle\mathbf{G}^T + \langle \mathbf{n}\mathbf{n}^T \rangle \\ &= \mathbf{G}[\text{cov } \mathbf{m}]\mathbf{G}^T + [\text{cov } \mathbf{n}] \end{aligned} \quad (8.114)$$

Similarly,

$$\begin{aligned} [\text{cov } \mathbf{m}\mathbf{d}] &= \langle \mathbf{m}\mathbf{d}^T \rangle \\ &= \langle \mathbf{m}[\mathbf{G}\mathbf{m} + \mathbf{n}]^T \rangle \\ &= \langle \mathbf{m}\mathbf{m}^T \rangle\mathbf{G}^T + \langle \mathbf{m}\mathbf{n}^T \rangle \\ &= [\text{cov } \mathbf{m}]\mathbf{G}^T \end{aligned} \quad (8.115)$$

if $\langle \mathbf{m}\mathbf{n}^T \rangle = 0$.

Replacing $[\text{cov } \mathbf{m}\mathbf{d}]$ and $[\text{cov } \mathbf{d}]$ in Equation (8.112) with expressions from Equations (8.114) and (8.115), respectively, gives the definition of the *stochastic inverse operator* \mathbf{G}_s^{-1} as

$$\mathbf{G}_s^{-1} = [\text{cov } \mathbf{m}]\mathbf{G}^T \{ \mathbf{G}[\text{cov } \mathbf{m}]\mathbf{G}^T + [\text{cov } \mathbf{n}] \}^{-1} \quad (8.116)$$

Then the stochastic inverse solution, \mathbf{m}_s , is given by

$$\begin{aligned}\mathbf{m}_s &= \mathbf{G}_s^{-1}\mathbf{d} \\ &= [\text{cov } \mathbf{m}]\mathbf{G}^T[\text{cov } \mathbf{d}]^{-1}\mathbf{d}\end{aligned}\quad (8.117)$$

It is possible to decompose the symmetric covariance matrices $[\text{cov } \mathbf{d}]$ and $[\text{cov } \mathbf{m}]$ in exactly the same manner as was done for the maximum likelihood operator [Equations (8.19) and (8.25)]:

$$[\text{cov } \mathbf{d}] = \mathbf{B}\Lambda_d\mathbf{B}^T = \{\mathbf{B}\Lambda_d^{1/2}\}\{\Lambda_d^{1/2}\mathbf{B}^T\} = \mathbf{D}^{-1}[\mathbf{D}^{-1}]^T \quad (8.118)$$

$$[\text{cov } \mathbf{d}]^{-1} = \mathbf{B}\Lambda_d^{-1}\mathbf{B}^T = \mathbf{D}^T\mathbf{D} \quad (8.119)$$

$$[\text{cov } \mathbf{m}] = \mathbf{M}\Lambda_m\mathbf{M}^T = \{\mathbf{M}\Lambda_m^{1/2}\}\{\Lambda_m^{1/2}\mathbf{M}^T\} = \mathbf{S}^{-1}[\mathbf{S}^{-1}]^T \quad (8.120)$$

$$[\text{cov } \mathbf{m}]^{-1} = \mathbf{M}\Lambda_m^{-1}\mathbf{M}^T = \mathbf{S}^T\mathbf{S} \quad (8.121)$$

where Λ_d and Λ_m are the eigenvalues of $[\text{cov } \mathbf{d}]$ and $[\text{cov } \mathbf{m}]$, respectively. The orthogonal matrices \mathbf{B} and \mathbf{M} are the associated eigenvectors.

At this point it is useful to reintroduce a set of transformations based on the decompositions in (8.118)–(8.121) that will transform \mathbf{d} , \mathbf{m} , and \mathbf{G} back and forth between the original coordinate system and a primed coordinate system.

$$\mathbf{m}' = \mathbf{S}\mathbf{m} \quad (8.122)$$

$$\mathbf{d}' = \mathbf{D}\mathbf{d} \quad (8.123)$$

$$\mathbf{G}' = \mathbf{D}\mathbf{G}\mathbf{S}^{-1} \quad (8.124)$$

$$\mathbf{m} = \mathbf{S}^{-1}\mathbf{m}' \quad (8.125)$$

$$\mathbf{d} = \mathbf{D}^{-1}\mathbf{d}' \quad (8.126)$$

$$\mathbf{G} = \mathbf{D}^{-1}\mathbf{G}'\mathbf{S} \quad (8.127)$$

Then, Equation (8.117), using primed coordinate variables, is given by

$$\begin{aligned}\mathbf{S}^{-1}\mathbf{m}'_s &= [\text{cov } \mathbf{m}]\mathbf{G}^T[\text{cov } \mathbf{d}]^{-1}\mathbf{d} \\ \mathbf{S}^{-1}\mathbf{m}'_s &= \mathbf{S}^{-1}[\mathbf{S}^{-1}]^T[\mathbf{D}^{-1}\mathbf{G}'\mathbf{S}]^T\mathbf{D}^T\mathbf{D}\mathbf{D}^{-1}\mathbf{d}' \\ &= \mathbf{S}^{-1}[\mathbf{S}^{-1}]^T\mathbf{S}^T[\mathbf{G}']^T[\mathbf{D}^{-1}]^T\mathbf{D}^T\mathbf{d}'\end{aligned}\quad (8.128)$$

but
$$[\mathbf{S}^{-1}]^T\mathbf{S}^T = \mathbf{I}_M \quad (8.129)$$

and
$$[\mathbf{D}^{-1}]^T \mathbf{D}^T = \mathbf{I}_N \quad (8.130)$$

and hence
$$\mathbf{S}^{-1} \mathbf{m}'_s = \mathbf{S}^{-1} [\mathbf{G}']^T \mathbf{d}' \quad (8.131)$$

Premultiplying both sides by \mathbf{S} yields

$$\mathbf{m}'_s = [\mathbf{G}']^T \mathbf{d}' \quad (8.132)$$

That is, the stochastic inverse in the primed coordinate system is simply the transpose of \mathbf{G} in the primed coordinate system. Once you have found \mathbf{m}'_s , you can transform back to the original coordinates to obtain the stochastic solution as

$$\mathbf{m}_s = \mathbf{S}^{-1} \mathbf{m}'_s \quad (8.133)$$

The stochastic inverse minimizes the sum of the weighted model parameter vector and the weighted data misfit. That is, the quantity

$$\mathbf{m}^T [\text{cov } \mathbf{m}]^{-1} \mathbf{m} + [\mathbf{d} - \hat{\mathbf{d}}]^T [\text{cov } \mathbf{d}]^{-1} [\mathbf{d} - \hat{\mathbf{d}}] \quad (8.134)$$

is minimized. The generalized inverse, or maximum likelihood, minimizes both individually but not the sum.

It is important to realize that the transformations introduced in Equations (8.118)–(8.121), while of the same form and nomenclature as those introduced in the weighted generalized inverse case in Equations (8.17) and (8.23), differ in an important aspect. Namely, as mentioned after Equation (8.110), $[\text{cov } \mathbf{d}]$ is now a derived quantity, given by Equation (8.114):

$$[\text{cov } \mathbf{d}] = \mathbf{G} [\text{cov } \mathbf{m}] \mathbf{G}^T + [\text{cov } \mathbf{n}] \quad (8.114)$$

The data covariance matrix $[\text{cov } \mathbf{d}]$ is only equal to the noise covariance matrix $[\text{cov } \mathbf{n}]$ if you assume that the noise, or errors, in \mathbf{m} are exactly zero. Thus, before doing a stochastic inverse analysis and the transformations given in Equations (8.118)–(8.121), $[\text{cov } \mathbf{d}]$ must be constructed from the noise covariance matrix $[\text{cov } \mathbf{n}]$ and the mapping of model parameter uncertainties in $[\text{cov } \mathbf{m}]$ as shown in Equation (8.114).

