Chapter 4

Solution of the Linear, Gaussian Inverse Problem, Viewpoint 2: Generalized Inverses

4.1 SOLUTIONS VERSUS OPERATORS

In the previous chapter, we derived methods of solving the linear inverse problem $\mathbf{Gm}=\mathbf{d}$ that were based on examining two properties of its solution: prediction error and solution simplicity (or length). Most of these solutions had a form that was linear in the data, $\mathbf{m}^{\text{est}}=\mathbf{Md}+\mathbf{v}$, where $\mathbf{M}$ is some matrix and $\mathbf{v}$ is some vector, both of which are independent of the data $\mathbf{d}$. This equation indicates that the estimate of the model parameters is controlled by some matrix $\mathbf{M}$ operating on the data (that is, multiplying the data). We therefore shift our emphasis from the estimates $\mathbf{m}^{\text{est}}$ to the operator matrix $\mathbf{M}$, with the expectation that by studying it we can learn more about the properties of inverse problems. Since the matrix $\mathbf{M}$ solves, or "inverts," the inverse problem $\mathbf{Gm}=\mathbf{d}$, it is often called the *generalized inverse* and given the symbol $\mathbf{G}^{-g}$. The exact form of the generalized inverse depends on the problem at hand. The generalized inverse of the overdetermined least squares problem is $\mathbf{G}^{-g}=[\mathbf{G}^{T}\mathbf{G}]^{-1}\mathbf{G}^{T}$, and for the minimum length underdetermined solution it is $\mathbf{G}^{-g}=\mathbf{G}^{T}[\mathbf{GG}^{T}]^{-1}$.

Note that in some ways the generalized inverse is analogous to the ordinary matrix inverse. The solution to the square (even-determined) matrix equation $\mathbf{Ax}=\mathbf{y}$ is $\mathbf{x}=\mathbf{A}^{-1}\mathbf{y}$, and the solution to the inverse problem $\mathbf{Gm}=\mathbf{d}$ is $\mathbf{m}^{\text{est}}=\mathbf{G}^{-g}\mathbf{d}$ (plus some vector, possibly). The analogy is very limited, however. The generalized inverse is not a matrix inverse in the usual sense. It is not square, and neither $\mathbf{G}^{-g}\mathbf{G}$ nor $\mathbf{GG}^{-g}$ need equal an identity matrix.

4.2 THE DATA RESOLUTION MATRIX

Suppose we have found a generalized inverse that in some sense solves the inverse problem $\mathbf{Gm}=\mathbf{d}$, yielding an estimate of the model parameters $\mathbf{m}^{\text{est}}=\mathbf{G}^{-g}\mathbf{d}$ (for the sake of simplicity we assume that there is no additive vector). We can then retrospectively ask how well this estimate of the model

Geophysical Data Analysis: Discrete Inverse Theory. DOI: 10.1016/B978-0-12-397160-9.00004-7

parameters fits the data. By plugging our estimate into the equation $\mathbf{Gm}=\mathbf{d}$ we conclude

$$\mathbf{d}^{\text{pre}} = \mathbf{Gm}^{\text{est}} = \mathbf{G}[\mathbf{G}^{-g}\mathbf{d}^{\text{obs}}] = [\mathbf{G}\mathbf{G}^{-g}]\mathbf{d}^{\text{obs}} = \mathbf{N}\mathbf{d}^{\text{obs}} \tag{4.1}$$

Here, the superscripts obs and pre mean observed and predicted, respectively. The $N \times N$ square matrix $\mathbf{N}=\mathbf{G}\mathbf{G}^{-g}$ is called the *data resolution matrix*. This matrix describes how well the predictions match the data. If $\mathbf{N}=\mathbf{I}$, then $\mathbf{d}^{\text{pre}}=\mathbf{d}^{\text{obs}}$ and the prediction error is zero. On the other hand, if the data resolution matrix is not an identity matrix, the prediction error is nonzero.

If the elements of the data vector $\mathbf{d}$ possess a natural ordering, then the data resolution matrix has a simple interpretation. Consider, for example, the problem of fitting a curve to (z, d) points, where the data have been ordered according to the value of the auxiliary variable z. If $\mathbf{N}$ is not an identity matrix but is close to an identity matrix (in the sense that its largest elements are near its main diagonal), then the configuration of the matrix signifies that averages of neighboring data can be predicted, whereas individual data cannot. Consider the ith row of $\mathbf{N}$. If this row contained all zeros except for a one in the ith column, then d_i would be predicted exactly. On the other hand, suppose that the row contained the elements

$$[\ldots 0 \quad 0 \quad 0 \quad 0.1 \quad 0.8 \quad 0.1 \quad 0 \quad 0 \quad 0 \ldots] \tag{4.2}$$

where the 0.8 is in the ith column. Then the ith datum is given by

$$d_i^{\text{pre}} = \sum_{j=1}^{N} N_{ij} d_j^{\text{obs}} = 0.1 d_{i-1}^{\text{obs}} + 0.8 d_i^{\text{obs}} + 0.1 d_{i+1}^{\text{obs}} \tag{4.3}$$

The predicted value is a weighted average of three neighboring observed data. If the true data vary slowly with the auxiliary variable, then such an average might produce an estimate reasonably close to the observed value.

The rows of the data resolution matrix $\mathbf{N}$ describe how well neighboring data can be independently predicted, or *resolved*. If the data have a natural ordering, then a graph of the elements of the rows of $\mathbf{N}$ against column indices illuminates the sharpness of the resolution (Figure 4.1A). If the graphs have a single sharp maximum centered about the main diagonal, then the data are well resolved. If the graphs are very broad, then the data are poorly resolved. Even in cases where there is no natural ordering of the data, the resolution matrix still shows how much weight each observation has in influencing the predicted value. There is then no special significance to whether large off-diagonal elements fall near to or far from the main diagonal.

A straight line has only two parameters and so cannot accurately preduct many independent data. Consequently, the data resolution matrix for the problem of fitting a straight line to data is not diagonal (Figure 4.1B). Its largest amplitudes are at its top-right and bottom-left corners, indicating that the points at the *ends* of the line are controlling the fit.

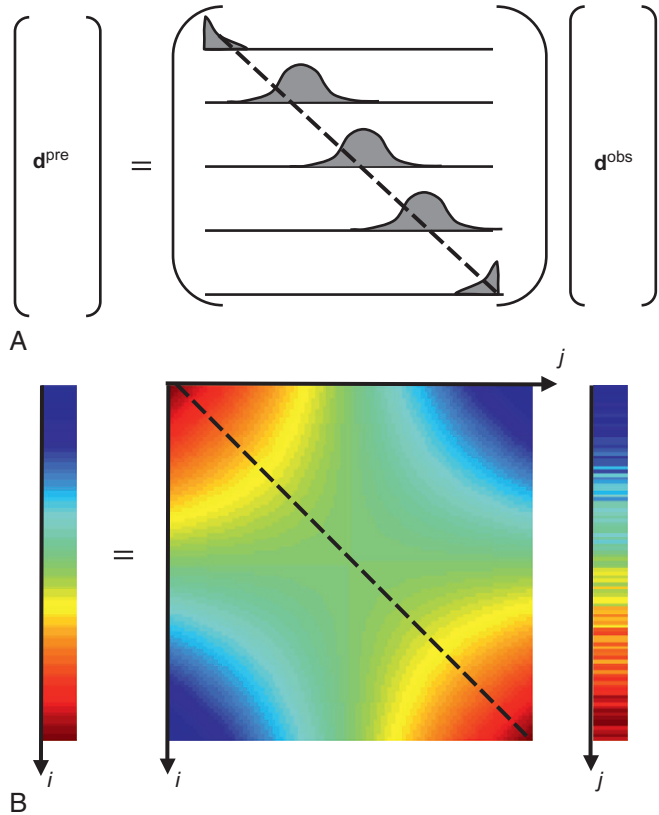

FIGURE 4.1 (A) Plots of selected rows of the data resolution matrix, **N**, indicate how well the data can be predicted. Narrow peaks occurring near the main diagonal of the matrix (dashed line) indicate that the resolution is good. (B) Actual **N** for the case of fitting a straight line to 100 data, equally spaced along the z-axis. Large values (red colors) occur only near the ends of the main diagonal (dashed line), indicating that the resolution is poor at intermediate values of z. *MatLab* script gda04_01.

