

Review of the Generalized Least Squares Method

William Menke

Received: 6 March 2014 / Accepted: 16 August 2014 / Published online: 10 September 2014
© Springer Science+Business Media Dordrecht 2014

Abstract The generalized least squares (GLS) method uses both data and prior information to solve for a best-fitting set of model parameters. We review the method and present simplified derivations of its essential formulas. Concepts of resolution and covariance—essential in all of inverse theory—are applicable to GLS, but their meaning, and especially that of resolution, must be carefully interpreted. We introduce derivations that show that the quantity being resolved is the deviation of the solution from the prior model and that the covariance of the model depends on both the uncertainty in the data and the uncertainty in the prior information. On face value, the GLS formulas for resolution and covariance seem to require matrix inverses that may be difficult to calculate for the very large (but often sparse) linear systems encountered in practical inverse problems. We demonstrate how to organize the computations in an efficient manner and present MATLAB code that implements them. Finally, we formulate the well-understood problem of interpolating data with minimum curvature splines as an inverse problem and use it to illustrate the GLS method.

Keywords Least squares · Prior information · Resolution · Covariance · Inverse theory · Tomography · Biconjugate gradient · Splines

1 Introduction

The principle of least squares underpins many types of geophysical data analysis, including tomography, geophysical inversion and signal processing. First stated more than 200 years

Electronic supplementary material The online version of this article (doi:[10.1007/s10712-014-9303-1](https://doi.org/10.1007/s10712-014-9303-1)) contains supplementary material, which is available to authorized users.

W. Menke (✉)
Lamont-Doherty Earth Observatory of Columbia University, 61 Route 9 W, Palisades, NY 10964,
USA
e-mail: menke@ldeo.columbia.edu

ago by Legendre (1805), numerous subsequent developments have expanded the technique, delineated its relationship to other areas of mathematics [e.g., coordinate transformations (Householder 1958)] and applied it to increasingly varied, complex and large problems.

Relatively recent developments were the recognition that data can be supplemented with prior information—expectations about the nature of the solution that are not directly linked to observations—and the methodology for solving problems that combines prior information with data in a way that accounts for the uncertainty of each (Lawson and Hanson 1974; Wiggins 1972; Tarantola and Valette 1982a, b; Menke 1984). The resulting theory, here called generalized least squares (GLS), is now central to much of the data processing that occurs in geophysics (and other fields, too). The first purpose of this paper is to review GLS and presents simplified derivations of its essential formulas. As is the case in many fields other fields, the approach of some of the seminal papers has turned out to be unnecessarily complicated.

Key ideas about the resolving power of data were developed independently of GLS, and especially through the study of continuous inverse problems (Backus and Gilbert 1968, 1970). These ideas have found fruitful application in GLS (e.g., Wiggins 1972; Yao et al. 1999), but at the same time have been the source of considerable confusion. The second purpose of this paper is to clarify the concept of resolution in problems containing prior information; that is, to rigorously define just what is being resolved.

Modern inverse problems are extremely large; datasets containing millions of observations and models containing tens of thousands of parameters are not uncommon. While GLS provides elegant formula for the resolution and covariance of an estimated model, those formulas seem, at first inspection, to be difficult to efficiently compute. As a result, the percentage of data analysis papers (at least in geophysics) that apply GLS but which omit discussion of resolution and covariance is unnecessarily large. The third purpose of this paper is to provide practical algorithms for computing them, along with sample MATLAB code.

No exposition of techniques is complete without an illustrative example. We formulate the well-understood problem of interpolating data with minimum curvature splines as an inverse problem and use it to illustrate the GLS method. This problem is chosen both because it is intuitively appealing and because it has, in the continuum limit, an analytic solution against which GLS results can be compared.

2 Review of Basis Inverse Theory Principles

2.1 Definition of the Canonical Linear Inverse Problem

We consider a linear forward problem,

$$\mathbf{d}^{\text{true}} = \mathbf{G}\mathbf{m}^{\text{true}} \text{ and } \mathbf{d}^{\text{obs}} = \mathbf{G}\mathbf{m}^{\text{true}} + \mathbf{n} \quad (2.1.1)$$

where a known $N \times M$ data kernel \mathbf{G} links noise-free true data \mathbf{d}^{true} to the true model parameters \mathbf{m}^{true} . Observational noise \mathbf{n} is always present, so that the observed data differ from the noise-free “true” data by $\mathbf{d}^{\text{obs}} = \mathbf{d}^{\text{true}} + \mathbf{n}$. The generalized inverse \mathbf{G}^{-g} turns these equations around, linking model parameters \mathbf{m} through to data \mathbf{d} to

$$\mathbf{m}^{\text{est}} = \mathbf{G}^{-g} \mathbf{d}^{\text{obs}} \text{ and } \mathbf{m}^{\text{true}} = \mathbf{G}^{-g} \mathbf{d}^{\text{obs}} - \mathbf{G}^{-g} \mathbf{n} \quad (2.1.2)$$

The generalized inverse, \mathbf{G}^{-g} , is an $M \times N$ matrix that is a function of the data kernel. Note that we have distinguished the true model \mathbf{m}^{true} , obtained by correcting the observed data for the noise \mathbf{n} , from an estimated model \mathbf{m}^{est} in which the correction is omitted. The noise is an unknown quantity, so only an estimate of the true model can be calculated; that is, $\mathbf{m}^{\text{est}} \neq \mathbf{m}^{\text{true}}$. Note that at this point, neither the method by which the generalized inverse \mathbf{G}^{-g} has been obtained its functional form has been specified. Many generalized inverses are possible, of which the GLS generalized inverse, discussed later in this paper, is but one.

2.2 Definition of Model Covariance

Observational error are assumed to be normally distributed with zero mean and with prior covariance \mathbf{C}_d . By “prior,” we mean that the covariance is assigned independently of the results of the inversion, using, say, an understanding of the limitations of the measurement process. In many instances, the errors will be statistically independent and with uniform variance σ_d^2 , in which case $\mathbf{C}_d = \sigma_d^2 \mathbf{I}$. These errors propagate through the inversion process, leading to estimated of model parameters with covariance \mathbf{C}_m . This error propagation is described by the rule for linear functions of random variables (e.g., Rektorys 1969, Sec 33.6):

$$\mathbf{C}_m = \mathbf{G}^{-g} \mathbf{C}_d \mathbf{G}^{-gT} \quad (2.2.1)$$

The variance of the model parameters is given by the diagonal elements of the covariance matrix:

$$\sigma_{m_i}^2 = [\mathbf{C}_m]_{ii} \quad (2.2.2)$$

and is typically used to state data confidence bounds for the model parameters, e.g.,

$$m_i^{\text{true}} = m_i^{\text{est}} \pm 2\sigma_{m_i} \text{ (95 \%)} \quad (2.2.3)$$

Note the distinction between the estimated model parameters, which are calculated during the inversion process and the true model parameters, which though bounded statistically, cannot be exactly known. The confidence bounds can be quite misleading in the case where the estimated model parameters are highly correlated.

An estimate of the model parameters, \mathbf{m}^{est} , has a corresponding prediction error $\mathbf{e} = \mathbf{d}^{\text{obs}} - \mathbf{G}\mathbf{m}^{\text{est}}$, where the superscript “obs” means “observed”; that is, the data measured during the experiment. The posterior variance of the error:

$$(\sigma_e^{\text{post}})^2 = \frac{\mathbf{e}^T \mathbf{e}}{\nu} \quad \text{with } \nu = (N - M) \quad (2.2.4)$$

is sometimes used as a proxy for the prior variance σ_d^2 , at least in cases where the data are believed to be uncorrelated and with uniform variance. Note, however, that this formula assumes that the estimated model is close to the true model, so that the error can be attributed solely to noise in the data—an assumption that is not always justified.

The ratio, $\rho^2 = (\sigma_e^{\text{post}})^2 / \sigma_d^2$, is a measure of how well the data are fitted by an estimated model. Models for which $\rho^2 \approx 1$ fit the data acceptably well and models for which $\rho^2 \gg 1$ fit them poorly. This notion can be quantified by recognizing that the quantity $X^2 = \nu\rho^2$ is chi-squared distributed with ν degrees of freedom and then using a standard chi-squared

test of significance. The chi-squared distribution is approximately normal for large ν with mean ν and variance 2ν , so in that case an unacceptably poor fit (at 95 % confidence) is one for which $\rho^2 > 1 + 2\sqrt{2/\nu}$. Models for which $\rho^2 \ll 1$ *over-fit* the data, which usually means that the model contains features whose presence is not justified by data with that noise level. At 95 % confidence, over-fit data satisfy $\rho^2 < 1 - 2\sqrt{2/\nu}$ (in the large ν limit).