8.3.3 Damped Least Squares

We are now ready to see how this applies to damped least squares. Suppose

$$[\text{cov } \mathbf{m}] = \sigma_m^2 \mathbf{I}_M \quad (8.135)$$

and

$$[\text{cov } \mathbf{n}] = \sigma_n^2 \mathbf{I}_N \quad (8.136)$$

Define a damping term ε^2 as

$$\varepsilon^2 = \sigma_n^2 / \sigma_m^2 \quad (8.137)$$

The stochastic inverse operator, from Equation (8.116), becomes

$$\mathbf{G}_s^{-1} = \mathbf{G}^T [\mathbf{G}\mathbf{G}^T + \varepsilon^2 \mathbf{I}_N]^{-1} \quad (8.138)$$

To determine the effect of adding the ε^2 term, consider the following

$$\mathbf{G}\mathbf{G}^T = \mathbf{U}_P \Lambda_P^2 \mathbf{U}_P^T \quad (7.43)$$

$[\mathbf{G}\mathbf{G}^T]^{-1}$ exists only when $P = N$, and is given by

$$[\mathbf{G}\mathbf{G}^T]^{-1} = \mathbf{U}_P \Lambda_P^{-2} \mathbf{U}_P^T \quad P = N \quad (7.44)$$

we can therefore write $\mathbf{G}\mathbf{G}^T + \varepsilon^2 \mathbf{I}$ as

$$\mathbf{G}\mathbf{G}^T + \varepsilon^2 \mathbf{I} = \begin{bmatrix} \mathbf{U}_P & \mathbf{U}_0 \end{bmatrix} \begin{bmatrix} \Lambda_P^2 + \varepsilon^2 \mathbf{I}_P & 0 \\ 0 & \varepsilon^2 \mathbf{I}_{N-P} \end{bmatrix} \begin{bmatrix} \mathbf{U}_P^T \\ \mathbf{U}_0^T \end{bmatrix} \quad (8.139)$$

$N \times N \quad N \times N \quad N \times N \quad N \times N$

Thus

$$[\mathbf{G}\mathbf{G}^T + \varepsilon^2 \mathbf{I}]^{-1} = \begin{bmatrix} \mathbf{U}_P & \mathbf{U}_0 \end{bmatrix} \begin{bmatrix} [\Lambda_P^2 + \varepsilon^2 \mathbf{I}_P]^{-1} & 0 \\ 0 & \varepsilon^{-2} \mathbf{I}_{N-P} \end{bmatrix} \begin{bmatrix} \mathbf{U}_P^T \\ \mathbf{U}_0^T \end{bmatrix} \quad (8.140)$$

Explicitly multiplying Equation (8.140) out gives

$$[\mathbf{G}\mathbf{G}^T + \varepsilon^2 \mathbf{I}]^{-1} = \mathbf{U}_P [\Lambda_P^2 + \varepsilon^2 \mathbf{I}_P]^{-1} \mathbf{U}_P^T + \mathbf{U}_0 [\varepsilon^{-2} \mathbf{I}_{N-P}] \mathbf{U}_0^T \quad (8.141)$$

Next, we write out Equation (8.138), using singular-value decomposition, as

$$\begin{aligned} \mathbf{G}_s^{-1} &= \mathbf{G}^T [\mathbf{G}\mathbf{G}^T + \varepsilon^2 \mathbf{I}_N]^{-1} \\ &= \{ \mathbf{V}_P \Lambda_P \mathbf{U}_P^T \} \{ \mathbf{U}_P [\Lambda_P^2 + \varepsilon^2 \mathbf{I}_P]^{-1} \mathbf{U}_P^T + \mathbf{U}_0 [\varepsilon^{-2} \mathbf{I}_{N-P}] \mathbf{U}_0^T \} \\ &= \mathbf{V}_P \frac{\Lambda_P}{\Lambda_P^2 + \varepsilon^2 \mathbf{I}_P} \mathbf{U}_P^T \end{aligned} \quad (8.142)$$

since $\mathbf{U}_P^T \mathbf{U}_0 = 0$.

Note the similarity between the stochastic inverse in Equation (8.142) and the generalized inverse

$$\mathbf{G}_g^{-1} = \mathbf{V}_P \Lambda_P^{-1} \mathbf{U}_P^T \quad (7.8)$$

The net effect of the stochastic inverse is to suppress the contributions of eigenvectors with singular values less than ε . To see this, let us write out $\Lambda_P / (\Lambda_P^2 + \varepsilon^2 \mathbf{I}_P)$ explicitly:

$$\frac{\Lambda_P}{\Lambda_P^2 + \varepsilon^2 \mathbf{I}_P} = \begin{bmatrix} \frac{\lambda_1}{\lambda_1^2 + \varepsilon^2} & 0 & \cdots & 0 \\ 0 & \frac{\lambda_2}{\lambda_2^2 + \varepsilon^2} & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \frac{\lambda_p}{\lambda_p^2 + \varepsilon^2} \end{bmatrix} \quad (8.143)$$

If $\lambda_i \gg \varepsilon$, then $\lambda_i / (\lambda_i^2 + \varepsilon^2) \rightarrow \lambda_i^{-1}$, the same as the generalized inverse. If $\lambda_i \ll \varepsilon$, then $\lambda_i / (\lambda_i^2 + \varepsilon^2) \rightarrow \lambda_i / \varepsilon^2 \rightarrow 0$. The stochastic inverse, then, dampens the contributions of eigenvectors associated with small singular values.

The stochastic inverse in Equation (8.138) looks similar to the minimum length inverse

$$\mathbf{G}_{ML}^{-1} = \mathbf{G}^T [\mathbf{G}\mathbf{G}^T]^{-1} \quad (3.75)$$

To see why the stochastic inverse is also called damped least squares, consider the following:

$$\begin{aligned} [\mathbf{G}^T \mathbf{G} + \varepsilon^2 \mathbf{I}_M]^{-1} \mathbf{G}^T &= \{\mathbf{V}_P [\Lambda_P^2 + \varepsilon^2 \mathbf{I}_P]^{-1} \mathbf{V}_P^T + \varepsilon^2 \mathbf{V}_0 \mathbf{V}_0^T\} \{\mathbf{V}_P \Lambda_P \mathbf{U}_P^T\} \\ &= \{\mathbf{V} [\Lambda_P^2 + \varepsilon^2 \mathbf{I}_P]^{-1}\} \{\mathbf{V}_P^T \mathbf{V}_P \Lambda_P \mathbf{U}_P^T\} + \varepsilon^2 \mathbf{V}_0 \mathbf{V}_0^T \mathbf{V}_P \Lambda_P \mathbf{U}_P^T \\ &= \mathbf{V}_P \frac{\Lambda_P}{\Lambda_P^2 + \varepsilon^2 \mathbf{I}_P} \mathbf{U}_P^T \\ &= \mathbf{G}^T [\mathbf{G}\mathbf{G}^T + \varepsilon^2 \mathbf{I}_N]^{-1} \end{aligned} \quad (8.144)$$

Thus

$$[\mathbf{G}^T \mathbf{G} + \varepsilon^2 \mathbf{I}_M]^{-1} \mathbf{G}^T = \mathbf{G}^T [\mathbf{G}\mathbf{G}^T + \varepsilon^2 \mathbf{I}_N]^{-1} \quad (8.145)$$

The choice of σ_m^2 is often arbitrary. Thus, ε^2 is often chosen arbitrarily to stabilize the problem. Solutions are obtained for a variety of ε^2 , and a final choice is made based on the *a posteriori* model covariance matrix.

The stability gained with damped least squares is not obtained without loss elsewhere. Specifically, resolution degrades with increased damping. To see this, consider the model resolution matrix for the stochastic inverse:

$$\begin{aligned}\mathbf{R} &= \mathbf{G}_s^{-1}\mathbf{G} \\ &= \mathbf{V}_P \frac{\Lambda_P^2}{\Lambda_P^2 + \varepsilon^2 \mathbf{I}_P} \mathbf{V}_P^T\end{aligned}\quad (8.146)$$

It is easy to see that the stochastic inverse model resolution matrix reduces to the generalized inverse case when ε^2 goes to 0, as expected.