Because the diagonal elements of the data resolution matrix indicate how much weight a datum has in its own prediction, these diagonal elements are often singled out and called the *importance* **n** of the data (Minster et al., 1974)

$$\mathbf{n} = \text{diag}(\mathbf{N}) \tag{4.4}$$

The data resolution matrix is not a function of the data but only of the data kernel **G** (which embodies the model and experimental geometry) and any *a priori* information applied to the problem. It can therefore be computed and studied without actually performing the experiment and can be a useful tool in experimental design.

4.3 THE MODEL RESOLUTION MATRIX

The data resolution matrix characterizes whether the data can be independently predicted, or resolved. The same question can be asked about the model parameters. To explore this question we imagine that there is a true, but unknown set of model parameters $\mathbf{m}^{\text{true}}$ that solve $\mathbf{G}\mathbf{m}^{\text{true}} = \mathbf{d}^{\text{obs}}$. We then inquire how closely a particular estimate of the model parameters $\mathbf{m}^{\text{est}}$ is to this true solution. Plugging the expression for the observed data $\mathbf{G}\mathbf{m}^{\text{true}} = \mathbf{d}^{\text{obs}}$ into the expression for the estimated model $\mathbf{m}^{\text{est}} = \mathbf{G}^{-g}\mathbf{d}^{\text{obs}}$ gives

$$\mathbf{m}^{\text{est}} = \mathbf{G}^{-g}\mathbf{d}^{\text{obs}} = \mathbf{G}^{-g}[\mathbf{G}\mathbf{m}^{\text{true}}] = [\mathbf{G}^{-g}\mathbf{G}]\mathbf{m}^{\text{true}} = \mathbf{R}\mathbf{m}^{\text{true}} \tag{4.5}$$

(Wiggins, 1972). Here $\mathbf{R}$ is the $M \times M$ *model resolution matrix*. If $\mathbf{R} = \mathbf{I}$, then each model parameter is uniquely determined. If $\mathbf{R}$ is not an identity matrix, then the estimates of the model parameters are really weighted averages of the true model parameters. If the model parameters have a natural ordering (as they would if they represented a discretized version of a continuous function), then plots of the rows of the resolution matrix can be useful in determining to what scale features in the model can actually be resolved (Figure 4.2A). Like the data resolution matrix, the model resolution is a function of only the data kernel and the *a priori* information added to the problem. It is therefore independent of the actual values of the data and can therefore be another important tool in experimental design.

As an example, we examine the resolution of the discrete version of the *Laplace transform*

$$d(c) = \int_0^\infty \exp(-cz)m(z)dz \rightarrow d_i = \sum_{j=1}^{M} \exp(-c_i z_j)m_j \tag{4.6}$$

Here, the datum d_i is a weighed average of the model parameters m_j, with weights that decline exponentially with depth z. The decay rate of the exponential is controlled by the constant, c_i, so that the smaller *i*s correspond to averages over a wider range of depths and the larger *i*s over a shallower range of depths. Not surprisingly, the shallow model parameters are better resolved (Figure 4.2B).

4.4 THE UNIT COVARIANCE MATRIX

The covariance of the model parameters depends on the covariance of the data and the way in which error is mapped from data to model parameters. This mapping is a function of only the data kernel and the generalized inverse, not of the data itself. A *unit covariance matrix* can be defined to characterize the degree of error amplification that occurs in the mapping. If the data are assumed to be uncorrelated and to have uniform variance σ^2, the unit covariance matrix is given by

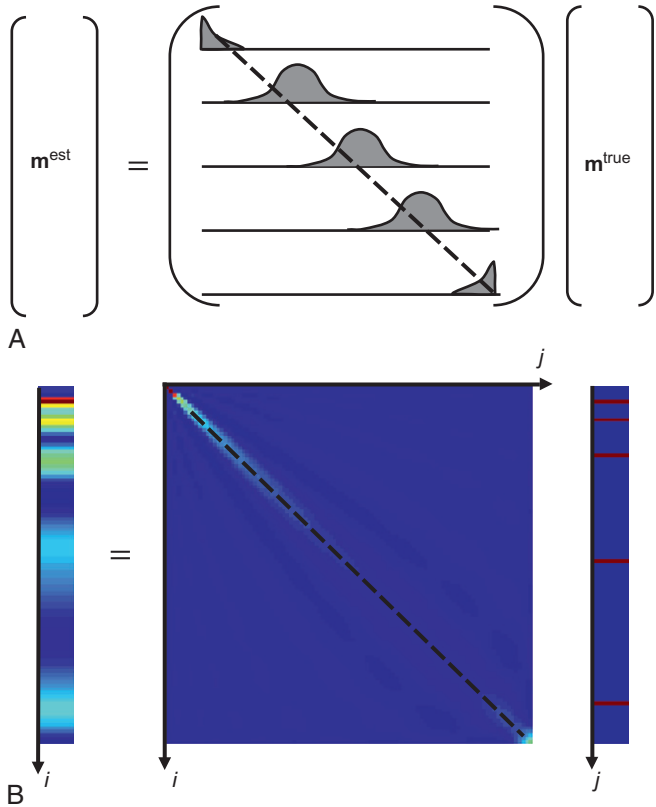

FIGURE 4.2 (A) Plots of selected rows of the model resolution matrix, **R**, indicate how well the model parameters can be resolved. Narrow peaks occurring near the main diagonal of the matrix (dashed line) indicate that the resolution is good. (B) Actual **R** for the case where the model parameters, $m_j(z_j)$, are related to the data through the kernel, $G_{ij}=\exp(-c_i z_j)$, where the cs are constants. Large values (red colors) occur only near the top (small z) of the main diagonal (dashed line), indicating that the resolution is poor at larger values of z. *MatLab* script gda04_02.

$$[\mathrm{cov}_{\mathrm{u}}\mathbf{m}] = \sigma^{-2}\mathbf{G}^{-\mathrm{g}}[\mathrm{cov}\ \mathbf{d}]\mathbf{G}^{-\mathrm{gT}} = \mathbf{G}^{-\mathrm{g}}\mathbf{G}^{-\mathrm{gT}} \tag{4.7}$$

Even if the data are correlated, one can often find some normalization of the data covariance matrix, so that one can define a *unit data covariance matrix* $[\mathrm{cov}_{\mathrm{u}}\ \mathbf{d}]$, related to the model covariance matrix by

$$[\mathrm{cov}_{\mathrm{u}}\mathbf{m}] = \mathbf{G}^{-\mathrm{g}}[\mathrm{cov}_{\mathrm{u}}\ \mathbf{d}]\mathbf{G}^{-\mathrm{gT}} \tag{4.8}$$

The unit covariance matrix is a useful tool in experimental design, especially because it is independent of the actual values and variances of the data themselves.