2.3 Simple Least Squares

The simple least squares solution for data with covariance \mathbf{C}_d is obtained by minimizing the prediction error:

$$\Phi_{\text{SLS}} = [\mathbf{d}^{\text{obs}} - \mathbf{G}\mathbf{m}]^T \mathbf{C}_d^{-1} [\mathbf{d}^{\text{obs}} - \mathbf{G}\mathbf{m}] \quad (2.3.1)$$

with respect to model parameters \mathbf{m} . This formula can be thought of as the “sum of squared prediction errors, with each error weighted by the *certainty* of the corresponding observation” (since variance quantifies uncertainty, its reciprocal quantifies certainty). The minimization of Φ_{SLS} yields an estimate for the model parameters (e.g., Lawson and Hanson 1974):

$$\mathbf{m}^{\text{est}} = [\mathbf{G}^T \mathbf{C}_d^{-1} \mathbf{G}]^{-1} \mathbf{G}^T \mathbf{C}_d^{-1} \mathbf{d}^{\text{obs}} \quad (2.3.2)$$

The generalized inverse is $\mathbf{G}^{-g} = [\mathbf{G}^T \mathbf{C}_d^{-1} \mathbf{G}]^{-1} \mathbf{G}^T \mathbf{C}_d^{-1}$. Note that in the case of uncorrelated data with uniform variance, $\mathbf{C}_d = \sigma_d^2 \mathbf{I}$, this formula reduces to $\mathbf{G}^{-g} = [\mathbf{G}^T \mathbf{G}]^{-1} \mathbf{G}^T$. One of the limitations of simple least squares is that this generalized inverse exists only when the observations are sufficient to uniquely specify a solution; else the matrix $[\mathbf{G}^T \mathbf{C}_d^{-1} \mathbf{G}]^{-1}$ does not exist. As discussed below, one of the purposes of GLS is to overcome this limitation.

In simple least squares, the covariance of the model parameters is:

$$\mathbf{C}_m = \mathbf{G}^{-g} \mathbf{C}_d \mathbf{G}^{-gT} = [\mathbf{G}^T \mathbf{C}_d^{-1} \mathbf{G}]^{-1} \mathbf{G}^T \mathbf{C}_d^{-1} \mathbf{C}_d \mathbf{C}_d^{-1} \mathbf{G} [\mathbf{G}^T \mathbf{C}_d^{-1} \mathbf{G}]^{-1} = [\mathbf{G}^T \mathbf{C}_d^{-1} \mathbf{G}]^{-1} \quad (2.3.3)$$

In general, the model parameters will be correlated and of unequal variance even when the data are independent and with uniform variance:

$$\mathbf{C}_m = \sigma_d^2 [\mathbf{G}^T \mathbf{G}]^{-1} \quad \text{when } \mathbf{C}_d = \sigma_d^2 \mathbf{I} \quad (2.3.4)$$

2.4 Definition of Model Resolution

The model resolution matrix $\mathbf{R}^G = \mathbf{G}^{-g} \mathbf{G}$ can be obtained using the fact that an asserted model \mathbf{m}^{ass} (that is, a hypothetical model put forward for discussion purposes) predicts data, $\mathbf{d}^{\text{pre}} = \mathbf{G}\mathbf{m}^{\text{ass}}$ and model \mathbf{m}^{rec} recovered from inverting those data is $\mathbf{m}^{\text{rec}} = \mathbf{G}^{-g} \mathbf{d}^{\text{pre}}$ (Wiggins 1972):

$$\mathbf{m}^{\text{rec}} = \mathbf{G}^{-g} \mathbf{G}\mathbf{m}^{\text{ass}} = \mathbf{R}^G \mathbf{m}^{\text{ass}} \quad (2.4.1)$$

The resolution matrix \mathbf{R}^G indicates that the recovered model parameters only equal the asserted model parameters in the special case where $\mathbf{R}^G = \mathbf{I}$. This notion can be extended

to the true and estimated model by including the effect of noise \mathbf{n} (Friedel 2003; Gunther 2004):

$$\mathbf{m}^{\text{est}} = \mathbf{G}^{-g} \mathbf{d}^{\text{obs}} = \mathbf{G}^{-g} (\mathbf{G} \mathbf{m}^{\text{true}} + \mathbf{n}) = \mathbf{R}^G \mathbf{m}^{\text{true}} + \mathbf{G}^{-g} \mathbf{n} \quad (2.4.2)$$

The estimated and true models are again related via the resolution matrix \mathbf{R}^G , up to a correction factor involving the noise. Unfortunately, this correction factor cannot be calculated, since the noise is unknown. In subsequent discussion, we will acknowledge this limitation by writing $\mathbf{m}^{\text{est}} \approx \mathbf{R}^G \mathbf{m}^{\text{true}}$, which is to say, focusing on the way in which \mathbf{m}^{est} and \mathbf{m}^{true} are related when the effect of noise is negligible.

In typical cases, the estimated model parameters are linear combinations (“weighted averages”) of the true model parameters. In general, \mathbf{R}^G has no special symmetry; it is neither symmetric nor antisymmetric. We note for future reference that the simple least squares solution, when it exists, has perfect resolution:

$$\mathbf{R}^G = \mathbf{G}^{-g} \mathbf{G} = [\mathbf{G}^T \mathbf{C}_d^{-1} \mathbf{G}]^{-1} \mathbf{G}^T \mathbf{C}_d^{-1} \mathbf{G} = \mathbf{I} \quad (2.4.3)$$

2.5 Meaning of the k -th Row of the Resolution Matrix

The k -th estimated model parameter satisfies:

$$m_k^{\text{est}} \approx \sum_i R_{ki}^G m_i^{\text{true}} \quad (2.5.1)$$

and so can be interpreted as being equal to a linear combination of the true model parameters, where the coefficients are given by the elements of the k -th row of the resolution matrix. Colloquially, we might speak of the estimated model parameters as *weighted averages* of the true model parameters. However, strictly speaking, they are only true weighted averages when the elements of the row are positive and sum to unity,

$$R_{ki}^G \geq 0 \text{ and } \sum_i R_{ki}^G = [1]_k \quad (2.5.2)$$

which are, in general, not the case.

2.6 Meaning of the k -th Column of the Resolution Matrix

The k -th column of the resolution matrix specifies how each of the estimated model parameters is influenced by the k -th true model parameter. This can be seen by setting $\mathbf{m}^{\text{true}} = \mathbf{s}^{(k)}$ with $s_i^{(k)} = \delta_{ik}$; that is, all the true model parameters are zero except the k -th, which is unity. Denoting the set of estimated model parameters associated with $\mathbf{s}^{(k)}$ as $\mathbf{m}^{\text{est}(k)}$, we have (Menke 2012):

$$\mathbf{m}^{\text{est}(k)} \approx \mathbf{R}^G \mathbf{s}^{(k)} \text{ or } m_j^{\text{est}(k)} \approx \sum_i R_{ji}^G \delta_{ik} = R_{jk}^G \quad (2.6.1)$$

Thus, the k -th column of the resolution matrix is analogous to the *point-spread function* encountered in image processing (e.g., Smith 1997, Chapter 24); that is, a single true model parameter spreads out into many estimated model parameters.

If instead of being set to a spike, \mathbf{m}^{true} is set some other pattern, then the resulting \mathbf{m}^{est} provides information on how well that pattern can be recovered. When the model

represents the discrete version of a function $m(x, y)$ of two spatial variables (x, y) , a common choice for the pattern is the *checkboard* (an alternating pattern of positive and negative fluctuations) and the procedure is called a *checkerboard test*.

2.7 Spread Resolution and the Size of Covariance Trade-Off

Spread of resolution can be quantified by the degree of departure of \mathbf{R}^G from an identity matrix and size of \mathbf{C}_m by the magnitude of its main diagonal, which represents variance. A general principle of inverse theory is that resolution trades off with variance (Backus and Gilbert 1970). A solution with small spread of resolution tends to have large variance and vice versa. Many inverse methods (including GLS) have a tunable parameter that defines a “trade-off curve” of allowable combinations of spread and size. Users can then select a combination of resolution and variance that is optimum for their particular use.

3 Generalized Least Squares (GLS)

3.1 Definition of Prior Information

Generalized least squares (Lawson and Hanson 1974; Wiggins 1972; Tarantola and Valette 1982a, b; Menke 1984; see also Menke 2012) improves upon simple least squares by supplementing the observations with prior information, represented by the linear equation $\mathbf{H}_0\mathbf{m} = \mathbf{h}_0^{\text{pri}}$, where the superscript “pri” indicates prior. This equation, assumed to be determined independently from any actual observations, encodes prior expectations about the behavior of the model parameters. Many (but not all) classes of prior information can be represented by judiciously choosing the matrix \mathbf{H}_0 and vector $\mathbf{h}_0^{\text{pri}}$; for example, the model parameters have specific values, $\langle \mathbf{m} \rangle$

$$\mathbf{H}_0 = \mathbf{I} \text{ and } \mathbf{h}_0^{\text{pri}} = \langle \mathbf{m} \rangle$$

the average of the model parameters have a specific value, $\langle \mathbf{m} \rangle$

$$\mathbf{H}_0 = M^{-1} [1 \quad 1 \quad \cdots \quad 1] \text{ and } \mathbf{h}_0^{\text{pri}} = \langle \mathbf{m} \rangle$$

the model parameters are flat

$$\mathbf{H}_0 \text{ with rows like } [0 \quad 0 \quad \cdots \quad 0 \quad 1 \quad -2 \quad 1 \quad 0 \quad \cdots \quad 0] \text{ and } \mathbf{h}_0^{\text{pri}} = 0$$

and the model parameters are smooth

$$\mathbf{H}_0 \text{ with rows like } [0 \quad 0 \quad \cdots \quad 0 \quad 1 \quad -2 \quad 1 \quad 0 \quad \cdots \quad 0] \text{ and } \mathbf{h}_0^{\text{pri}} = 0 \quad (3.1.1)$$

and so forth. The accuracy of the prior information is described by a covariance matrix \mathbf{C}_{h0} that represents the quality of the underlying expectations.