The reduction in model resolution can be seen by considering the trace of \mathbf{R} :

$$\text{trace}(\mathbf{R}) = \sum_{i=1}^P \frac{\lambda_i^2}{\lambda_i^2 + \varepsilon^2} \leq P \quad (8.147)$$

Similarly, the data resolution matrix \mathbf{N} is given by

$$\begin{aligned}\mathbf{N} &= \mathbf{G}\mathbf{G}_s^{-1} \\ &= \mathbf{U}_P \frac{\Lambda_P^2}{\Lambda_P^2 + \varepsilon^2 \mathbf{I}_P} \mathbf{U}_P^T\end{aligned}\quad (8.148)$$

$$\text{trace}(\mathbf{N}) = \sum_{i=1}^P \frac{\lambda_i^2}{\lambda_i^2 + \varepsilon^2} \leq P \quad (8.149)$$

Finally, consider the unit model covariance matrix $[\text{cov}_u \mathbf{m}]$, given by

$$\begin{aligned}[\text{cov}_u \mathbf{m}] &= \mathbf{G}_s^{-1}[\mathbf{G}_s^{-1}]^T \\ &= \mathbf{V}_P \frac{\Lambda_P^2}{[\Lambda_P^2 + \varepsilon^2 \mathbf{I}_P]^2} \mathbf{V}_P^T\end{aligned}\quad (8.150)$$

which reduces to the generalized inverse case when $\varepsilon^2 = 0$. The introduction of ε^2 reduces the size of the covariance terms, a reflection of the stability added by including a damping term.

An alternative approach to damped least squares is achieved by adding equations of the form

$$\varepsilon m_i = 0 \quad i = 1, 2, \dots, M \quad (8.151)$$

to the original set of equations

$$\mathbf{G}\mathbf{m} = \mathbf{d} \quad (1.13)$$

The combined set of equations can be written in partitioned form as

$$\begin{matrix} \begin{bmatrix} \mathbf{G} \\ \varepsilon \mathbf{I}_M \end{bmatrix} \mathbf{m} & = & \begin{bmatrix} \mathbf{d} \\ \mathbf{0} \end{bmatrix} \\ (N+M) \times M & & (N+M) \times 1 \end{matrix} \quad (8.152)$$

The least squares solution to Equation (8.152) is given by

$$\begin{aligned} \mathbf{m} &= \left\{ \begin{bmatrix} \mathbf{G}^T & \varepsilon \mathbf{I}_M \end{bmatrix} \begin{bmatrix} \mathbf{G} \\ \varepsilon \mathbf{I}_M \end{bmatrix} \right\}^{-1} \begin{bmatrix} \mathbf{G}^T & \varepsilon \mathbf{I}_M \end{bmatrix} \begin{bmatrix} \mathbf{d} \\ \mathbf{0} \end{bmatrix} \\ &= [\mathbf{G}^T \mathbf{G} + \varepsilon^2 \mathbf{I}_M]^{-1} \mathbf{G}^T \mathbf{d} \end{aligned} \quad (8.153)$$

The addition of $\varepsilon^2 \mathbf{I}_M$ insures a least squares solution because $\mathbf{G}^T \mathbf{G} + \varepsilon^2 \mathbf{I}_M$ will have no eigenvalues less than ε^2 , and hence is invertible.

In signal processing, the addition of ε^2 is equivalent to adding white noise to the signal. Consider transforming

$$[\mathbf{G}^T \mathbf{G} + \varepsilon^2 \mathbf{I}_M] \mathbf{m} = \mathbf{G}^T \mathbf{d} \quad (8.154)$$

into the frequency domain as

$$[F_i^*(\omega) F_i(\omega) + \varepsilon^2] M(\omega) = F_i^*(\omega) F_o(\omega) \quad (8.155)$$

where $F_i(\omega)$ is the Fourier transform of the input waveform to some filter, * represents complex conjugate, $F_o(\omega)$ is the Fourier transform of the output wave form from the filter, $M(\omega)$ is the Fourier transform of the impulse response of the filter, and ε^2 is a constant for all frequencies ω . Solving for \mathbf{m} as the inverse Fourier transform of Equation (8.155) gives

$$\mathbf{m} = \text{F.T.}^{-1} \left[\frac{F_i^*(\omega) F_o(\omega)}{F_i^*(\omega) F_i(\omega) + \varepsilon^2} \right] \quad (8.156)$$

The addition of ε^2 in the denominator assures that the solution is not dominated by small values of $F_i(\omega)$, which can arise when the signal-to-noise ratio is poor. Because the ε^2 term is added equally at all frequencies, this is equivalent to adding white light to the signal.

Damping is particularly useful in nonlinear problems. In nonlinear problems, small singular values can produce very large changes, or steps, during the iterative process. These large steps can easily violate the assumption of linearity in the region where the nonlinear problem was linearized. In order to limit step sizes, an ε^2 term can be added. Typically, one uses a fairly large value of ε^2 during the initial phase of the iterative procedure, gradually letting ε^2 go to zero as the solution is approached.

Recall that the generalized inverse minimized $[\mathbf{d} - \mathbf{G}\mathbf{m}]^T[\mathbf{d} - \mathbf{G}\mathbf{m}]$ and $\mathbf{m}^T\mathbf{m}$ individually. Consider a new function E to minimize, defined by

$$\begin{aligned} E &= [\mathbf{d} - \mathbf{G}\mathbf{m}]^T[\mathbf{d} - \mathbf{G}\mathbf{m}] + \varepsilon^2\mathbf{m}^T\mathbf{m} \\ &= \mathbf{m}^T\mathbf{G}^T\mathbf{G}\mathbf{m} - \mathbf{m}^T\mathbf{G}^T\mathbf{d} - \mathbf{d}^T\mathbf{G}\mathbf{m} + \mathbf{d}^T\mathbf{d} + \varepsilon^2\mathbf{m}^T\mathbf{m} \end{aligned} \quad (8.157)$$

Differentiating E with respect to \mathbf{m}^T and setting it equal to zero yields

$$\partial E/\partial\mathbf{m}^T = \mathbf{G}^T\mathbf{G}\mathbf{m} - \mathbf{G}^T\mathbf{d} + \varepsilon^2\mathbf{m} = 0 \quad (8.158)$$

or

$$[\mathbf{G}^T\mathbf{G} + \varepsilon^2\mathbf{I}_M]\mathbf{m} = \mathbf{G}^T\mathbf{d} \quad (8.159)$$

This shows why damped least squares minimized a weighted sum of the misfit and the length of the model parameter vector.

8.4 Ridge Regression

8.4.1 Mathematical Background

Recall the least squares operator

$$[\mathbf{G}^T\mathbf{G}]^{-1}\mathbf{G}^T \quad (8.160)$$

If the data covariance matrix $[\text{cov } \mathbf{d}]$ is given by

$$[\text{cov } \mathbf{d}] = \sigma^2\mathbf{I} \quad (8.161)$$

then the *a posteriori* model covariance matrix $[\text{cov } \mathbf{m}]$, also called the dispersion of \mathbf{m} , is given by

$$[\text{cov } \mathbf{m}] = \sigma^2[\mathbf{G}^T\mathbf{G}]^{-1} \quad (8.162)$$

In terms of singular-value decomposition, it is given by

$$[\text{cov } \mathbf{m}] = \sigma^2\mathbf{V}_p\mathbf{\Lambda}_p^{-2}\mathbf{V}_p^T \quad (8.163)$$

This can also be written as

$$[\text{cov } \mathbf{m}] = \sigma^2[\mathbf{V}_p|\mathbf{V}_0] \begin{bmatrix} \mathbf{\Lambda}_p^{-2} \\ \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{V}_p^T \\ \mathbf{V}_0^T \end{bmatrix} \quad (8.164)$$

The total variance is defined as the trace of the model covariance matrix, given by

$$\text{trace}[\text{cov } \mathbf{m}] = \sigma^2 \{\text{trace}[\mathbf{G}^T \mathbf{G}]^{-1}\} = \sigma^2 \sum_{i=1}^P \frac{1}{\lambda_i^2} \quad (8.165)$$

which follows from the fact that the trace of a matrix is invariant under an orthogonal coordinate transformation.