As an example, reconsider the problem of fitting a straight line to (z, d) data. The unit covariance matrix for intercept m_1 and slope m_2 is given by

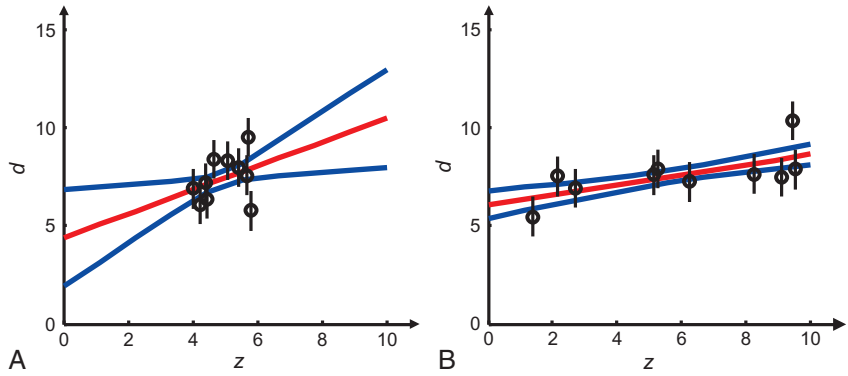

FIGURE 4.3 (A) The method of least squares is used to fit a straight line (red) to uncorrelated data (black circles) with uniform variance (vertical bars, 1σ confidence limits). Since the data are not well-separated in z, the variance of the slope and intercept is large, and consequently the variance of the predicted data is large as well (blue curves, 1σ confidence limits). (B) Same as (A) but with the data well-separated in z. Although the variance of the data is the same as in (A), the variance of the intercept and slope, and consequently the predicted data, is much smaller. *MatLab* script gda04_03.

$$[\mathrm{cov}_\mathrm{u}\ \mathbf{m}] = \frac{1}{N\sum z_i^2 - (\sum z_i)^2}\begin{bmatrix} \sum z_i^2 & -\sum z_i \\ -\sum z_i & N \end{bmatrix} \tag{4.9}$$

Note that the estimates of intercept and slope are uncorrelated only when the data are centered about $z=0$. The overall size of the variance is controlled by the denominator of the fraction. If all the z values are nearly equal, then the denominator of the fraction is small and the variance of the intercept and slope is large (Figure 4.3A). On the other hand, if the z values have a large spread, the denominator is large, and the variance is small (Figure 4.3B).

4.5 RESOLUTION AND COVARIANCE OF SOME GENERALIZED INVERSES

The data and model resolution and unit covariance matrices describe many interesting properties of the solutions to inverse problems. We therefore calculate these quantities for some of the simpler generalized inverses (with $[\mathrm{cov}_\mathrm{u}\ \mathbf{d}] = \mathbf{I}$).

4.5.1 Least Squares

$$\begin{aligned} \mathbf{G}^{-\mathrm{g}} &= [\mathbf{G}^\mathrm{T}\mathbf{G}]^{-1}\mathbf{G}^\mathrm{T} \\ \mathbf{N} = \mathbf{G}\mathbf{G}^{-\mathrm{g}} &= \mathbf{G}[\mathbf{G}^\mathrm{T}\mathbf{G}]^{-1}\mathbf{G}^\mathrm{T} \\ \mathbf{R} = \mathbf{G}^{-\mathrm{g}}\mathbf{G} &= [\mathbf{G}^\mathrm{T}\mathbf{G}]^{-1}\mathbf{G}^\mathrm{T}\mathbf{G} = \mathbf{I} \\ [\mathbf{cov}_\mathrm{u}\ \mathbf{m}] = \mathbf{G}^{-\mathrm{g}}\mathbf{G}^{-\mathrm{gT}} &= [\mathbf{G}^\mathrm{T}\mathbf{G}]^{-1}\mathbf{G}^\mathrm{T}\mathbf{G}[\mathbf{G}^\mathrm{T}\mathbf{G}]^{-1} = [\mathbf{G}^\mathrm{T}\mathbf{G}]^{-1} \end{aligned} \tag{4.10}$$

4.5.2 Minimum Length

$$\mathbf{G}^{-g} = \mathbf{G}^{T}[\mathbf{G}\mathbf{G}^{T}]^{-1}$$

$$\mathbf{N} = \mathbf{G}\mathbf{G}^{-g} = \mathbf{G}\mathbf{G}^{T}[\mathbf{G}\mathbf{G}^{T}]^{-1} = \mathbf{I}$$

$$\mathbf{R} = \mathbf{G}^{-g}\mathbf{G} = \mathbf{G}^{T}[\mathbf{G}\mathbf{G}^{T}]^{-1}\mathbf{G}$$

$$[\mathrm{cov}_{u}\ \mathbf{m}] = \mathbf{G}^{-g}\mathbf{G}^{-gT} = \mathbf{G}^{T}[\mathbf{G}\mathbf{G}^{T}]^{-\mathbf{1}}[\mathbf{G}\mathbf{G}^{T}]^{-\mathbf{1}}\mathbf{G} = \mathbf{G}^{T}[\mathbf{G}^{T}\mathbf{G}]^{-2}\mathbf{G} \quad (4.11)$$

Note that there is a great deal of symmetry between the least squares and minimum length solutions. Least squares solves the completely overdetermined problem and has perfect model resolution; minimum length solves the completely underdetermined problem and has perfect data resolution. As we shall see later, generalized inverses that solve the intermediate mixed-determined problems will have data and model resolution matrices that are intermediate between these two extremes.

4.6 MEASURES OF GOODNESS OF RESOLUTION AND COVARIANCE

Just as we were able to quantify the goodness of the model parameters by measuring their overall prediction error and simplicity, we shall develop techniques that quantify the goodness of data and model resolution matrices and unit covariance matrices. Because the resolution is best when the resolution matrices are identity matrices, one possible measure of resolution is based on the size, or *spread*, of the off-diagonal elements.

$$\begin{aligned} \mathrm{spread}(\mathbf{N}) &= \|\mathbf{N} - \mathbf{I}\|_2^2 = \sum_{i=1}^{N}\sum_{j=1}^{N}[N_{ij} - \delta_{ij}]^2 \\ \mathrm{spread}(\mathbf{R}) &= \|\mathbf{R} - \mathbf{I}\|_2^2 = \sum_{i=1}^{M}\sum_{j=1}^{M}[R_{ij} - \delta_{ij}]^2 \end{aligned} \quad (4.12)$$

Here δ_{ij} are the elements of the identity matrix $\mathbf{I}$. These measures of the goodness of the resolution spread are based on the L_2 norm of the difference between the resolution matrix and an identity matrix. They are sometimes called the *Dirichlet spread functions*. When $\mathbf{R} = \mathbf{I}$, spread$(\mathbf{R}) = 0$.

Since the unit standard deviation of the model parameters is a measure of the amount of error amplification mapped from data to model parameters, this quantity can be used to estimate the size of the unit covariance matrix as

$$\mathrm{size}\ ([\mathrm{cov}_{u}\mathbf{m}]) = \left\|[\mathrm{var}_{u}\mathbf{m}]^{1/2}\right\|_2^2 = \sum_{i=1}^{M}[\mathrm{cov}_{u}\mathbf{m}]_{ii} \quad (4.13)$$

where the square root is interpreted element by element. Note that this measure of covariance size does not take into account the size of the off-diagonal elements in the unit covariance matrix.

4.7 GENERALIZED INVERSES WITH GOOD RESOLUTION AND COVARIANCE

Having found a way to measure quantitatively the goodness of the resolution and covariance of a generalized inverse, we now consider whether it is possible to use these measures as guiding principles for deriving generalized inverses. This procedure is analogous to that of Chapter 3, which involves first defining measures of solution prediction error and simplicity and then using those measures to derive the least squares and minimum length estimates of the model parameters.