In many cases, the prior information will be uncorrelated, implying that \mathbf{C}_{h0} is a diagonal matrix [but see Abers (1994) for an interesting counterexample]. In many cases, all the prior information will also be equally uncertain, in which case $\mathbf{C}_{h0} = \sigma_h^2 \mathbf{I}$, where σ_h^2 is the variance of the prior information.

The case of a non-uniform prior variance has wide uses, because it can be used to assert that one part of the model obeys prior information, like $m_i \approx \langle m_i \rangle$, more strongly than others. This allows a kind of hypothesis testing that is called *squeezing* (Lerner-lam and Jordan 1987). The presence of a certain feature in the model is tested by designing prior information that asserts that the feature is *absent* and assigning that feature—and that feature only—low variance. The feature will be suppressed (squeezed) in a solution that includes this prior information, relative to one that does not. The feature is only accepted as significant if the posterior variance σ_e^2 of the error (see 2.2.1) is significantly more for the squeezed model than the unsqueezed model. The ratio $F = (\sigma_e^2)^{\text{squeezed}} / (\sigma_e^2)^{\text{unsqueezed}}$ is F -distributed with $\nu_1 = \nu_2 = N - M$ degrees of freedom, so significance can be assessed with a standard F test.

In general, the prior information needs to be neither consistent nor sufficient to uniquely determine the model parameters. However, it is always possible to add additional—but very weak—information to uniquely determine a set or prior model parameters, say \mathbf{m}^H that can be used for reference, that is, allowing us to answer the question of how much the data changed our preconceptions about the model. We take the approach of augmenting $\mathbf{H}_0\mathbf{m} = \mathbf{h}_0^{\text{pri}}$ to

$$\mathbf{H}\mathbf{m} = \mathbf{h}^{\text{pri}} = \begin{bmatrix} \mathbf{H}_0 \\ \mathbf{I} \end{bmatrix} \mathbf{m} = \begin{bmatrix} \mathbf{h}_0^{\text{pri}} \\ 0 \end{bmatrix} \quad (3.1.2)$$

This modification adds the information that the model parameters are close to zero. When assigned high variance, it will force to zero linear combinations of model parameters that are not resolved by $\mathbf{H}_0\mathbf{m} = \mathbf{h}_0^{\text{pri}}$ while having negligible effect on the others. We define the corresponding covariance to be

$$\mathbf{C}_h = \begin{bmatrix} \mathbf{C}_{h0} & 0 \\ 0 & \varepsilon^{-2}\mathbf{I} \end{bmatrix} \quad (3.1.3)$$

Here, ε^{-2} is the variance of the additional information, which is presumed very large (implying that ε is very small). We can then define the reference model \mathbf{m}^H to be the simple least squares solution to the augmented system:

$$\mathbf{m}^H = [\mathbf{H}^T\mathbf{C}_h^{-1}\mathbf{H}]^{-1}\mathbf{H}^T\mathbf{C}_h^{-1}\mathbf{h}^{\text{pri}} = [\mathbf{H}_0^T\mathbf{C}_{h0}^{-1}\mathbf{H}_0 + \varepsilon^2\mathbf{I}]^{-1}\mathbf{H}_0^T\mathbf{C}_{h0}^{-1}\mathbf{h}_0^{\text{pri}} \quad (3.1.4)$$

We will call \mathbf{m}^H the *prior model*, for it is the one predicted by the prior information, acting alone. We can define the data predicted by the prior information as:

$$\mathbf{d}^H = \mathbf{G}\mathbf{m}^H \quad (3.1.5)$$

3.2 The Generalized Least Squares Solution

The GLS is solution obtained by minimizing the generalized error, that is, the sum of the simple least squares prediction error Φ_{SLS} and the error in prior information, $[\mathbf{h} - \mathbf{H}\mathbf{m}]^T\mathbf{C}_h^{-1}[\mathbf{h} - \mathbf{H}\mathbf{m}]$ (Tarantola and Valette 1982a, b):

$$\Phi_{\text{GLS}} = [\mathbf{d}^{\text{obs}} - \mathbf{G}\mathbf{m}]^T\mathbf{C}_d^{-1}[\mathbf{d}^{\text{obs}} - \mathbf{G}\mathbf{m}] + [\mathbf{h}^{\text{pri}} - \mathbf{H}\mathbf{m}]^T\mathbf{C}_h^{-1}[\mathbf{h}^{\text{pri}} - \mathbf{H}\mathbf{m}] \quad (3.2.1)$$

Note that no cross-terms appear in this equation. We have assumed that observational errors do not correlate with errors in the prior information. By defining:

$$\mathbf{F} = \begin{bmatrix} \mathbf{C}_d^{-1/2} \mathbf{G} \\ \mathbf{C}_h^{-1/2} \mathbf{H} \end{bmatrix} \text{ and } \mathbf{f}^{\text{obs}} = \begin{bmatrix} \mathbf{C}_d^{-1/2} \mathbf{d}^{\text{obs}} \\ \mathbf{C}_h^{-1/2} \mathbf{h}^{\text{pri}} \end{bmatrix}, \tag{3.2.2}$$

we can write $\Phi_{\text{GLS}} = [\mathbf{f}^{\text{obs}} - \mathbf{Fm}]^T \mathbf{C}_f^{-1} [\mathbf{f}^{\text{obs}} - \mathbf{Fm}]$, with $\mathbf{C}_f = \mathbf{I}$, which is in the form of a simple least squares minimization problem. Note that the covariance of \mathbf{f}^{obs} is, indeed, $\mathbf{C}_f = \mathbf{I}$; the data and prior information have been weighted so as to produce uncorrelated random variables with unit variance. Here, \mathbf{d}^{obs} denotes the observed values of the data and \mathbf{h}^{pri} the prior values of the information (that is, the values that are asserted). The combined vector \mathbf{f}^{obs} includes both observations and prior information, but we simplify its superscript to “obs.”

The solution is given by the simple least squares formula:

$$\mathbf{m}^{\text{est}} = [\mathbf{F}^T \mathbf{F}]^{-1} \mathbf{F}^T \mathbf{f}^{\text{obs}} \tag{3.2.3}$$

or

$$\mathbf{m}^{\text{est}} = \mathbf{G}^{-g} \mathbf{d}^{\text{obs}} + \mathbf{H}^{-g} \mathbf{h}^{\text{pri}}$$

with $\mathbf{G}^{-g} = \mathbf{A}^{-1} \mathbf{G}^T \mathbf{C}_d^{-1}$ and $\mathbf{H}^{-g} = \mathbf{A}^{-1} \mathbf{H}^T \mathbf{C}_h^{-1}$ and $\mathbf{A} = \mathbf{G}^T \mathbf{C}_d^{-1} \mathbf{G} + \mathbf{H}^T \mathbf{C}_h^{-1} \mathbf{H}$ (3.2.4)

The presumption in GLS is that the addition of prior information to the problem is sufficient to eliminate any non-uniqueness that would have been present had only observations been used. Thus, the inverse of \mathbf{A} is presumed to exist. Note that since \mathbf{A} is symmetric, its inverse \mathbf{A}^{-1} will also be symmetric.

3.3 Variance of Generalized Least Squares

The standard formula for error propagation gives:

$$\mathbf{C}_m = \mathbf{F}^{-g} \mathbf{C}_f \mathbf{F}^{-gT} = [\mathbf{F}^T \mathbf{F}]^{-1} \mathbf{F}^T \mathbf{I} \mathbf{F} [\mathbf{F}^T \mathbf{F}]^{-1} = [\mathbf{F}^T \mathbf{F}]^{-1} = \mathbf{A}^{-1} \tag{3.3.1}$$

Note that this covariance depends on both the prior covariance of the data, \mathbf{C}_d , and the variance of the prior information, \mathbf{C}_h . We can identify the contribution of the two sources of error as:

$$\mathbf{C}_m = \mathbf{G}^{-g} \mathbf{C}_d \mathbf{G}^{-gT} + \mathbf{H}^{-g} \mathbf{C}_h \mathbf{H}^{-gT} \tag{3.3.2}$$

since

$$\begin{aligned} \mathbf{C}_m &= \mathbf{G}^{-g} \mathbf{C}_d \mathbf{G}^{-gT} + \mathbf{H}^{-g} \mathbf{C}_h \mathbf{H}^{-gT} = \mathbf{A}^{-1} \mathbf{G}^T \mathbf{C}_d^{-1} \mathbf{C}_d \mathbf{C}_d^{-1} \mathbf{G} \mathbf{A}^{-1} + \mathbf{A}^{-1} \mathbf{H}^T \mathbf{C}_h^{-1} \mathbf{C}_h \mathbf{C}_h^{-1} \mathbf{H} \mathbf{A}^{-1} \\ &= \mathbf{A}^{-1} \mathbf{G}^T \mathbf{C}_d^{-1} \mathbf{G} \mathbf{A}^{-1} + \mathbf{A}^{-1} \mathbf{H}^T \mathbf{C}_h^{-1} \mathbf{H} \mathbf{A}^{-1} = \mathbf{A}^{-1} [\mathbf{G}^T \mathbf{C}_d^{-1} \mathbf{G} + \mathbf{H}^T \mathbf{C}_h^{-1} \mathbf{H}] \mathbf{A}^{-1} \\ &= \mathbf{A}^{-1} \mathbf{A} \mathbf{A}^{-1} = \mathbf{A}^{-1} \end{aligned} \tag{3.3.3}$$