It is clear from Equation (8.165) that the total variance will get large as λ_i gets small. We saw that the stochastic inverse operator

$$\mathbf{G}_s^{-1} = [\mathbf{G}^T \mathbf{G} + \varepsilon^2 \mathbf{I}_M]^{-1} \mathbf{G}^T = \mathbf{G}^T [\mathbf{G} \mathbf{G}^T + \varepsilon^2 \mathbf{I}_N]^{-1} \quad (8.145)$$

resulted in a reduction of the model covariance (8.107). In fact, the addition of ε^2 to each diagonal entry $\mathbf{G}^T \mathbf{G}$ results in a total variance defined by

$$\text{trace}[\text{cov } \mathbf{m}] = \sigma^2 \{\text{trace}[\mathbf{G}^T \mathbf{G} + \varepsilon^2 \mathbf{I}]^{-1}\} = \sigma^2 \sum_{i=1}^P \frac{\lambda_i^2}{(\lambda_i^2 + \varepsilon^2)^2} \quad (8.166)$$

Clearly, Equation (8.166) is less than (8.165) for all $\varepsilon^2 > 0$.

8.4.2 The Ridge Regression Operator

The stochastic inverse operator of Equation (8.145) is also called ridge regression for reasons that I will explain shortly. The ridge regression operator is derived as follows. We seek an operator that finds a solution \mathbf{m}_{RR} that is closest to the origin (as in the minimum length case), subject to the constraint that the solution lie on an ellipsoid defined by

$$\begin{matrix} [\mathbf{m}_{RR} - \mathbf{m}_{LS}]^T & \mathbf{G}^T \mathbf{G} & [\mathbf{m}_{RR} - \mathbf{m}_{LS}] & = & \phi_0 \\ 1 \times M & M \times M & M \times 1 & & 1 \times 1 \end{matrix} \quad (8.167)$$

where \mathbf{m}_{LS} is the least squares solution (i.e., obtained by setting ε^2 equal to 0). Equation (8.167) represents a single-equation quadratic in \mathbf{m}_{RR} .

The ridge regression operator \mathbf{G}_{RR}^{-1} is obtained using Lagrange multipliers. We form the function

$$\Psi(\mathbf{m})_{RR} = \mathbf{m}_{RR}^T \mathbf{m}_{RR} + \lambda \{ [\mathbf{m}_{RR} - \mathbf{m}_{LS}]^T \mathbf{G}^T \mathbf{G} [\mathbf{m}_{RR} - \mathbf{m}_{LS}] - \phi_0 \} \quad (8.168)$$

and differentiate with respect to \mathbf{m}_{RR}^T to obtain

$$\mathbf{m}_{RR} + \lambda \mathbf{G}^T \mathbf{G} [\mathbf{m}_{RR} - \mathbf{m}_{LS}] = 0 \quad (8.169)$$

Solving Equation (8.169) for \mathbf{m}_{RR} gives

$$[\lambda \mathbf{G}^T \mathbf{G} + \mathbf{I}_M] \mathbf{m}_{RR} = \lambda \mathbf{G}^T \mathbf{G} \mathbf{m}_{LS}$$

or

$$\mathbf{m}_{RR} = [\lambda \mathbf{G}^T \mathbf{G} + \mathbf{I}_M]^{-1} \lambda \mathbf{G}^T \mathbf{G} \mathbf{m}_{LS} \quad (8.170)$$

The least squares solution \mathbf{m}_{LS} is given by

$$\mathbf{m}_{LS} = [\mathbf{G}^T \mathbf{G}]^{-1} \mathbf{G}^T \mathbf{d} \quad (3.31)$$

Substituting \mathbf{m}_{LS} from Equation (3.31) into (8.170)

$$\begin{aligned} \mathbf{m}_{RR} &= [\lambda \mathbf{G}^T \mathbf{G} + \mathbf{I}_M]^{-1} \lambda \mathbf{G}^T \mathbf{G} [\mathbf{G}^T \mathbf{G}]^{-1} \mathbf{G}^T \mathbf{d} \\ &= [\lambda \mathbf{G}^T \mathbf{G} + \mathbf{I}_M]^{-1} \lambda \mathbf{G}^T \mathbf{d} \\ &= \frac{1}{\lambda} \left[\mathbf{G}^T \mathbf{G} + \frac{1}{\lambda} \mathbf{I}_M \right]^{-1} \lambda \mathbf{G}^T \mathbf{d} \quad \lambda \neq 0 \\ &= \left[\mathbf{G}^T \mathbf{G} + \frac{1}{\lambda} \mathbf{I}_M \right]^{-1} \mathbf{G}^T \mathbf{d} \end{aligned} \quad (8.171)$$

If we let $1/\lambda = \varepsilon^2$, then Equation (8.171) becomes

$$\mathbf{m}_{RR} = [\mathbf{G}^T \mathbf{G} + \varepsilon^2 \mathbf{I}_M]^{-1} \mathbf{G}^T \mathbf{d} \quad (8.172)$$

and the ridge regression operator \mathbf{G}_{RR}^{-1} is defined as

$$\mathbf{G}_{RR}^{-1} = [\mathbf{G}^T \mathbf{G} + \varepsilon^2 \mathbf{I}_M]^{-1} \mathbf{G}^T \quad (8.173)$$

In terms of singular-value decomposition, the ridge regression operator \mathbf{G}_{RR}^{-1} is identical to the stochastic inverse operator, and following Equation (8.142),

$$\mathbf{G}_{RR}^{-1} = \mathbf{V}_P \frac{\Lambda_P}{\Lambda_P^2 + \varepsilon^2 \mathbf{I}_P} \mathbf{U}_P^T \quad (8.174)$$

In practice, we determine ε^2 (and thus λ) by trial and error, with the attendant trade-off between resolution and stability. As defined, however, every choice of ε^2 is associated with a particular ϕ_0 and hence a particular ellipsoid from Equation (8.167). Changing ϕ_0 does not change the orientation of the ellipsoid; it simply stretches or contracts the major and minor axes. We can think of the family of ellipsoids defined by varying ε^2 (or ϕ_0) as a ridge in solution space, with

each particular ε^2 (or ϕ_0) being a contour of the ridge. We then obtain the ridge regression solution by following one of the contours around the ellipsoid until we find the point closest to the origin, hence the name ridge regression.

8.4.3 An Example of Ridge Regression Analysis

A simple example will help clarify the ridge regression operator. Consider the following:

$$\begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} m_1 \\ m_2 \end{bmatrix} = \begin{bmatrix} 8 \\ 4 \end{bmatrix} \quad (8.175)$$

G m d

Singular-value decomposition gives

$$\mathbf{U}_p = \mathbf{U} = \mathbf{I}_2 \quad (8.176)$$

$$\mathbf{V}_p = \mathbf{V} = \mathbf{I}_2 \quad (8.177)$$

$$\Lambda_p = \Lambda = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix} \quad (8.178)$$

The generalized inverse \mathbf{G}_g^{-1} is given by

$$\begin{aligned} \mathbf{G}_g^{-1} &= \mathbf{V}_p \Lambda_p^{-1} \mathbf{U}_p^T \\ &= \mathbf{I}_2 \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & 1 \end{bmatrix} \mathbf{I}_2^T \\ &= \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & 1 \end{bmatrix} \end{aligned} \quad (8.179)$$

The generalized inverse solution (also the exact, or least squares, solution) is

$$\mathbf{m}_{LS} = \mathbf{G}_g^{-1} \mathbf{d} = \begin{bmatrix} 4 \\ 4 \end{bmatrix} \quad (8.180)$$