4.7.1 Overdetermined Case

We first consider a purely overdetermined problem of the form $\mathbf{Gm}=\mathbf{d}$. We postulate that this problem has a solution of the form $\mathbf{m}^{\text{est}}=\mathbf{G}^{-g}\mathbf{d}$ and try to determine $\mathbf{G}^{-g}$ by minimizing some combination of the above measures of goodness. Since we previously noted that the overdetermined least squares solution had perfect model resolution, we shall try to determine $\mathbf{G}^{-g}$ by minimizing only the spread of the data resolution. We begin by examining the spread of the kth row of $\mathbf{N}$, say, J_k:

$$J_k = \sum_{i=1}^{N} (N_{ki} - \delta_{ki})^2 = \sum_{i=1}^{N} N_{ki}{}^2 - 2\sum_{i=1}^{N} N_{ki}\delta_{ki} + \sum_{i=1}^{N} \delta_{ki}^2 \tag{4.14}$$

Since each of the J_ks is positive, we can minimize the total spread$(\mathbf{N}) = \sum J_k$ by minimizing each individual J_k. We therefore insert the definition of the data resolution matrix $\mathbf{N}=\mathbf{GG}^{-g}$ into the formula for J_k and minimize it with respect to the elements of the generalized inverse matrix:

$$\frac{\partial J_k}{\partial G_{qr}^{-g}} = 0 \tag{4.15}$$

We shall perform the differentiation separately for each of the three terms of J_k. The first term is given by

$$\begin{aligned}\frac{\partial}{\partial G_{qr}^{-g}}\left[\sum_{i=1}^{N}\left[\sum_{j=1}^{M} G_{kj}G_{ji}^{-g}\right]\left[\sum_{p=1}^{M} G_{kp}G_{pi}^{-g}\right]\right] &= \frac{\partial}{\partial G_{qr}^{-g}}\left[\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{p=1}^{M} G_{ji}^{-g}G_{pi}^{-g}G_{kj}G_{kp}\right] \\ &= 2\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{p=1}^{M}\delta_{jq}\delta_{ir}G_{pi}^{-g}G_{kj}G_{kp} \\ &= 2\sum_{p=1}^{M} G_{pr}^{-g}G_{kq}G_{kp} \end{aligned} \tag{4.16}$$

The second term is given by

$$-2\frac{\partial}{\partial G_{qr}^{-g}}\sum_{i=1}^{N}\sum_{j=1}^{M}G_{kj}G_{ji}^{-g}\delta_{ki}=-2\sum_{i=1}^{N}\sum_{j=1}^{M}G_{kj}\delta_{jq}\delta_{ir}\delta_{ki}=-2G_{kq}\delta_{kr} \quad (4.17)$$

The third term is zero, since it is not a function of the generalized inverse. The complete equation is $\sum_{p=1}^{M}G_{kq}G_{kp}G_{pr}^{-g}=G_{kq}\delta_{kr}$. After summing over k and converting to matrix notation, we obtain

$$\mathbf{G}^{\mathrm{T}}\mathbf{G}\mathbf{G}^{-\mathrm{g}}=\mathbf{G}^{\mathrm{T}} \quad (4.18)$$

Since $\mathbf{G}^{\mathrm{T}}\mathbf{G}$ is square, we can premultiply by its inverse to solve for the generalized inverse, $\mathbf{G}^{-\mathrm{g}}=[\mathbf{G}^{\mathrm{T}}\mathbf{G}]^{-1}\mathbf{G}^{\mathrm{T}}$, which is precisely the same as the formula for the least squares generalized inverse. The least squares generalized inverse can be interpreted either as the inverse that minimizes the L_2 norm of the prediction error or as the inverse that minimizes the Dirichlet spread of the data resolution.

4.7.2 Underdetermined Case

The data can be satisfied exactly in a purely underdetermined problem. The data resolution matrix is, therefore, precisely an identity matrix and its spread is zero. We might therefore try to derive a generalized inverse for this problem by minimizing the spread of the model resolution matrix with respect to the elements of the generalized inverse. It is perhaps not particularly surprising that the generalized inverse obtained by this method is exactly the minimum length generalized inverse $\mathbf{G}^{-\mathrm{g}}=\mathbf{G}^{\mathrm{T}}[\mathbf{G}\mathbf{G}^{\mathrm{T}}]^{-1}$. The minimum length solution can be interpreted either as the inverse that minimizes the L_2 norm of the solution length or as the inverse that minimizes the Dirichlet spread of the model resolution. This is another aspect of the symmetrical relationship between the least squares and minimum length solutions.

4.7.3 The General Case with Dirichlet Spread Functions

We seek the generalized inverse $\mathbf{G}^{-\mathrm{g}}$ that minimizes the weighted sum of Dirichlet measures of resolution spread and covariance size.

$$\text{Minimize: } \alpha_1\,\text{spread}(\mathbf{N})+\alpha_2\,\text{spread}(\mathbf{R})+\alpha_3\,\text{size}([\text{cov}_{\mathrm{u}}\mathbf{m}]) \quad (4.19)$$

where the αs are arbitrary weighting factors. This problem is done in exactly the same fashion as the one in Section 4.7.1, except that there is now three times as much algebra. The result is an equation for the generalized inverse:

$$\alpha_1[\mathbf{G}^{\mathrm{T}}\mathbf{G}]\mathbf{G}^{-\mathrm{g}}+\mathbf{G}^{-\mathrm{g}}\left[\alpha_2[\mathbf{G}\mathbf{G}^{\mathrm{T}}]+\alpha_3[\text{cov}_{\mathrm{u}}\mathbf{d}]\right]=[\alpha_1+\alpha_2]\mathbf{G}^{\mathrm{T}} \quad (4.20)$$

An equation of this form is called a *Sylvester equation*. It is just a set of linear equations in the elements of the generalized inverse $\mathbf{G}^{-\mathrm{g}}$ and so could be solved by writing the elements of $\mathbf{G}^{-\mathrm{g}}$ as a vector in a huge $NM\times NM$ matrix equation, but it has no explicit solution in terms of algebraic functions of the component matrices. Explicit solutions can be written, however, for a variety of

special choices of the weighting factors. The least squares solution is recovered if $\alpha_1 = 1$ and $\alpha_2 = \alpha_3 = 0$, and the minimum length solution is recovered if $\alpha_1 = 0$, $\alpha_2 = 1$, and $\alpha_3 = 0$. Of more interest is the case in which $\alpha_1 = 1$, $\alpha_2 = 0$, α_3 equals some constant (say, ε^2) and $[\text{cov}_u\, \mathbf{d}] = \mathbf{I}$. The generalized inverse is then given by

$$\mathbf{G}^{-g} = [\mathbf{G}^T\mathbf{G} + \varepsilon^2\mathbf{I}]^{-1}\mathbf{G}^T \tag{4.21}$$

This formula is precisely the damped least squares inverse, which we derived in the previous chapter by minimizing a combination of prediction error and solution length. The damped least squares solution can also be interpreted as the inverse that minimizes a weighted combination of data resolution spread and covariance size.

Another interesting solution is obtained when a weighted combination of model resolution spread and covariance size is minimized. Setting $\alpha_1 = 0$, $\alpha_2 = 1$, $\alpha_2 = \varepsilon^2$, and $[\text{cov}_u\, \mathbf{d}] = \mathbf{I}$, we find

$$\mathbf{G}^{-g} = \mathbf{G}^T[\mathbf{G}\mathbf{G}^T + \varepsilon^2\mathbf{I}]^{-1} \tag{4.22}$$

This solution might be termed *damped minimum length*. It will be important in the discussion later in this chapter, because it is the Dirichlet analog to the Backus-Gilbert generalized inverse that will be introduced there.