Thus, the covariance of the model parameters consists of the sum of a term, $\mathbf{G}^{-g} \mathbf{C}_d \mathbf{G}^{-gT}$, which is identical in form (but not in value) to the one encountered in simple least squares, and an analogous term arising from the prior information. In the limit of $\|\mathbf{C}_d\| \rightarrow \infty$ (very noisy data), \mathbf{C}_m depends only upon \mathbf{C}_h , and in the limit of $\|\mathbf{C}_h\| \rightarrow \infty$ (very weak information), \mathbf{C}_m depends only upon \mathbf{C}_d .

Since the predicted data are linear functions of the estimated model parameters through the equation, $\mathbf{d}^{\text{pre}} = \mathbf{G}\mathbf{m}^{\text{est}}$, their covariance is given by:

$$\mathbf{C}_{d^{\text{pre}}} = \mathbf{G}\mathbf{C}_m\mathbf{G}^T = \mathbf{G}\mathbf{A}^{-1}\mathbf{G}^T \tag{3.3.4}$$

3.4 Resolution of Generalized Least Squares

Generalized least squares does not distinguish the weighted data equation $\mathbf{C}_d^{-1/2}\{\mathbf{G}\mathbf{m} = \mathbf{d}^{\text{obs}}\}$ from the weighted prior information equation $\mathbf{C}_h^{-1/2}\{\mathbf{H}\mathbf{m} = \mathbf{h}^{\text{pri}}\}$; the latter is simply appended to the bottom of the former to create the combined equation $\mathbf{F}\mathbf{m} = \mathbf{f}^{\text{obs}}$. Consequently, in analogy to the simple least squares case, we can define a generalized inverse \mathbf{F}^{-g} and a resolution matrix \mathbf{R}^F as:

$$\begin{aligned} \mathbf{F}^{-g} &= [\mathbf{F}^T\mathbf{F}]^{-1}\mathbf{F}^T \quad \text{so that } \mathbf{m}^{\text{est}} = \mathbf{F}^{-g}\mathbf{f}^{\text{pre}} \\ \mathbf{R}^F &= \mathbf{F}^{-g}\mathbf{F} \quad \text{so that } \mathbf{m}^{\text{est}} = \mathbf{R}^F\mathbf{m}^{\text{true}} + \mathbf{F}^{-g}\mathbf{n} \text{ and } \mathbf{m}^{\text{rec}} = \mathbf{R}^F\mathbf{m}^{\text{ass}} \end{aligned} \tag{3.4.1}$$

However, when defined in this way, the resolution of GLS is perfect, since

$$\mathbf{R}^F = \mathbf{F}^{-g}\mathbf{F} = [\mathbf{F}^T\mathbf{F}]^{-1}\mathbf{F}^T\mathbf{F} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I} \tag{3.4.2}$$

In general, the estimated model parameters depend upon both \mathbf{d}^{obs} and \mathbf{h}^{pri} ; that is, $\mathbf{m}^{\text{est}} = \mathbf{G}^{-g}\mathbf{d}^{\text{obs}} + \mathbf{H}^{-g}\mathbf{h}^{\text{pri}}$. Consider for the moment the special case of $\mathbf{h}^{\text{pri}} = 0$ (we will relax this requirement in the next section). This case commonly arises in practice, e.g., for the prior information of smoothness. The estimated model parameters depend only upon \mathbf{d}^{obs} ; that is, $\mathbf{m}^{\text{est}} = \mathbf{G}^{-g}\mathbf{d}^{\text{obs}} + 0$. We can use the forward equation to predict data associated with an asserted model parameters, $\mathbf{d}^{\text{pre}} = \mathbf{G}\mathbf{m}^{\text{ass}}$, and then invert these predictions back recovered model parameters, $\mathbf{m}^{\text{rec}} = \mathbf{G}^{-g}\mathbf{d}^{\text{pre}} + 0$. Hence, we obtain the usual formula for resolution:

$$\mathbf{m}^{\text{rec}} = \mathbf{R}^G\mathbf{m}^{\text{ass}} \quad \text{with } \mathbf{R}^G = \mathbf{G}^{-g}\mathbf{G} \tag{3.4.3}$$

Superficially, we seemed to have achieved contradictory results, as the two resolution matrices have radically different properties:

$$\mathbf{R}^F = \mathbf{I} \quad \text{and } \mathbf{R}^G \neq \mathbf{I} \tag{3.4.4}$$

However, one step in the derivations is critically different. During the derivation of \mathbf{R}^G , we asserted that $\mathbf{h}^{\text{pri}} = 0$, even though an arbitrary \mathbf{m}^{ass} predicts $\mathbf{h}^{\text{pre}} = \mathbf{H}\mathbf{m}^{\text{ass}} \neq 0$. During the derivation of \mathbf{R}^F , we made no such assertion; the \mathbf{h}^{pre} imbedded in \mathbf{f}^{pre} arises from $\mathbf{h}^{\text{rec}} = \mathbf{H}\mathbf{m}^{\text{ass}}$ and is not equal to zero.

That the \mathbf{R}^G and not \mathbf{R}^F is the proper definition of resolution can be understood from the following scenario. Suppose that the model \mathbf{m} represents a discrete version of a continuous function $m(x)$ and that one is trying to find an \mathbf{m}^{est} that approximately satisfies $\mathbf{G}\mathbf{m} = \mathbf{d}^{\text{obs}}$ but is smooth. Smoothness is the opposite of roughness, and the roughness of a function can be quantified by the mean-squared value of its second derivative. Thus, we take \mathbf{H}_0 to be the second-derivative operator (i.e., with rows like $[0 \ \cdots \ 0 \ 1 \ -2 \ 1 \ 0 \ \cdots \ 0]$) and $\mathbf{h}^{\text{pri}} = 0$, which leads to the minimization of $[\mathbf{H}_0\mathbf{m}]^T[\mathbf{H}_0\mathbf{m}]$, a quantity proportional to the r.m.s. average of the second derivative. Now suppose that the asserted solution is the spike $\mathbf{m}^{\text{ass}} = \mathbf{s}^{(k)}$ (that is, zero except for the k -th

element, which is unity). We want to know how this spike spreads out during the inversion process, presuming that an experiment produced the data $\mathbf{d}^{\text{pre}} = \mathbf{G}\mathbf{s}^{(k)}$ that this model predicts. What values should one use for \mathbf{h} in such an inversion? The model predicts $\mathbf{h}^{\text{pre}} = \mathbf{H}_0\mathbf{s}^{(k)}$, but these are the actual values of the second derivative. To use them in the inversion would be to assert that second derivatives are known—much stronger information than merely the assertion that their mean-squared average is small. One should, therefore, use $\mathbf{h}^{\text{pri}} = 0$, which leads to a solution that is a column of \mathbf{R}^{G} , not \mathbf{R}^{F} .

So far, our discussion has been limited to the special case of $\mathbf{h}^{\text{pri}} = 0$. We now relax that condition (Kalscheuer 2008; Kalscheuer et al. 2010), but still require that the prior information is complete (as above), so that it implies a specific model, $\mathbf{m}^{\text{H}} = [\mathbf{H}^{\text{T}}\mathbf{C}_h^{-1}\mathbf{H}]^{-1}\mathbf{H}^{\text{T}}\mathbf{C}_h^{-1}\mathbf{h}^{\text{pri}}$. We now use the prior model \mathbf{m}^{H} as a reference model, defining the deviation of a given model from it as $\Delta\mathbf{m} = \mathbf{m} - \mathbf{m}^{\text{H}}$. The GLS solution can be rewritten in terms of this deviation:

$$\begin{aligned} \Delta\mathbf{m}^{\text{est}} &= \mathbf{m}^{\text{est}} - \mathbf{m}^{\text{H}} = \mathbf{A}^{-1}(\mathbf{G}^{\text{T}}\mathbf{C}_d^{-1}\mathbf{d}^{\text{obs}} + \mathbf{H}^{\text{T}}\mathbf{C}_h^{-1}\mathbf{h}^{\text{pri}}) - \mathbf{m}^{\text{H}} \\ &= \mathbf{A}^{-1}(\mathbf{G}^{\text{T}}\mathbf{C}_d^{-1}\mathbf{d}^{\text{obs}} + \mathbf{H}^{\text{T}}\mathbf{C}_h^{-1}\mathbf{h}^{\text{pri}}) - \mathbf{A}^{-1}\mathbf{A}\mathbf{m}^{\text{H}} \\ &= \mathbf{A}^{-1}(\mathbf{G}^{\text{T}}\mathbf{C}_d^{-1}\mathbf{d}^{\text{obs}} + \mathbf{H}^{\text{T}}\mathbf{C}_h^{-1}\mathbf{h}^{\text{pri}} - \mathbf{A}\mathbf{m}^{\text{H}}) \\ &= \mathbf{A}^{-1}(\mathbf{G}^{\text{T}}\mathbf{C}_d^{-1}\mathbf{d}^{\text{obs}} + \mathbf{H}^{\text{T}}\mathbf{C}_h^{-1}\mathbf{h}^{\text{pri}} - \mathbf{G}^{\text{T}}\mathbf{C}_d^{-1}\mathbf{G}\mathbf{m}^{\text{H}} - \mathbf{H}^{\text{T}}\mathbf{C}_h^{-1}\mathbf{H}[\mathbf{H}^{\text{T}}\mathbf{C}_h^{-1}\mathbf{H}]^{-1}\mathbf{H}^{\text{T}}\mathbf{C}_h^{-1}\mathbf{h}^{\text{pri}}) \\ &= \mathbf{A}^{-1}(\mathbf{G}^{\text{T}}\mathbf{C}_d^{-1}\mathbf{d}^{\text{obs}} + \mathbf{H}^{\text{T}}\mathbf{C}_h^{-1}\mathbf{h}^{\text{pri}} - \mathbf{G}^{\text{T}}\mathbf{C}_d^{-1}\mathbf{G}\mathbf{m}^{\text{H}} - \mathbf{H}^{\text{T}}\mathbf{C}_h^{-1}\mathbf{h}^{\text{pri}}) \\ &= \mathbf{G}^{-\text{g}}(\mathbf{d}^{\text{obs}} - \mathbf{G}\mathbf{m}^{\text{H}}) = \mathbf{G}^{-\text{g}}(\mathbf{d}^{\text{obs}} - \mathbf{d}^{\text{H}}) \end{aligned} \tag{3.4.5}$$

Thus, the deviation of the model from \mathbf{m}^{H} depends only on the deviation of the data from those predicted by \mathbf{m}^{H} :

$$\Delta\mathbf{m} = \mathbf{G}^{-\text{g}}\Delta\mathbf{d} \text{ with } \Delta\mathbf{m} = \mathbf{m} - \mathbf{m}^{\text{H}} \text{ and } \Delta\mathbf{d} = \mathbf{d} - \mathbf{d}^{\text{H}} \tag{3.4.6}$$

and furthermore

$$\mathbf{G}\Delta\mathbf{m} = \Delta\mathbf{d} \text{ since } \mathbf{G}\Delta\mathbf{m} = \mathbf{G}(\mathbf{m} - \mathbf{m}^{\text{H}}) = \mathbf{G}\mathbf{m} - \mathbf{G}\mathbf{m}^{\text{H}} = \mathbf{d} - \mathbf{d}^{\text{H}} = \Delta\mathbf{d} \tag{3.4.7}$$

Once again, we can combine $\Delta\mathbf{d}^{\text{pre}} = \mathbf{G}\Delta\mathbf{m}^{\text{ass}}$ with $\Delta\mathbf{m}^{\text{rec}} = \mathbf{G}^{-\text{g}}\Delta\mathbf{d}^{\text{pre}}$ into the usual statement about resolution,

$$\Delta\mathbf{m}^{\text{rec}} = \mathbf{R}^{\text{G}}\Delta\mathbf{m}^{\text{ass}} \text{ with } \mathbf{R}^{\text{G}} = \mathbf{G}^{-\text{g}}\mathbf{G} \tag{3.4.8}$$

In this case, too, \mathbf{R}^{G} is the correct choice for quantifying resolution. However, the quantity being resolved is the deviation of the model from the reference model, \mathbf{m}^{H} , and not the model itself. The distinction, while of minor significance in cases where \mathbf{m}^{H} has a simple shape, is more important when \mathbf{m}^{H} is complicated.

3.5 Linearized GLS

In many cases, the relationship between data and model is nonlinear:

$$d_i = g_i(\mathbf{m}) + n_i \tag{3.5.1}$$

Here, \mathbf{n} represents observational noise. This equation can be compared to the linear case in (2.1.1). A common approach is to use Taylor's theorem to linearize this equation about a trial solution $\mathbf{m}^{(p)}$:

$$d_i = g(\mathbf{m}^{(p)}) + \sum_j \left. \frac{\partial g_i}{\partial m_j} \right|_{\mathbf{m}^{(p)}} (m_j - m_j^{(p)}) + n_i \text{ or } \delta \mathbf{d}^{(p)} = \mathbf{G}^{(p)} \delta \mathbf{m} + \mathbf{n}$$

with

$$\delta \mathbf{d}^{(p)} = \mathbf{d} - \mathbf{g}(\mathbf{m}^{(p)}) \text{ and } \delta \mathbf{m}^{(p)} = \mathbf{m} - \mathbf{m}^{(p)} \text{ and } \left[\mathbf{G}^{(p)} \right]_{ij} = \left. \frac{\partial g_i}{\partial m_j} \right|_{\mathbf{m}^{(p)}} \quad (3.5.2)$$

Here, $\delta \mathbf{m}$ is the perturbation of the model parameters from the trial solution $\mathbf{m}^{(p)}$, and $\delta \mathbf{d}$ is the perturbation of the data from those predicted by the trial solution. This approach leads to a standard linear equation of the form $\delta \mathbf{d} = \mathbf{G}^{(p)} \delta \mathbf{m}$. We now need to combine this equation with prior information. We first write the prior information equation, $\mathbf{Hm} = \mathbf{h}$ (Sect. 3.1) in terms of perturbations:

$$\mathbf{Hm} = \mathbf{h} \text{ becomes } \mathbf{H}(\mathbf{m}^{(p)} + \delta \mathbf{m}^{(p)}) = \mathbf{h} \text{ becomes } \mathbf{H} \delta \mathbf{m}^{(p)} = \mathbf{h} - \mathbf{Hm}^{(p)} \quad (3.5.3)$$

The GLS Eq. (3.2.2) then becomes:

$$\begin{aligned} \mathbf{F}^{(p)} \delta \mathbf{m}^{(p)} &= \mathbf{f}^{(p)} \text{ with} \\ \mathbf{F}^{(p)} &= \begin{bmatrix} \mathbf{C}_d^{-1/2} \mathbf{G}^{(p)} \\ \mathbf{C}_h^{-1/2} \mathbf{H} \end{bmatrix} \text{ and } \mathbf{f}^{(p)} = \begin{bmatrix} \mathbf{C}_d^{-1/2} \delta \mathbf{d}^{(p)} \\ \mathbf{C}_h^{-1/2} \delta \mathbf{h}^{(p)} \end{bmatrix} \\ \text{and } \delta \mathbf{d}^{(p)} &= \mathbf{d}^{\text{obs}} - \mathbf{g}(\mathbf{m}^{(p)}) \text{ and } \delta \mathbf{h}^{(p)} = \mathbf{h}^{\text{pri}} - \mathbf{Hm}^{(p)} \end{aligned} \quad (3.5.4)$$

Note that the quantity $\delta \mathbf{h}^{(p)}$ represents the perturbation of the prior information from that predicted by the trial model. An initial solution $\mathbf{m}^{(0)}$ can be iterated to produce a sequence of solutions, $\mathbf{m}^{(1)} = \mathbf{m}^{(0)} + \delta \mathbf{m}^{(0)}$, $\mathbf{m}^{(2)} = \mathbf{m}^{(1)} + \delta \mathbf{m}^{(1)}$..., which under favorable circumstances will converge to the solution, \mathbf{m}^{est} , that minimizes the generalized error Φ_{GLS} .

After any iteration, the covariance of $\delta \mathbf{m}^{(p)}$ can be calculated as $\mathbf{C}_m = [\mathbf{F}^{(p)T} \mathbf{F}^{(p)}]^{-1}$ (see 3.3.1). Since $\mathbf{m}^{(p+1)} = \mathbf{m}^{(p)} + \delta \mathbf{m}^{(p)}$ and $\mathbf{m}^{(p)}$ is a constant, it is also the covariance of $\mathbf{m}^{(p+1)}$. Its value after the final iteration can be used as an estimate of the covariance of \mathbf{m}^{est} . However, this estimate must be used cautiously because it is based on a linear approximation.

The issue of resolution is treated exactly parallel to its handling in the linear problem. The linearized Eq. (3.5.4) has exactly the same form as the original linear version (3.2.2), except that \mathbf{m} is replaced with $\delta \mathbf{m}^{(p)}$, \mathbf{d}^{obs} is replaced with $\delta \mathbf{d}^{(p)}$ and \mathbf{h}^{pri} is replaced with $\delta \mathbf{h}^{(p)}$.