The ridge regression solution is given by

$$\begin{aligned}
 \mathbf{m}_{RR} &= \mathbf{G}_{RR}^{-1} \mathbf{d} = \mathbf{V}_P \frac{\Lambda_P}{\Lambda_P^2 + \varepsilon^2 \mathbf{I}_P} \mathbf{U}_P^T \begin{bmatrix} 8 \\ 4 \end{bmatrix} \\
 &= \mathbf{I}_2 \begin{bmatrix} \frac{2}{4 + \varepsilon^2} & 0 \\ 0 & \frac{1}{1 + \varepsilon^2} \end{bmatrix} \mathbf{I}_2 \begin{bmatrix} 8 \\ 4 \end{bmatrix} \\
 &= \begin{bmatrix} \frac{2}{4 + \varepsilon^2} & 0 \\ 0 & \frac{1}{1 + \varepsilon^2} \end{bmatrix} \begin{bmatrix} 8 \\ 4 \end{bmatrix}
 \end{aligned} \tag{8.181}$$

Note that for $\varepsilon^2 = 0$, the least squares solution is recovered. Also, as $\varepsilon^2 \rightarrow \infty$, the solution goes to the origin. Thus, as expected, the solution varies from the least squares solution to the origin as more and more weight is given to minimizing the length of the solution vector.

We can now determine the ellipsoid associated with a particular value of ε^2 . For example, let $\varepsilon^2 = 1$. Then the ridge regression solution, from Equation (8.181), is

$$\begin{bmatrix} m_1 \\ m_2 \end{bmatrix}_{RR} = \begin{bmatrix} \frac{16}{4 + \varepsilon^2} \\ \frac{4}{1 + \varepsilon^2} \end{bmatrix} = \begin{bmatrix} 3.2 \\ 2 \end{bmatrix} \tag{8.182}$$

Now, returning to the constraint Equation (8.167), we have that

$$\begin{bmatrix} m_1 - 4.0 \\ m_2 - 4.0 \end{bmatrix}^T \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} m_1 - 4.0 \\ m_2 - 4.0 \end{bmatrix} = \phi_0$$

or

$$4(m_1 - 4.0)^2 + (m_2 - 4.0)^2 = \phi_0 \tag{8.183}$$

To find ϕ_0 , we substitute the solution from Equation (8.182) into (8.167) and

$$4(3.2 - 4.0)^2 + (2.0 - 4.0)^2 = \phi_0 \tag{8.184}$$

or

$$\phi_0 = 6.56 \tag{8.185}$$

Substituting ϕ_0 from Equation (8.185) back into (8.183) and rearranging gives

$$\frac{(m_1 - 4.0)^2}{1.64} + \frac{(m_2 - 4.0)^2}{6.56} = 1.0 \tag{8.186}$$

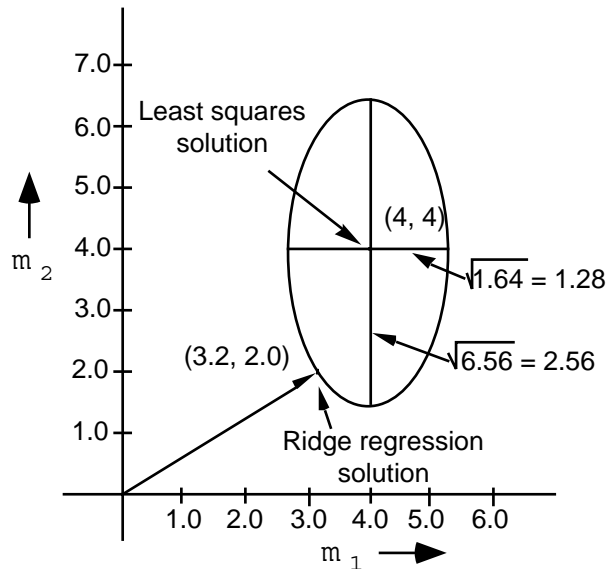
Equation (8.186) is of the form

$$\frac{(x-h)^2}{b^2} + \frac{(y-k)^2}{a^2} = 1.0 \quad (8.187)$$

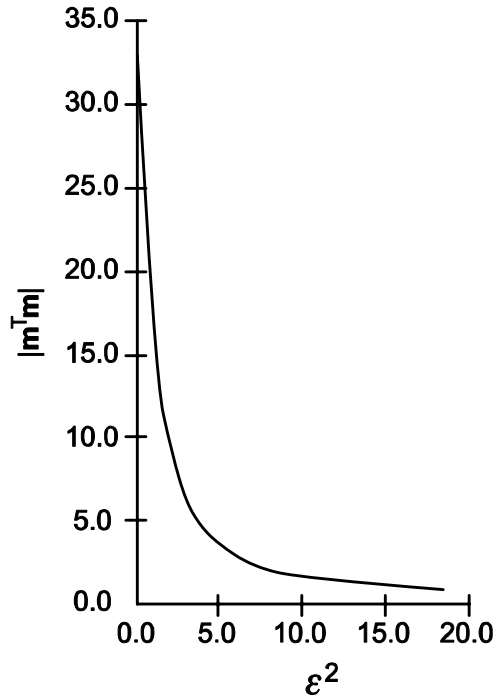
which represents an ellipse centered at (h, k) , with semimajor and semiminor axes a and b parallel to the y and x axes, respectively. Thus, for the current example, the lengths of the semimajor and semiminor axes are 2.56 and 1.28, respectively. The axes of the ellipse are parallel to the m_2 and m_1 axes, and the ellipse is centered at $(4, 4)$. Different choices for ϵ^2 will produce a family of ellipses centered on $(4, 4)$, with semimajor and semiminor axes parallel to the m_2 and m_1 axes, respectively, and with the semimajor axis always twice the length of the semiminor axis.

The shape and orientation of the family of ellipses follow completely from the structure of the original \mathbf{G} matrix. The axes of the ellipse coincide with the m_1 and m_2 axes because the original \mathbf{G} matrix was diagonal. If the original \mathbf{G} matrix had not been diagonal, the axes of the ellipse would have been inclined to the m_1 and m_2 axes. The center of the ellipse, given by the least squares solution, is, of course, both a function of \mathbf{G} and the data vector \mathbf{d} .

The graph below illustrates this particular problem for $\epsilon^2 = 1$.

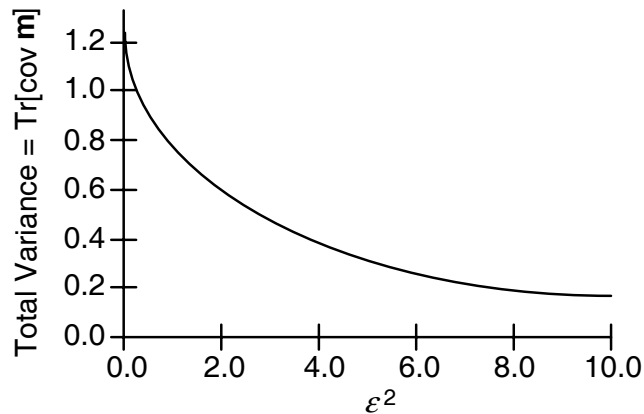


It is also instructive to plot the length squared of the solution, $\mathbf{m}^T\mathbf{m}$, as a function of ϵ^2 :



This figure shows that adding ε^2 damps the solution from least squares toward zero length as ε^2 increases.

Next consider a plot of the total variance from Equation (8.166) as a function of ε^2 for data variance $\sigma^2 = 1$.