Note that it is quite possible for these generalized inverses to possess resolution matrices containing *negative* off-diagonal elements. Physically, an average makes most sense when it contained only positive weighting factors, so negative elements interfere with the interpretation of the rows of $\mathbf{R}$ as localized averages. In principle, it is possible to include non-negativity as a constraint when choosing the generalized inverse by minimizing the spread functions. However, in practice, this constraint is never implemented because it makes the calculation of the generalized inverse very difficult. Furthermore, the more constraints that one places on $\mathbf{R}$, the less localized it tends to become.

4.8 SIDELOBES AND THE BACKUS-GILBERT SPREAD FUNCTION

The Dirichlet spread function is not a particularly appropriate measure of the goodness of resolution when the data or model parameters have a natural ordering because the off-diagonal elements of the resolution matrix are all weighted equally, regardless of whether they are close or far from the main diagonal. We would much prefer that any large elements be close to the main diagonal when there is a natural ordering (Figure 4.4) because the rows of the resolution matrix then represent *localized* averaging functions.

If one uses the Dirichlet spread function to compute a generalized inverse, it will often have *sidelobes*, that is, large amplitude regions in the resolution matrices far from the main diagonal. We would prefer to find a generalized inverse

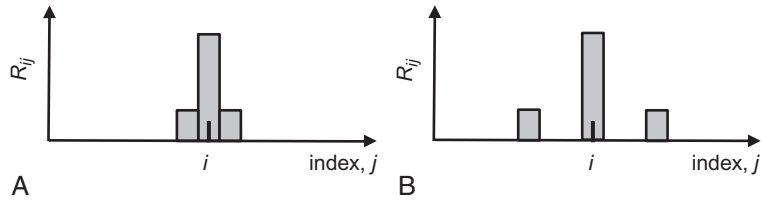

FIGURE 4.4 (A, B) Resolution matrices have the same spread when measured by the Dirichlet spread function. Nevertheless, if the model parameters possess a natural ordering, then (A) is better resolved. The Backus-Gilbert spread function is designed to measure (A) as having a smaller spread than (B).

without sidelobes, even at the expense of widening the band of nonzero elements near the main diagonal, since a solution with such a resolution matrix is then interpretable as a localized average of physically adjacent model parameters.

We therefore add a weighting factor $w(i, j)$ to the measure of spread that weights the (i, j) element of $\mathbf{R}$ according to its physical distance from the diagonal element. This weighting preferentially selects resolution matrices that are "spiky," or "deltalike." If the natural ordering were a simple linear one, then the choice $w(i, j) = (i-j)^2$ would be reasonable. If the ordering is multidimensional, a more complicated weighting factor is needed. It is usually convenient to choose the spread function so that the diagonal elements have no weight, i.e., $w(i, i) = 0$, and so that $w(i, j)$ is always non-negative and symmetric in i and j. The new spread function, often called the Backus-Gilbert spread function (Backus and Gilbert, 1967, 1968), is then given by

$$\text{spread}(\mathbf{R}) = \sum_{i=1}^{M}\sum_{j=1}^{M} w(i,j)[R_{ij} - \delta_{ij}]^2 = \sum_{i=1}^{M}\sum_{j=1}^{M} w(i,j)R_{ij}^2 \tag{4.23}$$

A similar expression holds for the spread of the data resolution. One can now use this measure of spread to derive new generalized inverses. Their sidelobes will be smaller than those based on the Dirichlet spread functions. On the other hand, they are sometimes worse when judged by other criteria. As we shall see, the Backus-Gilbert generalized inverse for the completely underdetermined problem does not exactly satisfy the data, even though the analogous minimum length generalized inverse does. These facts demonstrate that there are unavoidable trade-offs inherent in finding solutions to inverse problems.

4.9 THE BACKUS-GILBERT GENERALIZED INVERSE FOR THE UNDERDETERMINED PROBLEM

This problem is analogous to deriving the minimum length solution by minimizing the Dirichlet spread of model resolution. Since it is very easy to satisfy the data when the problem is underdetermined (so that the data resolution has small

spread), we shall find a generalized inverse that minimizes the spread of the model resolution alone.

We seek the generalized inverse $\mathbf{G}^{-g}$ that minimizes the Backus-Gilbert spread of model resolution. Since the diagonal elements of the model resolution matrix are given no weight, we also require that the resulting model resolution matrix satisfy the equation

$$\sum_{j=1}^{M} R_{ij} = [1]_i \tag{4.24}$$

This constraint ensures that the diagonal of the resolution matrix is finite and that the rows are unit averaging functions acting on the true model parameters. Writing the spread of one row of the resolution matrix as J_k and inserting the expression for the resolution matrix, we have

$$\begin{aligned} J_k &= \sum_{l=1}^{M} w(l,k) R_{kl} R_{kl} \\ &= \sum_{l=1}^{M} w(l,k) \left[\sum_{i=1}^{N} G_{ki}^{-g} G_{il} \right] \left[\sum_{j=1}^{N} G_{kj}^{-g} G_{jl} \right] \\ &= \sum_{i=1}^{N} \sum_{j=1}^{N} G_{ki}^{-g} G_{kj}^{-g} \sum_{l=1}^{M} w(l,k) G_{il} G_{jl} \\ &= \sum_{i=1}^{N} \sum_{j=1}^{N} G_{ki}^{-g} G_{kj}^{-g} S_{ij}^{(k)} \end{aligned} \tag{4.25}$$

where the quantity $S_{ij}^{(k)}$ is defined as

$$S_{ij}^{(k)} = \sum_{l=1}^{M} w(l,k) G_{il} G_{jl} \tag{4.26}$$

The left-hand side of the constraint equation $\sum_j R_{ij} = [1]_i$ can also be written in terms of the generalized inverse

$$\sum_{k=1}^{M} R_{ik} = \sum_{k=1}^{M} \left[\sum_{j=1}^{N} G_{ij}^{-g} G_{jk} \right] = \sum_{j=1}^{N} G_{ij}^{-g} \sum_{k=1}^{M} G_{jk} = \sum_{j=1}^{N} G_{ij}^{-g} u_j \tag{4.27}$$

Here the quantity u_j is defined as

$$u_j = \sum_{k=1}^{M} G_{jk} \tag{4.28}$$

The problem of minimizing J_k with respect to the elements of the generalized inverse (under the given constraints) can be solved through the use of Lagrange multipliers. We first define a Lagrange function Φ such that

$$\Phi = \sum_{i=1}^{N}\sum_{j=1}^{M} G_{ki}^{-g} G_{kj}^{-g} S_{ij}^{(k)} + 2\lambda \sum_{j=1}^{N} G_{kj}^{-g} u_j \tag{4.29}$$

where 2λ is the Lagrange multiplier. We then differentiate Φ with respect to the elements of the generalized inverse and set the result equal to zero as

$$\frac{\partial \Phi}{\partial G_{kp}^{-g}} = 2\sum_{i=1}^{N} S_{pi}^{(k)} G_{ki}^{-g} + 2\lambda u_p = 0 \tag{4.30}$$

(Note that one can solve for each row of the generalized inverse separately, so that it is only necessary to take derivatives with respect to the elements in the kth row.) The above equation must be solved along with the original constraint equation. Treating the kth row of $\mathbf{G}^{-g}$ as the transform of a column-vector $\mathbf{g}^{(k)}$ and the quantity $S_{ij}^{(k)}$ as a matrix $\mathbf{S}^{(k)}$, we can write these equations as the matrix equation

$$\begin{bmatrix} \mathbf{S}^{(k)} & \mathbf{u} \\ \mathbf{u}^{T} & 0 \end{bmatrix} \begin{bmatrix} \mathbf{g}^{(k)} \\ \lambda \end{bmatrix} = \begin{bmatrix} \mathbf{O} \\ 1 \end{bmatrix} \tag{4.31}$$

This is a square $(N+1)\times(N+1)$ system of linear equations that must be solved for the N elements of the kth row of the generalized inverse and for the one Lagrange multiplier λ.