Therefore, we can arrive at the proper formula for resolution by making these substitutions in the results of the original linear derivation. In analogy to the linear case, we define quantities:

$$\delta \mathbf{m}^{\text{H}(p)} = [\mathbf{H}^T \mathbf{C}_h^{-1} \mathbf{H}]^{-1} \mathbf{H}^T \mathbf{C}_h^{-1} \delta \mathbf{h}^{(p)} = \mathbf{m}^{\text{H}} - \mathbf{m}^{(p)} \text{ and } \delta \mathbf{d}^{\text{H}(p)} = \mathbf{G}^{(p)} \delta \mathbf{m}^{\text{H}(p)} \quad (3.5.5)$$

The first equation represents the perturbation of the prior model from the trial model. The second equation represents the data perturbation predicted by that model perturbation. In further analogy to the linear case, we define the deviations:

$$\Delta\delta\mathbf{m}^{(p)} = \delta\mathbf{m}^{(p)} - \delta\mathbf{m}^{H(p)} \text{ and } \Delta\delta\mathbf{d}^{(p)} = \delta\mathbf{d}^{(p)} - \delta\mathbf{d}^{H(p)} \tag{3.5.6}$$

As in the linear case (3.4.7), $\Delta\delta\mathbf{m}^{(p)}$ is predicted from $\Delta\delta\mathbf{d}^{(p)}$ via a generalized inverse analogous to (3.2.4):

$$\begin{aligned} \Delta\delta\mathbf{m}^{(p)} &= [\mathbf{G}^{(p)}]^{-g} \Delta\delta\mathbf{d}^{(p)} \text{ with} \\ [\mathbf{G}^{(p)}]^{-g} &= [\mathbf{G}^{(p)T} \mathbf{C}_d^{-1} \mathbf{G}^{(p)} + \mathbf{H}^T \mathbf{C}_h^{-1} \mathbf{H}]^{-1} \mathbf{G}^{(p)T} \mathbf{C}_d^{-1} \end{aligned} \tag{3.5.7}$$

We can then define a resolution matrix:

$$\mathbf{R}^{G(p)} = [\mathbf{G}^{(p)}]^{-g} \mathbf{G}^{(p)} \text{ that satisfies } (\Delta\delta\mathbf{m}^{(p)})^{rec} = \mathbf{R}^{G(p)} (\Delta\delta\mathbf{m}^{(p)})^{ass} \tag{3.5.8}$$

This resolution matrix is exactly analogous to the linear case (3.4.8); one merely substitutes the linearized data kernel $\mathbf{G}^{(p)}$ for the usual and linear data kernel, \mathbf{G} . Superficially, the quantity being resolved looks complicated—the deviation between two perturbations. However, closer examination reveals:

$$\Delta\delta\mathbf{m}^{(p)} = \delta\mathbf{m}^{(p)} - \delta\mathbf{m}^{H(p)} = (\mathbf{m} - \mathbf{m}^{(p)}) - (\mathbf{m}^H - \mathbf{m}^{(p)}) = \mathbf{m} - \mathbf{m}^H \tag{3.5.9}$$

This is the same quantity as in the linear case (3.4.6), that is, the deviation of the model from the prior model.

3.6 Symmetric Resolution in the Special Case of Convolutions

Let us consider the special case where \mathbf{m} represents the discrete version of a continuous function $m(x)$ and where \mathbf{G} and \mathbf{H} represent convolutions (Bracewell 1986; Claerbout 1976; see also Menke and Menke 2011). That is, $\mathbf{G}\mathbf{m}$ is the discrete version of $G(t) * m(t)$, where $*$ is the convolution operator. Furthermore, let us assume that the data and prior information are uncorrelated and with uniform variances, $\mathbf{C}_d = \sigma_d^2 \mathbf{I}$ and $\mathbf{C}_h = \sigma_h^2 \mathbf{I}$. Convolutions commute; that is, $G(t) * H(t) = H(t) * G(t)$. Consequently, the corresponding matrices will commute as well (except possibly for “edge effects”); that is, $\mathbf{GH} = \mathbf{HG}$. Furthermore, the transpose of a convolution matrix is itself a convolution—namely the original convolution backward in time; that is, $\mathbf{H}^T \rightarrow \mathbf{H}(-t)$. These properties imply that the resolution matrix:

$$\mathbf{R}^G = \mathbf{G}^{-g} \mathbf{G} = \mathbf{A}^{-1} \mathbf{G}^T \sigma_d^{-2} \mathbf{G} \tag{3.6.1}$$

is symmetric, since:

$$\begin{aligned}
\mathbf{R}^G &\stackrel{?}{=} (\mathbf{R}^G)^T \equiv \mathbf{A}^{-1} \mathbf{G}^T \sigma_d^{-2} \mathbf{G} \stackrel{?}{=} \mathbf{G}^T \sigma_d^{-2} \mathbf{G} \mathbf{A}^{-1} \equiv \mathbf{G}^T \mathbf{G} \mathbf{A} \stackrel{?}{=} \mathbf{A} \mathbf{G}^T \mathbf{G} \equiv \\
&\mathbf{G}^T \mathbf{G} [\sigma_d^{-2} \mathbf{G}^T \mathbf{G} + \sigma_h^{-2} \mathbf{H}^T \mathbf{H}] \stackrel{?}{=} [\sigma_d^{-2} \mathbf{G}^T \mathbf{G} + \sigma_h^{-2} \mathbf{H}^T \mathbf{H}] \mathbf{G}^T \mathbf{G} \equiv \\
&\sigma_h^{-2} \mathbf{G}^T \mathbf{G} \mathbf{H}^T \mathbf{H} \stackrel{?}{=} \sigma_h^{-2} \mathbf{H}^T \mathbf{H} \mathbf{G}^T \mathbf{G} \equiv \mathbf{H}^T \mathbf{H} \mathbf{G}^T \mathbf{G} = \mathbf{H}^T \mathbf{H} \mathbf{G}^T \mathbf{G}
\end{aligned} \tag{3.6.2}$$

4 Computational Efficiencies

4.1 Calculating the Generalized Least Squares Solution

In practice, the matrix inverse $\mathbf{A}^{-1} = [\mathbf{F}^T \mathbf{F}]^{-1}$ is not needed when computing an estimate of the model parameters from data; instead, once solves the linear system:

$$[\mathbf{F}^T \mathbf{F}] \mathbf{m}^{\text{est}} = \mathbf{F}^T \mathbf{f}^{\text{obs}} \tag{4.1.1}$$

Furthermore, when the biconjugate gradient solver (Press et al. 2007) is used, the matrix $\mathbf{A} = \mathbf{F}^T \mathbf{F} = \mathbf{G}^T \mathbf{C}_d^{-1} \mathbf{G} + \mathbf{H}^T \mathbf{C}_h^{-1} \mathbf{H}$ needs never to be explicitly calculated, since it is only used by the solver to multiply a known vector, say \mathbf{v} (Menke 2005). This product can be written as:

$$\mathbf{A} \mathbf{v} = \mathbf{A}^T \mathbf{v} = \mathbf{G}^T [\mathbf{C}_d^{-1} (\mathbf{G} \mathbf{v})] + \mathbf{H}^T [\mathbf{C}_h^{-1} (\mathbf{H} \mathbf{v})]; \tag{4.1.2}$$

that is, each intermediate result is a vector, not a matrix. This technique can lead to substantial efficiencies in speed and memory requirements, especially when the matrices are very large but sparse. The exemplary MATLAB code, below, calculates the solution \mathbf{m}^{est} , assuming that \mathbf{C}_d and \mathbf{C}_h are diagonal matrices with main diagonals vard and varh .

```
global G H vard varh
```

```
% define N, M, dobs, vard, hpri, varh, G, H, varh here
```

```
TOL=1e-5;
```

```
MAXI=4*N;
```

```
mest = bicg( @glsfcn, G'*(dobs./vard)+ H'*(hpri./varh), TOL, MAXI); \tag{4.1.3}
```

Here, `@glsfcn` is a handle to a function `glsfcn` that implements the multiplication shown in (4.1.2). Note that this function accesses the matrices \mathbf{G} and \mathbf{H} and the vectors vard and varh through their having been declared global variables in the main program.

```
function y = glsfcn(v,transp_flag)
```

```
global G H vard varh
```

```
y = G'*(G*v)./vard) + H'*(H*v)./varh); \tag{4.1.4}
```

```
return
```

4.2 Linearized GLS

Exemplary MATLAB code for the key part of the linearized inverse problem is shown below:

```

global G H vard varh

% define N, M, dobs, vard, hpri, varh, G, H, varh here
m_p = m_0;
for itt=[1:10]
    d_p = g(m_p);
    dd = dobs - d_p;
    dh = hpri - H*m_p;
    G = dgdgm(m_p); % matrix of derivatives dg/dm
    TOL=1e-5;
    MAXI=4*N;
    dm = bicg(@glsfcn,G'*(dd./vard)+H'*(dh./varh),TOL,MAXI);
    m_p = m_p+dm;
    if( ((dm'*dm)/(m_p'*m_p)) <= TOL )
        break;
    end
end
mest = m_p;

```

(4.2.1)

Here, $g(\mathbf{m}_p)$ and $dgdgm(\mathbf{m}_p)$ are two user-supplied functions that return $\mathbf{g}(\mathbf{m}^{(p)})$ and $\mathbf{G}(\mathbf{m}^{(p)})$, respectively. The maximum number of iterations is set here to 10, and the loop terminates early if the fractional change in the solution, from one iteration to the next, drops below 1×10^{-5} . These limits may need to be modified (by trial and error) to reflect the convergence properties of the actual problem being solved.