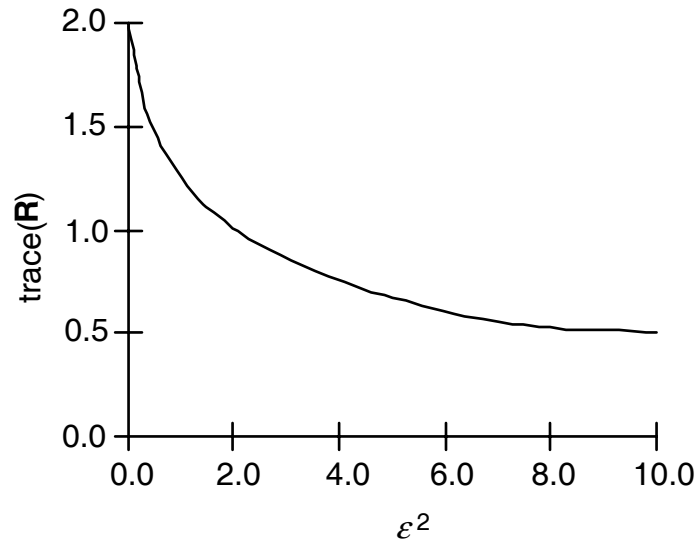


The total variance decreases, as expected, as more damping is included.

Finally, consider the model resolution matrix \mathbf{R} given by

$$\begin{aligned} \mathbf{R} &= \mathbf{G}_{\text{RR}}^{-1} \mathbf{G} \\ &= \mathbf{V}_P \frac{\Lambda_P^2}{\Lambda_P^2 + \varepsilon^2 \mathbf{I}_P} \mathbf{V}_P^T \end{aligned} \tag{8.188}$$

We can plot trace (\mathbf{R}) as a function of ε^2 and get



For $\varepsilon^2 = 0$, we have perfect model resolution, with trace (\mathbf{R}) = $P = 2 = M = N$. As ε^2 increases, the model resolution decreases. Comparing the plots of total variance and the trace of the model resolution matrix, we see that as ε^2 increases, stability improves (total variance decreases) while resolution degrades. This is an inevitable trade-off.

In this particular simple example, it is hard to choose the most appropriate value for ε^2 because, in fact, the sizes of the two singular values differ very little. In general, when the singular values differ greatly, the plots for total variance and trace (\mathbf{R}) can help us choose ε^2 . If the total variance initially diminishes rapidly and then very slowly for increasing ε^2 , choosing ε^2 near the bend in the total variance curve is most appropriate.

We have shown in this section how the ridge regression operator is formed and how it is equivalent to damped least squares and the stochastic inverse operator.

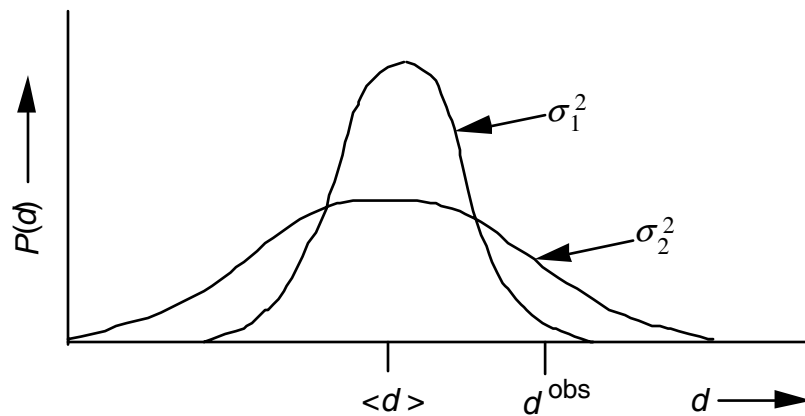
8.5 Maximum Likelihood

8.5.1 Background

The maximum likelihood approach is fundamentally probabilistic in nature. A probability density function (PDF) is created in data space that assigns a probability $P(\mathbf{d})$ to every point in data space. This PDF is a function of the model parameters, and hence $P(\mathbf{d})$ may change with each choice of \mathbf{m} . The underlying principle of the maximum likelihood approach is to find a solution \mathbf{m}_{MX} such that $P(\mathbf{d})$ is maximized at the observed data \mathbf{d}^{obs} . Put another way, a solution \mathbf{m}_{MX} is sought such that the probability of observing the observed data is maximized. At first thought, this may not seem very satisfying. After all, in some sense there is a 100% chance that the observed data are observed, simply because they are the observed data. The point

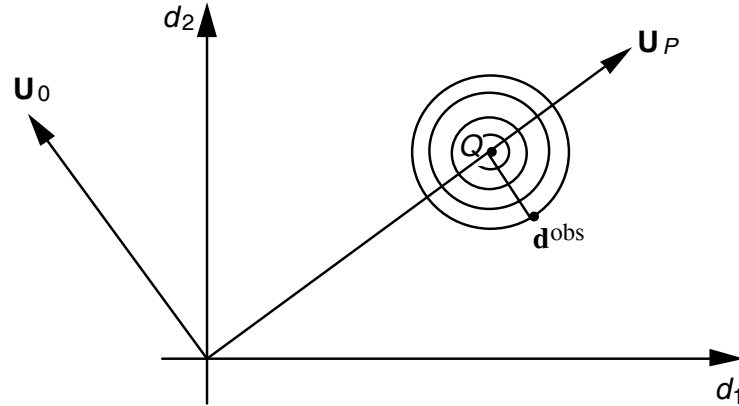
is, however, that $P(\mathbf{d})$ is a calculated quantity, which varies over data space as a function of \mathbf{m} . Put this way, does it make sense to choose \mathbf{m} such that $P(\mathbf{d}^{\text{obs}})$ is small, meaning that the observed data are an unlikely outcome of some experiment? This is clearly not ideal. Rather, it makes more sense to choose \mathbf{m} such that $P(\mathbf{d}^{\text{obs}})$ is as large as possible, meaning that you have found an \mathbf{m} for which the observed data, which exist with 100% certainty, are as likely an outcome as possible.

Imagine a very simple example with a single observation where $P(d)$ is Gaussian with fixed mean $\langle d \rangle$ and a variance σ^2 that is a function of some model parameter m . For the moment we need not worry about how m affects σ^2 , other than to realize that as m changes, so does σ^2 . Consider the diagram below, where the vertical axis is probability, and the horizontal axis is d . Shown on the diagram are d^{obs} , the observed datum; $\langle d \rangle$, the mean value for the Gaussian $P(d)$; and two different $P(d)$ curves based on two different variance estimates σ_1^2 and σ_2^2 , respectively.



The area under both $P(d)$ curves is equal to one, since this represents integrating $P(d)$ over all possible data values. The curve for σ_1^2 , where σ_1^2 is small, is sharply peaked at $\langle d \rangle$, but is very small at d^{obs} . In fact, d^{obs} appears to be several standard deviations σ from $\langle d \rangle$, indicating that d^{obs} is a very unlikely outcome. $P(d)$ for σ_2^2 , on the other hand, is not as sharply peaked at $\langle d \rangle$, but because the variance is larger, $P(d)$ is larger at the observed datum, d^{obs} . You could imagine letting σ^2 get very large, in which case values far from $\langle d \rangle$ would have $P(d)$ larger than zero, but no value of $P(d)$ would be very large. In fact, you could imagine $P(d^{\text{obs}})$ becoming smaller than the case for σ_2^2 . Thus, the object would be to vary m , and hence σ^2 , such that $P(d^{\text{obs}})$ is maximized. Of course, for this simple example we have not worried about the mechanics of finding m , but we will later for more realistic cases.