The matrix equation can be solved explicitly using a variant of the *bordering method* of linear algebra, which is used to construct the inverse of a matrix by partitioning it into submatrices with simple properties. Suppose that the inverse of the symmetric matrix in Equation (4.31) exists and that we partition it into an $N\times N$ symmetric square matrix $\mathbf{A}$, vector $\mathbf{b}$, and scalar c. By assumption, premultiplication by the inverse yields the identity matrix

$$\begin{bmatrix} \mathbf{A} & \mathbf{b} \\ \mathbf{b}^{T} & c \end{bmatrix} \begin{bmatrix} \mathbf{S}^{(k)} & \mathbf{u} \\ \mathbf{u}^{T} & 0 \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{O} \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} \mathbf{A}\mathbf{S}^{(k)} + \mathbf{b}\mathbf{u}^{T} & \mathbf{A}\mathbf{u} \\ \mathbf{b}^{T}\mathbf{S}^{(k)} + \mathbf{c}\mathbf{u}^{T} & \mathbf{b}^{T}\mathbf{u} \end{bmatrix} \tag{4.32}$$

The unknown submatrices $\mathbf{A}$, $\mathbf{b}$, and c can now be determined by equating the submatrices

$$\mathbf{A}\mathbf{S}^{(k)} + \mathbf{b}\mathbf{u}^{T} = \mathbf{I} \quad \text{so that } \mathbf{A} = \left[\mathbf{S}^{(k)}\right]^{-1}\left[\mathbf{I} - \mathbf{b}\mathbf{u}^{T}\right]$$

$$\mathbf{A}\mathbf{u} = \mathbf{O} \quad \text{so that } \left[\mathbf{S}^{(k)}\right]^{-1}\mathbf{u} = \mathbf{b}\mathbf{u}^{T}\mathbf{S}^{(k)}\mathbf{u} \quad \text{and} \quad \mathbf{b} = \frac{\left[\mathbf{S}^{(k)}\right]^{-1}\mathbf{u}}{\mathbf{u}^{T}\left[\mathbf{S}^{(k)}\right]^{-1}\mathbf{u}}$$

$$\mathbf{b}^{T}\mathbf{S}^{(k)} + c\mathbf{u}^{T} = 0 \quad \text{so that } c = \frac{-1}{\mathbf{u}^{T}\left[\mathbf{S}^{(k)}\right]^{-1}\mathbf{u}} \tag{4.33}$$

Multiplying Equation (4.31) by the inverse matrix yields $\mathbf{g}^{(k)} = \mathbf{b}$ and $\lambda = c$. The generalized inverse, written with summations, is

$$G_{kl}^{-g} = \frac{\sum_{i=1}^{N} u_i \left[\left(\mathbf{S}^{(k)} \right)^{-1} \right]_{il}}{\sum_{i=1}^{N} \sum_{j=1}^{N} u_i \left[\left(\mathbf{S}^{(k)} \right)^{-1} \right]_{ij} u_j} \tag{4.34}$$

This generalized inverse is the Backus-Gilbert analog to the minimum length solution.

As an example, we compare the Dirichlet and Backus-Gilbert solutions for the Laplace transform problem discussed in Section 4.3 (Figure 4.5). The Backus-Gilbert solution is the smoother of the two and has a corresponding model resolution matrix that consists of a single band along the main diagonal. The Dirichlet solution has more details but also more artifacts (such as negative values at $z \approx 3$). They are associated with the large-amplitude sidelobes in the corresponding model resolution matrix.

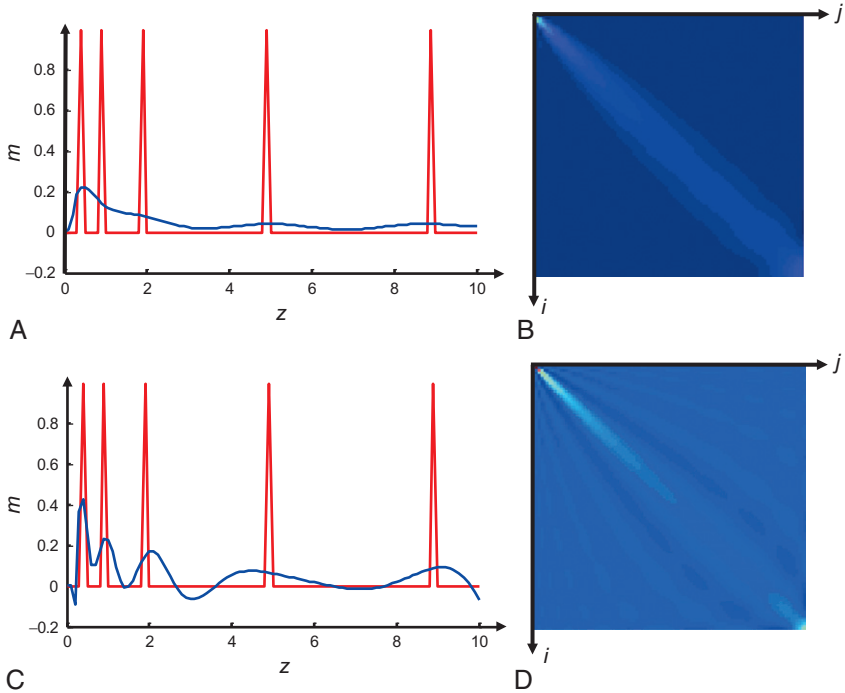

FIGURE 4.5 Comparison of the Backus-Gilbert and Dirichlet solutions of the inverse problem described in Figure 4.2. (A) The true model (red) contains a series of sharp spikes. The estimated model (blue) using the Backus-Gilbert spread function is much smoother, with the width of the smoothing increasing with z. (B) Corresponding model resolution matrix, $\mathbf{R}$. (C, D) Same, but for a Dirichlet spread function. Note that the Backus-Gilbert resolution matrix has the lower intensity sidelobes, but a wider central band. *MatLab* scripts gda04_02 and gda04_04.

4.10 INCLUDING THE COVARIANCE SIZE

The measure of goodness that was used to determine the Backus-Gilbert inverse can be modified to include a measure of the covariance size of the model parameters (Backus and Gilbert, 1970). We shall use the same measure as we did when considering the Dirichlet spread functions, so that goodness is measured by

$$\alpha\,\text{spread}(\mathbf{R}) + (1-\alpha)\,\text{size}([\text{cov}_\text{u}\mathbf{m}]) = \alpha\sum_{i=1}^{M}\sum_{j=1}^{M} w(i,j)R_{ij}^2 + (1-\alpha)\sum_{i=1}^{M}[\text{cov}_\text{u}\mathbf{m}]_{ii} \tag{4.35}$$

where $0 \le \alpha \le 1$ is a weighting factor that determines the relative contribution of model resolution and covariance to the measure of the goodness of the generalized inverse. The goodness J'_k of the kth row is then

$$\begin{aligned} J'_k &= \alpha\sum_{l=1}^{M} w(k,l)R_{kl}^2 + (1-\alpha)\,[\text{cov}_\text{u}\mathbf{m}]_{kk} \\ &= \alpha\sum_{i=1}^{N}\sum_{j=1}^{N} G_{ki}^{-g}G_{kj}^{-g}\,[S_{ij}]_k + (1-\alpha)\sum_{i=1}^{N}\sum_{j=1}^{N} G_{ki}^{-g}G_{kj}^{-g}\,[\text{cov}_\text{u}\mathbf{d}]_{ij} \\ &= \sum_{i=1}^{N}\sum_{j=1}^{N} G_{ki}^{-g}G_{kj}^{-g}S_{ij}^{\prime(k)} \end{aligned} \tag{4.36}$$

where the quantity $S_{ij}^{\prime(k)}$ is defined by the equation

$$S_{ij}^{\prime(k)} = \alpha\,S_{ij}^{(k)} + (1-\alpha)[\text{cov}_\text{u}\mathbf{d}]_{ij} \tag{4.37}$$