4.3 Calculating the k -th Row (or Column) of \mathbf{A}^{-1}

Note that the equation:

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{I} \text{ or } \sum_j A_{ij}A_{jk}^{-1} = \delta_{ik} \quad (4.3.1)$$

can be read as a sequence of vector equations:

$$\mathbf{A}\mathbf{v}^{(k)} = \mathbf{s}^{(k)} \text{ with } \left[\mathbf{v}^{(kj)} \right]_i = \left[\mathbf{A}^{-1} \right]_{ik} \text{ and } \left[\mathbf{s}^{(k)} \right]_i = \delta_{ik} \quad (4.3.2)$$

That is, $\mathbf{v}^{(k)}$ is the k -th column of \mathbf{A}^{-1} and $\mathbf{s}^{(k)}$ is the corresponding column of the identity matrix. Hence, we can solve for the k -th column of \mathbf{A}^{-1} by solving the system $\mathbf{A}\mathbf{v}^{(k)} = \mathbf{s}^{(k)}$. As in the previous section, the biconjugate gradient solver can be used to solve this system very efficiently. Finally, note that since \mathbf{A}^{-1} is symmetric, its k -th row is $\mathbf{v}^{(k)\text{T}}$. The exemplary MATLAB code, below, calculates the \mathbf{v}_k , assuming (as before) that \mathbf{C}_d and \mathbf{C}_h are diagonal matrices with main diagonals `vard` and `varh`.

```
global G H vard varh
% define N, M, dobs, vard, hpri, varh, G, H, varh, k here
s = zeros(M,1);
s(k)=1;
TOL=1e-5;
MAXI=4*N;
vk = bicg( @glsfcn, s, TOL, MAXI);
```

(4.3.3)

4.4 Calculating the k -th Row or Column of \mathbf{C}_m

In some instances, it is sufficient to compute a few representative elements of \mathbf{C}_m , as contrasted to the complete matrix. The results of the last section can be used directly, since $\mathbf{C}_m = \mathbf{A}^{-1}$. Exemplary MATLAB code for calculating the 95 % confidence intervals of model parameter m_j of the data is shown below:

```
% Note: k must match value used in calculation of vk
sigmamest = sqrt(vk(k));
% 95% confidence interval
mlow = mest(k)-2*sigmamest;
mhigh = mest(k)+2*sigmamest;
```

(4.4.1)

4.5 Calculating the k -th Row of the Generalized Inverse

Notice that:

$$\mathbf{G}^{-g} = \mathbf{A}^{-1}\mathbf{G}^T\mathbf{C}_d^{-1} \text{ or } [\mathbf{G}^{-g}]_{kj} = \sum_i A_{ki}^{-1} [\mathbf{G}^T\mathbf{C}_d^{-1}]_{ij} \quad (4.5.1)$$

Hence, the k -th row of the generalized inverse \mathbf{G}^{-g} and equals the k -th row of \mathbf{A}^{-1} dotted into $[\mathbf{G}^T\mathbf{C}_d^{-1}]$. We can construct the k -th row of the generalized inverse after using

the method of the previous section to calculate the k -th row of \mathbf{A}^{-1} . In many cases, \mathbf{C}_d is a diagonal matrix, which substantially simplifies the process of computing $[\mathbf{G}^T \mathbf{C}_d^{-1}]$:

$$[\mathbf{G}^T \mathbf{C}_d^{-1}]_{jk} = \sum_i G_{ij} \sigma_{d_i}^{-2} \delta_{ik} = G_{kj} \sigma_{d_k}^{-2} \tag{4.5.2}$$

Exemplary MATLAB code is shown below:

```
% Note: k matches the value used in the calculation of vk
gi = ((vk')*(G'))./(vard'); \tag{4.5.3}
```

4.6 Calculating the k -th Row of the Resolution Matrix

In some instances, it is sufficient to compute a few representative rows of \mathbf{R}^G , as contrasted to the complete matrix. The resolution matrix is formed from the generalized inverse and data kernel through:

$$\mathbf{R}^G = \mathbf{G}^{-g} \mathbf{G} \text{ or } R_{kj}^G = \sum_i G_{ki}^{-g} G_{ij} \tag{4.6.1}$$

Thus, the k -th row of the resolution matrix is the k -th row of the generalized inverse dotted into the data kernel. We can construct the k -th row of the resolution matrix after using the method of the previous section to calculate the k -th row of the generalized inverse.

Exemplary MATLAB code is shown below:

```
% Note: k matches the value used in the calculation of gi
r_row = gi*G; \tag{4.6.2}
```

4.7 Calculating the k -th Column of the Resolution Matrix.

Let us define the k -th column of the resolution matrix as the vector $\mathbf{r}^{(k)}$; that is,

$$r_i^{(k)} = R_{ik}^G \tag{4.7.1}$$

Then, notice that the definition $\mathbf{R}^G = \mathbf{G}^{-g} \mathbf{G}$ can be written as

$$\begin{aligned} \mathbf{R}^G = \mathbf{G}^{-g} \mathbf{G} \text{ or } r_i^{(k)} = R_{ik}^G &= \sum_p G_{ip}^{-g} \sum_q G_{pq} \delta_{qk} = \sum_p G_{pq}^{-g} d_q^{(k)} \\ \text{where } d_q^{(k)} &= \sum_q G_{pq} \delta_{qk} = \sum_q G_{pq} s_p^{(k)} \end{aligned} \tag{4.7.2}$$

As before, $\mathbf{s}^{(k)}$ is the k -th column of the identity matrix. The quantity $\mathbf{d}^{(k)} = \mathbf{G} \mathbf{s}^{(k)}$ is the data predicted by a set of model parameters $\mathbf{m}^{\text{true}(k)} = \mathbf{s}^{(k)}$ that are all zero, except for the k -th, which is unity. Thus, the two-step process:

$$\mathbf{d}^{(k)} = \mathbf{G}\mathbf{s}^{(k)} \text{ and } \mathbf{r}^{(k)} = \mathbf{G}^{-g}\mathbf{d}^{(k)} \quad (4.7.3)$$

forms the k -th column of the resolution matrix. In practice, the linear system $\mathbf{A}\mathbf{r}^{(k)} = \mathbf{G}^T\mathbf{C}_d^{-1}\mathbf{d}^{(k)}$ is solved (e.g., with a biconjugate gradients solver) instead of the equation containing the generalized inverse. Exemplary MATLAB code is shown below:

```
% unit spike at k
s = zeros(M,1);
s(k)=1;
% corresponding spiked data
ds = G*s;
% solve for column r_col of resolution matrix
TOL=1e-5;
MAXI=4*N;
r_col = bicg( @glsfcn, G'*(ds./vard), TOL, MAXI);
```

(4.7.4)

5 Minimum Curvature Splines as an Illustrative Example

5.1 Statement of the Problem

Let model parameters \mathbf{m} and data \mathbf{d} represent discrete version of continuous functions $m(x)$ and $d(x)$, say with spacing Δx . We would like to find model parameters that are approximately equal to the data [that is, $m(x) \approx d(x)$], but which are smoother (that is, have smaller second derivative). This is, of course, a data interpolation problem, and in that context, its solution would be called a minimum curvature spline (Schoenberg 1946; Briggs 1974; Smith and Wessel 1990).

Since the model parameters are direct estimates of the data, we set $\mathbf{G} = \mathbf{I}$. Smoothing is achieved by setting $\mathbf{h}_0 = 0$ and \mathbf{H}_0 to the second-derivative operator, with rows like:

$$(\Delta x)^{-2}[0 \quad \dots \quad 0 \quad 1 \quad -2 \quad 1 \quad 0 \quad \dots \quad 0] \quad (5.1.1)$$

Thus, $\mathbf{H}_0\mathbf{m} \approx 0$ for a smooth function. The data \mathbf{d} are uncorrelated with unit variance, $\sigma_d^2 = 1$, and the prior information $\mathbf{h}_0 = 0$ is uncorrelated with uniform variance, σ_h^2 . The degree of smoothing increases with $\gamma^2 = \sigma_d^2/\sigma_h^2$, that is, as the variance of the prior information of smoothness is decreased.