A second example, after one near the beginning of Menke's Chapter 5, is also very illustrative. Imagine collecting a single datum N times in the presence of Gaussian noise. The observed data vector \mathbf{d}^{obs} has N entries and hence lies in an N -dimensional data space. You can think of each observation as a random variable with the same mean $\langle d \rangle$ and variance σ^2 , both of which are unknown. The goal is to find $\langle d \rangle$ and σ^2 . We can cast this problem in our familiar $\mathbf{Gm} = \mathbf{d}$ form by associating \mathbf{m} with $\langle d \rangle$ and noting that $\mathbf{G} = (1/N)[1, 1, \dots, 1]^T$. Consider the simple case where $N = 2$, shown on the next page:



The observed data \mathbf{d}^{obs} are a point in the d_1d_2 plane. If we do singular-value decomposition on \mathbf{G} , we see immediately that, in general, $\mathbf{U}_P = (1/\sqrt{N})[1, 1, \dots, 1]^T$, and for our $N=2$ case, $\mathbf{U}_P = [1/\sqrt{2}, 1/\sqrt{2}]^T$, and $\mathbf{U}_0 = [-1/\sqrt{2}, 1/\sqrt{2}]^T$. We recognize that all predicted data must lie in \mathbf{U}_P space, which is a single vector. Every choice of $\mathbf{m} = \langle d \rangle$ gives a point on the line $d_1 = d_2 = \dots = d_N$. If we slide $\langle d \rangle$ up to the point Q on the diagram, we see that all the misfit lies in \mathbf{U}_0 space, and we have obtained the least squares solution for $\langle d \rangle$. Also shown on the figure are contours of $P(\mathbf{d})$ based on σ^2 . If σ^2 is small, the contours will be close together, and $P(\mathbf{d}^{\text{obs}})$ will be small. The contours are circular because the variance is the same for each d_i . Our $N=2$ case has thus reduced to the one-dimensional case discussed on the previous page, where some value of σ^2 will maximize $P(\mathbf{d}^{\text{obs}})$. Menke (Chapter 5) shows that $P(\mathbf{d})$ for the N -dimensional case with Gaussian noise is given by

$$P(d) = \frac{1}{\sigma^N (2\pi)^{N/2}} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^N (d_i - \langle d \rangle)^2\right] \quad (8.189)$$

where d_i are the observed data and $\langle d \rangle$ and σ are the unknown model parameters. The solution for $\langle d \rangle$ and σ is obtained by maximizing $P(\mathbf{d})$. That is, the partials of $P(\mathbf{d})$ with respect to $\langle d \rangle$ and σ are formed and set to zero. Menke shows that this leads to

$$\langle d \rangle^{\text{est}} = \frac{1}{N} \sum_{i=1}^N d_i \quad (8.190)$$

$$\sigma^{\text{est}} = \left[\frac{1}{N} \sum_{i=1}^N (d_i - \langle d \rangle)^2 \right]^{1/2} \quad (8.191)$$

We see that $\langle d \rangle$ is found independently of σ , and this shows why the least squares solution (point Q on the diagram) seems to be found independently of σ . Now, however, Equation (8.191) indicates that σ^{est} will vary for different choices of $\langle d \rangle$ affecting $P(\mathbf{d}^{\text{obs}})$.

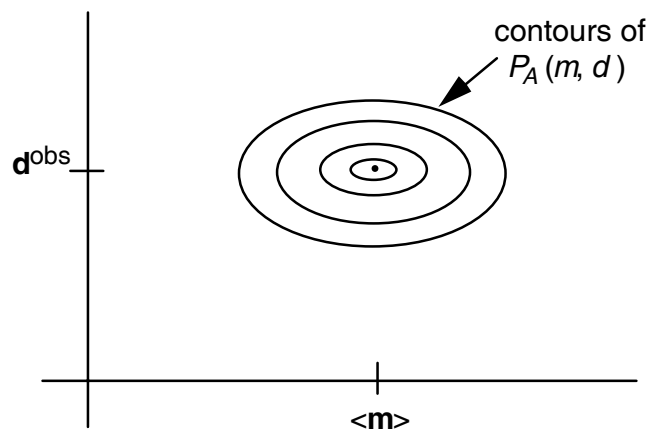
The example can be extended to the general data vector \mathbf{d} case where the Gaussian noise (possibly correlated) in \mathbf{d} is described by the data covariance matrix $[\text{cov } \mathbf{d}]$. Then it is possible to assume that $P(\mathbf{d})$ has the form

$$P(\mathbf{d}) \propto \exp\left\{-\frac{1}{2}[\mathbf{d} - \mathbf{Gm}]^T[\text{cov } \mathbf{d}]^{-1}[\mathbf{d} - \mathbf{Gm}]\right\} \quad (8.192)$$

We note that the exponential in Equation (8.192) reduces to the exponential in Equation (8.189) when $[\text{cov } \mathbf{d}] = \sigma^2\mathbf{I}$, and \mathbf{Gm} gives the predicted data, given by $\langle d \rangle$. $P(\mathbf{d})$ in Equation (8.192) is maximized when $[\mathbf{d} - \mathbf{Gm}]^T[\text{cov } \mathbf{d}]^{-1}[\mathbf{d} - \mathbf{Gm}]$ is minimized. This is, of course, exactly what is minimized in the weighted least squares [Equations (3.89) and (3.90)] and weighted generalized inverse [Equation (8.13)] approaches. We can make the very important conclusion that maximum likelihood approaches are equivalent to weighted least squares or weighted generalized inverse approaches when the noise in the data is Gaussian.

8.5.2 The General Case

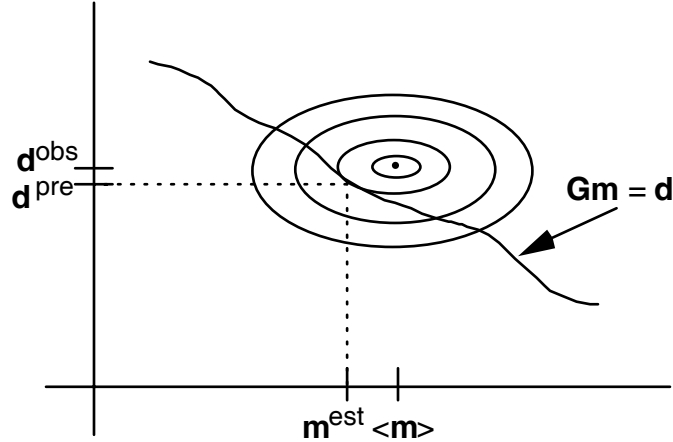
We found in the generalized inverse approach that whenever $P < M$, the solution is nonunique. The equivalent viewpoint with the maximum likelihood approach is that $P(\mathbf{d})$ does not have a well-defined peak. In this case, prior information (such as minimum length for the generalized inverse) must be added. We can think of \mathbf{d}^{obs} and $[\text{cov } \mathbf{d}]$ as prior information for the data, which we could summarize as $P_A(\mathbf{d})$. The prior information about the model parameters could also be summarized as $P_A(\mathbf{m})$ and could take the form of a prior estimate of the solution $\langle m \rangle$ and a covariance matrix $[\text{cov } \mathbf{m}]_A$. Graphically (after Figure 5.9 in Menke) you can represent the joint distribution $P_A(m, d) = P_A(m) P_A(d)$ detailing the prior knowledge of data and model spaces as



where $P_A(m, d)$ is contoured about $(\mathbf{d}^{\text{obs}}, \langle \mathbf{m} \rangle)$, the most likely point in the prior distribution. The contours are not inclined to the model or data axes because we assume that there is no correlation between our prior knowledge of \mathbf{d} and \mathbf{m} . As shown, the figure indicates less confidence in $\langle m \rangle$ than in the data. Of course, if the maximum likelihood approach were applied

to $P_A(m, d)$, it would return $(\mathbf{d}^{\text{obs}}, \langle m \rangle)$ because there has not been any attempt to include the forward problem $\mathbf{G}\mathbf{m} = \mathbf{d}$.

Each choice of \mathbf{m} leads to a predicted data vector \mathbf{d}^{pre} . In the schematic figure below, the forward problem $\mathbf{G}\mathbf{m} = \mathbf{d}$ is thus shown as a line in the model space–data space plane:



The maximum likelihood solution \mathbf{m}^{est} is the point where the $P(\mathbf{d})$ obtains its maximum value along the $\mathbf{G}\mathbf{m} = \mathbf{d}$ curve. If you imagine that $P(\mathbf{d})$ is very elongated along the model-space axis, this is equivalent to saying that the data are known much better than the prior model parameter estimate $\langle \mathbf{m} \rangle$. In this case \mathbf{d}^{pre} will be very close to the observed data \mathbf{d}^{obs} , but the estimated solution \mathbf{m}^{est} may be very far from $\langle \mathbf{m} \rangle$. Conversely, if $P(\mathbf{d})$ is elongated along the data axis, then the data uncertainties are relatively large compared to the confidence in $\langle \mathbf{m} \rangle$, and \mathbf{m}^{est} will be close to $\langle \mathbf{m} \rangle$, while \mathbf{d}^{pre} may be quite different from \mathbf{d}^{obs} .