Since the function J'_k has exactly the same form as J_k had in the previous section, the generalized inverse is just the previous result with $S_{ij}^{(k)}$ replaced by $S_{ij}^{\prime(k)}$:

$$G_{kl}^{-g} = \frac{\sum_{i=1}^{N} u_i\left[\left(\mathbf{S}^{\prime(k)}\right)^{-1}\right]_{il}}{\sum_{i=1}^{N}\sum_{j=1}^{N} u_i\left[\left(\mathbf{S}^{\prime(k)}\right)^{-1}\right]_{ij} u_j} \tag{4.38}$$

This generalized inverse is the Backus-Gilbert analog to the damped minimum length solution. In *MatLab*, the one-dimensional Backus-Gilbert generalized inverse `GMG` (that is, for the $w(i, j) = (i - j)^2$ weight function) is calculated as

```
GMG = zeros(M,N);
u = G*ones(M,1);
for k = [1:M]
    S = G * diag(([1:M]-k).^2) * G';
    Sp = alpha*S + (1-alpha)*eye(N,N);
    uSpinv = u'/Sp;
    GMG(k,:) = uSpinv / (uSpinv*u);
end
```

(*MatLab* script gda04_05)

In higher dimensions, the definition of $\mathbf{S}$ is more complicated, since the weight function must represent the physical distance between model pararameters. In two dimensions, a reasonable choice is

```
S = G * diag( (abs(ixofj([1:M])-ixofj(k))).^2 + ...
              (abs(iyofj([1:M])-iyofj(k))).^2 ) * G';
```

Here the index vectors `ixofj(k)` and `iyofj(k)` give the x and y values of model parameter $\mathbf{k}$.

4.11 THE TRADE-OFF OF RESOLUTION AND VARIANCE

Suppose that one is attempting to determine a set of model parameters that represents a discretized version of a continuous function, such as X-ray opacity in the medical tomography problem (Figure 4.6). If the discretization is made very fine, then the X-rays will not sample every box; the problem will be underdetermined. If we try to determine the opacity of each box individually, then estimates of opacity will tend to have rather large variance. Few boxes will have several X-rays passing through them, so that little averaging out of the errors will take place. On the other hand, the boxes are very small—and very small features can be detected (the resolution is very good). The large variance can be reduced by increasing the box size (or alternatively, averaging several neighboring boxes). Each of these larger regions will then contain several X-rays, and noise will tend to be averaged out. But because the regions are now larger, small features can no longer be detected and the resolution of the X-ray opacity has become poorer.

This scenario illustrates an important trade-off between model resolution spread and variance size. One can be decreased only at the expense of increasing

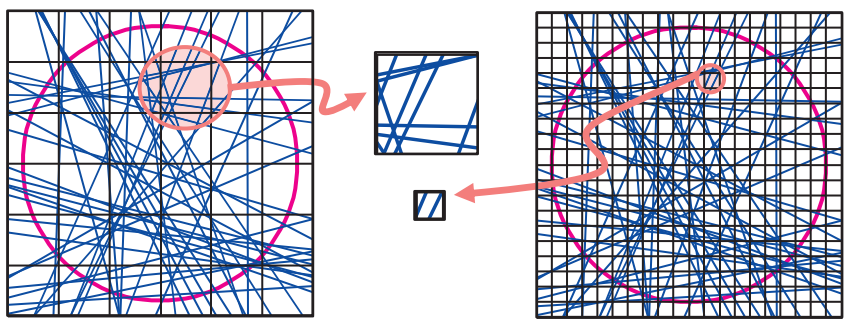

FIGURE 4.6 Hypothetical tomography experiment with (A) large voxels and (B) small voxels. *MatLab* Script. The small voxel case not only has better spatial resolution but also higher variance, as fewer rays pass through each voxel, leaving less opportunity for measurement error to average out. *MatLab* script gda04_06.

the other. We can study this trade-off by choosing a generalized inverse that minimizes a weighted sum of resolution spread and covariance size:

$$\alpha \text{ spread}(\mathbf{R}) + (1-\alpha)\text{size}([\text{cov}_{\mathbf{u}}\mathbf{m}]) \tag{4.39}$$

If the weighting parameter α is set near 1, then the model resolution matrix of the generalized inverse will have small spread, but the model parameters will have large variance. If α is set close to 0, then the model parameters will have a relatively small variance, but the resolution will have a large spread. A *trade-off curve* can be defined by varying α on the interval (0, 1) (Figure 4.7). Such curves can be helpful in choosing a generalized inverse that has an optimum trade-off in model resolution and variance (judged by criteria appropriate to the problem at hand).

Trade-off curves play an important role in continuous inverse theory, where the discretization is (so to speak) infinitely fine, and all problems are underdetermined. It is known that in this continuous limit the curves are monotonic and possess asymptotes in resolution and variance (Figure 4.8). The process of approximating a continuous function by a finite set of discrete parameters somewhat complicates this picture. The resolution and variance, and indeed the solution itself, are dependent on the parameterization, so it is difficult to make any definitive statement regarding the properties of the trade-off curves. Nevertheless, if the discretization is sufficiently fine, the discrete trade-off curves are usually close to ones obtained with the use of continuous inverse theory. Therefore, discretizations should always be made as fine as computational considerations permit.

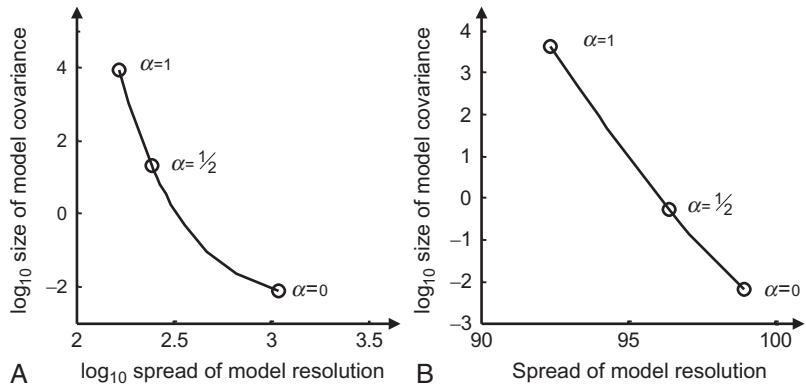

FIGURE 4.7 Trade-off curves of resolution and variance for the inverse problem shown in Figure 4.2. (A) Backus-Gilbert solution, (B) damped minimum length solution. The larger the parameter α, the more weight resolution is given (relative to variance) when forming the generalized inverse. The details of the trade-off curve depend upon the parameterization. The resolution can be no better than the smallest element in the parameterization and no worse than the sum of all the elements. *MatLab* script gda04_05.