Menke also points out that there may be uncertainties in the theoretical forward relationship $\mathbf{G}\mathbf{m} = \mathbf{d}$. These may be expressed in terms of an $N \times N$ inexact-theory covariance matrix $[\text{cov } \mathbf{g}]$. This covariance matrix deserves some comment. As in any covariance matrix of a single term (e.g., \mathbf{d} , \mathbf{m} , or \mathbf{G}), the diagonal entries are variances, and the off-diagonal terms are covariances. What does the (1, 1) entry of $[\text{cov } \mathbf{g}]$ refer to, however? It turns out to be the variance of the first equation (row) in \mathbf{G} . Similarly, each diagonal term in $[\text{cov } \mathbf{g}]$ refers to an uncertainty of a particular equation (row) in \mathbf{G} , and off-diagonal terms are covariances between rows in \mathbf{G} . Each row in \mathbf{G} times \mathbf{m} gives a predicted datum. For example, the first row of \mathbf{G} times \mathbf{m} gives d_1^{pre} . Thus a large variance for the (1, 1) term in $[\text{cov } \mathbf{g}]$ would imply that we do not have much confidence in the theory's ability to predict the first observation. It is easy to see that this is equivalent to saying that not much weight should be given to the first observation. We will see, then, that $[\text{cov } \mathbf{g}]$ plays a role similar to $[\text{cov } \mathbf{d}]$.

We are now in a position to give the maximum likelihood operator $\mathbf{G}_{\text{MX}}^{-1}$ in terms of \mathbf{G} , and the data ($[\text{cov } \mathbf{d}]$), model parameter ($[\text{cov } \mathbf{m}]$), and theory ($[\text{cov } \mathbf{g}]$) covariance matrices as

$$\mathbf{G}_{\text{MX}}^{-1} = [\text{cov } \mathbf{m}]^{-1} \mathbf{G}^T \{ [\text{cov } \mathbf{d}] + [\text{cov } \mathbf{g}] + \mathbf{G} [\text{cov } \mathbf{m}]^{-1} \mathbf{G}^T \}^{-1} \quad (8.193)$$

$$= [\mathbf{G}^T \{ [\text{cov } \mathbf{d}] + [\text{cov } \mathbf{g}] \}^{-1} \mathbf{G} + [\text{cov } \mathbf{m}]^{-1}]^{-1} \mathbf{G}^T \{ [\text{cov } \mathbf{d}] + [\text{cov } \mathbf{g}] \}^{-1} \quad (8.194)$$

where Equations (8.193) and (8.194) are equivalent. There are several points to make. First, as mentioned previously, [cov **d**] and [cov **g**] appear everywhere as a pair. Thus, the two covariance matrices play equivalent roles. Second, if we ignore all of the covariance information, we see that Equation (8.193) looks like $\mathbf{G}^T[\mathbf{G}\mathbf{G}^T]^{-1}$, which is the minimum length operator. Third, if we again ignore all covariance information, Equation (8.194) looks like $[\mathbf{G}^T\mathbf{G}]^{-1}\mathbf{G}^T$, which is the least squares operator. Thus, we see that the maximum likelihood operator can be viewed as some kind of a combined weighted least squares and weighted minimum length operator.

The maximum likelihood solution \mathbf{m}_{MX} is given by

$$\begin{aligned}\mathbf{m}_{MX} &= \langle \mathbf{m} \rangle + \mathbf{G}_{MX}^{-1} [\mathbf{d} - \mathbf{G}\langle \mathbf{m} \rangle] \\ &= \langle \mathbf{m} \rangle + \mathbf{G}_{MX}^{-1} \mathbf{d} - \mathbf{G}_{MX}^{-1} \mathbf{G} \langle \mathbf{m} \rangle \\ &= \mathbf{G}_{MX}^{-1} \mathbf{d} + [\mathbf{I} - \mathbf{R}] \langle \mathbf{m} \rangle\end{aligned}\quad (8.195)$$

where \mathbf{R} is the model resolution matrix. Equation (8.195) explicitly shows the dependence of \mathbf{m}_{MX} on the prior estimate of the solution $\langle \mathbf{m} \rangle$. If there is perfect model resolution, then $\mathbf{R} = \mathbf{I}$, and \mathbf{m}_{MX} is independent of $\langle \mathbf{m} \rangle$. If the i th row of \mathbf{R} is equal to the i th row of the identity matrix, then there will be no dependence on the i th entry in \mathbf{m}_{MX} on the i th entry in $\langle \mathbf{m} \rangle$.

Menke points out that there are several interesting limiting cases for the maximum likelihood operator. We begin by assuming some simple forms for the covariance matrices:

$$[\text{cov } \mathbf{g}] = \sigma_g^2 \mathbf{I}_N \quad (8.196)$$

$$[\text{cov } \mathbf{m}] = \sigma_m^2 \mathbf{I}_M \quad (8.197)$$

$$[\text{cov } \mathbf{d}] = \sigma_d^2 \mathbf{I}_N \quad (8.198)$$

In the first case we assume that the data and theory are much better known than $\langle \mathbf{m} \rangle$. In the limiting case we can assume $\sigma_d^2 = \sigma_g^2 = 0$. If we do, then Equation (8.193) reduces to $\mathbf{G}_{MX}^{-1} = \mathbf{G}^T[\mathbf{G}\mathbf{G}^T]^{-1}$, the minimum length operator. If we assume that [cov \mathbf{m}] still has some structure, then Equation (8.193) reduces to

$$\mathbf{G}_{MX}^{-1} = [\text{cov } \mathbf{m}]^{-1} \mathbf{G}^T \{ \mathbf{G} [\text{cov } \mathbf{m}]^{-1} \mathbf{G}^T \}^{-1} \quad (8.150)$$

the weighted minimum length operator. If we assume only that σ_d^2 and σ_g^2 are much less than σ_m^2 and that $1/\sigma_m^2$ goes to 0, then Equation (8.194) reduces to $\mathbf{G}_{MX}^{-1} = [\mathbf{G}^T\mathbf{G}]^{-1}\mathbf{G}^T$, or the least squares operator. It is important to realize that $[\mathbf{G}^T\mathbf{G}]^{-1}$ only exists when $P = M$, and $[\mathbf{G}\mathbf{G}^T]^{-1}$ only exists when $P = N$. Thus, either form, or both, may fail to exist, depending on P . The

simplifying assumptions about σ_d^2 , σ_g^2 , and σ_m^2 can thus break down the equivalence between Equations (8.193) and (8.194).

A second limiting case involves assuming no confidence in either (or both) the data or theory. That is, we let σ_d^2 and/or σ_g^2 go to infinity. Then we see that \mathbf{G}_{MX}^{-1} goes to $\mathbf{0}$ and $\mathbf{m}_{MX} = \langle \mathbf{m} \rangle$. This makes sense if we realize that we have assumed the data are useless (and/or the theory), and hence we do not have a useful forward problem to move us away from our prior estimate $\langle \mathbf{m} \rangle$.

We have assumed in deriving Equations (8.193) and (8.194) that all of the covariance matrices represent Gaussian processes. In this case, we have shown that maximum likelihood approaches will yield the same solution as weighted least squares ($P = M$), weighted minimum length ($P = N$), or weighted generalized inverse approaches. If the probability density functions are not Gaussian, then maximum likelihood approaches can lead to different solutions. If the distributions are Gaussian, however, then all of the modifications introduced in Section 8.2 for the generalized inverse can be thought of as the maximum likelihood approach.