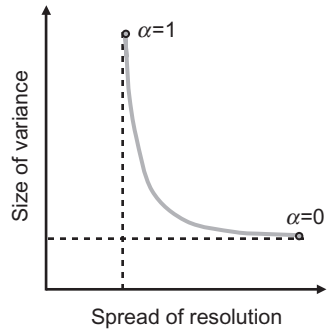

FIGURE 4.8 Trade-off curve of resolution and variance has two asymptotes in the case when the model parameter is a continuous function.

4.12 TECHNIQUES FOR COMPUTING RESOLUTION

In very large problems, the model resolution matrix $\mathbf{R}$ can be cumbersome to compute, owing to its large $M \times M$ size and non-sparse character. Furthermore, time is rarely available for examining all of its rows in detail. Plots of just a few rows, corresponding to model parameters located at strategically chosen points within the model volume, are usually sufficient.

Suppose that we call the kth row of $\mathbf{R}$ the vector $\mathbf{r}^{(k)T}$. Then the identity $\mathbf{R} = \mathbf{RI}$ can be rewritten as

$$R_{ik} = \sum_{j=1}^{M} R_{ij}\delta_{jk} \rightarrow r_i^{(k)} = \sum_{j=1}^{N} R_{ij} m_j^{(k)} \quad \text{with} \quad m_j^{(k)} = \delta_{jk} \tag{4.40}$$

Here we have identified the kth column of $\mathbf{I}$ as "model parameter" vector $m_j^{(k)}$ that is zero except for its kth element, which is unity. Recalling that $\mathbf{R} = \mathbf{G}^{-g}\mathbf{G}$, we can write

$$\mathbf{r}^{(k)} = \mathbf{R}\mathbf{m}^{(k)} = \mathbf{G}^{-g}\mathbf{G}\mathbf{m}^{(k)} = \mathbf{G}^{-g}\mathbf{d}^{(k)} \quad \text{with} \quad \mathbf{d}^{(k)} = \mathbf{G}\mathbf{m}^{(k)} \tag{4.41}$$

Thus, the kth row of the model resolution matrix solves the inverse problem for synthetic data $\mathbf{d}^{(k)}$ corresponding to a specific model parameter vector $\mathbf{m}^{(k)}$, one that is zero except for its kth element, which is unity (that is, a unit spike at row k). This suggests a procedure for calculating the resolution: construct the desired $\mathbf{m}^{(k)}$, solve the forward problem to generate $\mathbf{d}^{(k)}$, solve the inverse problem, and then interpret the result as the kth row of the resolution matrix (Figure 4.9A, B). The great advantage of this technique is that $\mathbf{R}$ in its entirety need not be constructed. Furthermore, the technique will work when the inverse problem is solved by an iterative method, such as the biconjugate gradient method, that does not explicitly construct $\mathbf{G}^{-g}$.

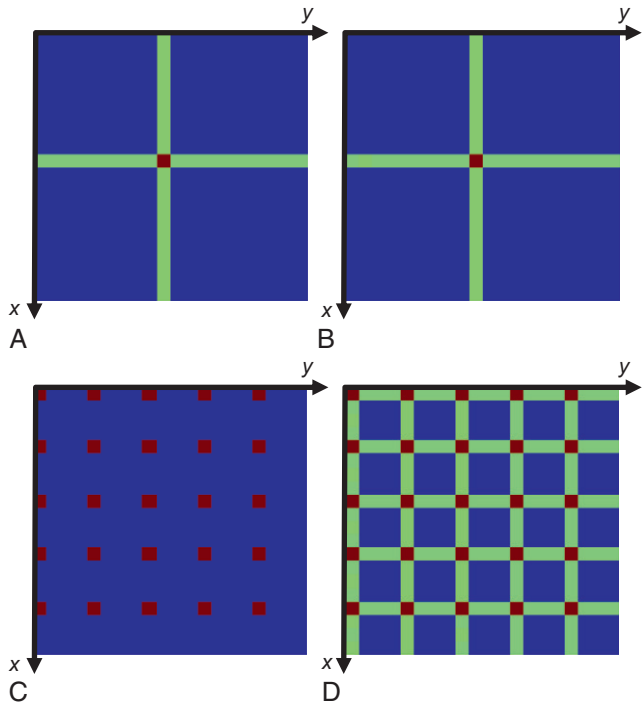

FIGURE 4.9 Resolution of an acoustic tomography problem solved with the minimum length method. The physical model space is a 20×20 grid of pixels on an (x,y) grid. Data are measured only along rows and columns, as in Figure 1.2. (Top row) One row of the resolution matrix, for a model parameter near the center of the (x,y) grid, calculated using two methods, (A) by computing the complete matrix **R** and extracting one row and (B) by calculating the row separately. (Botton row) Checkerboard resolution test showing (C) true checkerboard and (D) reconstructed checkerboard. *MatLab* scripts gda04_07 and gda04_08.

If the resolution of a problem is sufficiently good that the pattern for two well-separated model parameters does not overlap, or overlaps only minimally, then the calculation of two rows of the resolution matrix can be combined into one. One merely solves the inverse problem for synthetic data corresponding to a model parameter vector containing two unit spikes. Nor need one stop with two; a model parameter vector corresponding to a grid of spikes (that is, a *checkerboard*) allows the resolution to be assessed throughout the model volume (Figure 4.9C, D). If the problem has perfect resolution, this checkerboard pattern will be perfectly reproduced. If not, portions of the model volume with poor resolution will contain fuzzy spikes. The main limitation of this technique is that it makes the detection of unlocalized side lobes very difficult, since an unlocalized sidelobe associated with a particular spike will tend to be confused with a localized sidelobe of another spike.

4.13 PROBLEMS

4.1. Consider an underdetermined problem in which each datum is the sum of three neighboring model parameters, that is, $d_i = m_{i-1} + m_i + m_{i+1}$ for $2 \leq i \leq (M-1)$ with $M = 100$. Compute and plot both the Dirichlet and Backus-Gilbert model resolution matrices. Use the standard Backus-Gilbert weight function $w(i, j) = (i-j)^2$. Interpret the results.

4.2. This problem builds upon Problem 4.1. How does the Backus-Gilbert result change if you use the weight function $w(i, j) = |i-j|^{1/2}$, which gives less weight to distant sidelobes?

4.3. This problem is especially difficult. Consider a two-dimensional acoustic tomography problem like the one discussed in Section 1.3.3, consisting of a 20×20 rectangular array of pixels, with observations only along rows and columns. (A) Design an appropriate Backus-Gilbert weight function that quantifies the spread of resolution. (B) Write a *MatLab* script that calculates the model resolution matrix $\mathbf{R}$. (C) Plot a few representative rows of $\mathbf{R}$, but where each row is reorganized into a two-dimensional image, using the same scheme that was applied to the model parameters. Interpret the results. (Hint: You will need to switch back and forth between a 20×20 rectangular array of model parameters and a length $M = 400$ vector of model parameters, as in Figure 10.12.)

REFERENCES

Backus, G.E., Gilbert, J.F., 1967. Numerical application of a formalism for geophysical inverse problems. Geophys. J. Roy. Astron. Soc. 13, 247–276.

Backus, G.E., Gilbert, J.F., 1968. The resolving power of gross earth data. Geophys. J. Roy. Astron. Soc. 16, 169–205.

Backus, G.E., Gilbert, J.F., 1970. Uniqueness in the inversion of gross Earth data. Philos. Trans. R. Soc. Lond. A 266, 123–192.

Minster, J.F., Jordan, T.J., Molnar, P., Haines, E., 1974. Numerical modelling of instantaneous plate tectonics. Geophys. J. Roy. Astron. Soc. 36, 541–576.

Wiggins, R.A., 1972. The general linear inverse problem: Implication of surface waves and free oscillations for Earth structure. Rev. Geophys. Space Phys. 10, 251–285.