5.2 GLS Solution

Since $\mathbf{h}_0 = 0$, the solution has the form $\mathbf{m}^{\text{est}} = \mathbf{G}^{-g}\mathbf{d}^{\text{obs}}$. The generalized inverse \mathbf{G}^{-g} is:

$$\mathbf{G}^{-g} = \mathbf{A}^{-1}\sigma_d^{-2} \text{ with } \mathbf{A} = [\sigma_d^{-2}\mathbf{I} + \sigma_h^{-2}\mathbf{H}_0^T\mathbf{H}_0 + \varepsilon^2\mathbf{I}]$$

or

$$\mathbf{G}^{-g} = \mathbf{B}^{-1} \text{ with } \mathbf{B} = \sigma_d^2 \mathbf{A} = [(1 + \varepsilon^2 \sigma_d^2)I + \gamma^2 \mathbf{H}_0^T \mathbf{H}_0] \tag{5.2.1}$$

The model \mathbf{m}^H implied by the prior information alone can be deduced directly, because it corresponds to the smallest model with zero second derivative. The condition of a zero second derivative implies that \mathbf{m}^H is linear, and the condition that \mathbf{m}^H is small selects the linear function $\mathbf{m}^H = 0$. Hence, we do not need to distinguish between \mathbf{m} and $\Delta\mathbf{m}$ and will continue to use \mathbf{m} . Having resolved this issue, we can set $\varepsilon^2 = 0$ without affecting subsequent results. The variance of the model parameters is:

$$\mathbf{C}_m = \mathbf{A}^{-1} = \sigma_d^2 \mathbf{B}^{-1} \tag{5.2.2}$$

and the resolution matrix is:

$$\mathbf{R}^G = \mathbf{G}^{-g} \mathbf{G} = \mathbf{B}^{-1} \tag{5.2.3}$$

Note that $\mathbf{R}^G = \mathbf{I}$ (perfect resolution) when $\gamma^2 = 0$. Furthermore, \mathbf{R}^G is symmetric, since \mathbf{B} is symmetric.

5.3 Numerical Example

We used MATLAB to solve the exemplary smoothing problem, using synthetic data consisting of a sinusoidal function $m(x)$ with additive normally distributed noise (Figs. 1, 2, 3). The example shown here is for $N = M = 101$, which executes in 0.45 s on a notebook computer with a 2.5-GHz Intel Core i5-3210 M CPU. Test runs for $N = M = 5,001$ (not shown) were also successful and executed in a few minutes. The MATLAB code for this example is provided as supplementary material.

5.4 Analytic Analysis for Weak Smoothing

In most practical cases, a purely numerical solution to a GLS problem would be sufficient. However, many inverse problems (such as this one) have a sufficiently simple structure that analytic analysis adds important insights. Furthermore, it provides a reference against which the numerical results can be compared.

We first consider the case where the smoothness constraint is weak ($\gamma^2 \ll 1$). We can expand the matrix \mathbf{B}^{-1} in a Taylor series (e.g., Menke and Abbott 1989, Exercise 2.1), keeping only the first two terms:

$$\mathbf{B}^{-1} = \mathbf{G}^{-g} = \mathbf{R}^G = \sigma_d^{-2} \mathbf{C}_m = [\mathbf{I} + \gamma^2 \mathbf{H}_0^T \mathbf{H}_0]^{-1} \approx \mathbf{I} - \gamma^2 \mathbf{H}_0^T \mathbf{H}_0 \tag{5.4.1}$$

A row of \mathbf{B}^{-1} looks like:

$$\left[0 \quad \dots \quad 0 \quad -(\gamma \Delta x)^{-2} \quad 4(\gamma \Delta x)^{-2} \quad \left(1 - 6(\gamma \Delta x)^{-2}\right) \quad 4(\gamma \Delta x)^{-2} \quad -(\gamma \Delta x)^{-2} \quad 0 \quad \dots \quad 0 \right] \tag{5.4.2}$$

In the absence of the smoothness information ($\gamma^2 = 0$), $\mathbf{B}^{-1} = \mathbf{I}$, implying that estimated model parameters are uncorrelated and with uniform variance σ_d^2 and that the resolution is perfect. As the strength of smoothness information is increased, the magnitude

Fig. 1 (Top) Numerical test of the exemplary smoothing problem discussed in the text, for $\gamma = 0.05$. The observed data (black circles) and estimated model parameters (red curve), with selected 95 % confidence intervals (blue bars), are shown. (Bottom) Similar example, but for $\gamma = 0.005$. Note that the smaller γ implies that less weight is given to the prior information of smoothness, leading to a rougher curve. The confidence intervals are also wider, a manifestation of the trade-off of resolution and variance

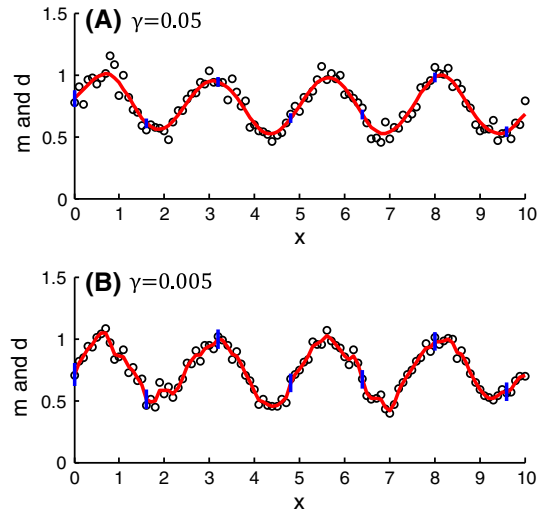


Fig. 2 Resolution of the exemplary smoothing problem discussed in the text. Selected rows of the resolution matrix \mathbf{R}^G for the case $\gamma = 0.05$. Rows (black) are calculated individually, according to the method described in the text. In this example, the resolution matrix is symmetric, so transposed columns, computed individually using the method described in the text, are also plotted (green). Results from the continuum limit, where the inverse problem is converted into a differential equation, are also shown (red). As expected, all curves agree

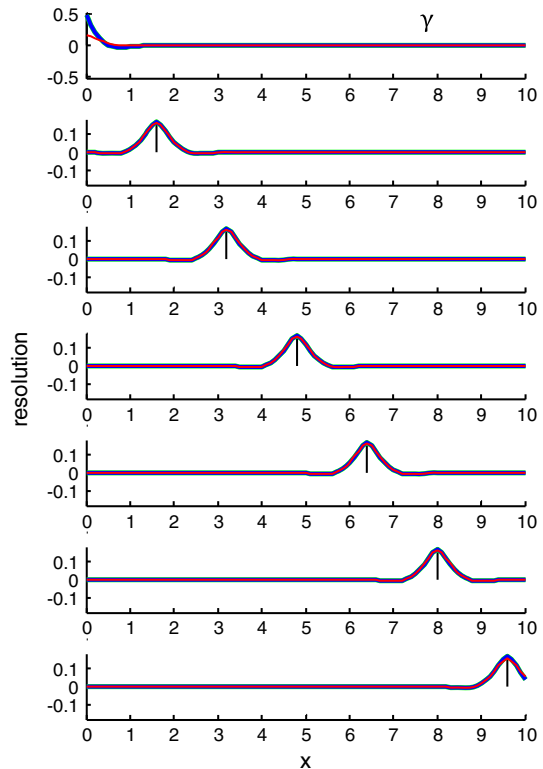
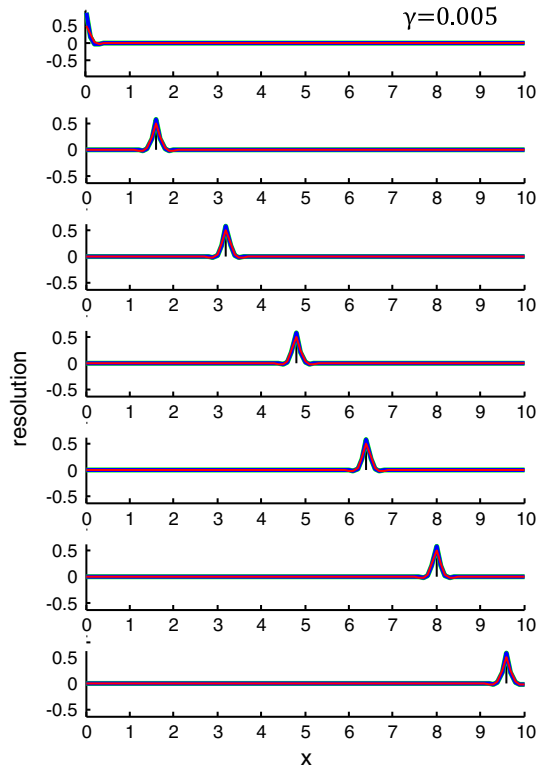


Fig. 3 Same as Fig. 2, but for $\gamma = 0.005$. Results for the smaller γ have the smaller spread



of the central value decreases and the nearest neighbor values become positive. For example, when $(\gamma\Delta x)^{-2} = 0.01$, a row of \mathbf{B}^{-1} looks like:

$$[0 \quad \dots \quad 0 \quad -0.01 \quad 0.04 \quad 0.94 \quad 0.04 \quad -0.01 \quad 0 \quad \dots \quad 0] \tag{5.4.3}$$

The smoothing has caused the variance of each model parameter (the central value) to decrease from σ_d^2 to $0.94\sigma_d^2$. However, the smoothing has also created covariance between model parameters, which decreases with separation. The smoothing has also caused the spread of the resolution to increase. However, although the row-sum is unity, the outermost nonzero values are negative, indicating that the smoothing cannot be interpreted as a weighted average in the normal sense.

5.5 Resolution in the Continuum Limit

Suppose, as before, the vector $\mathbf{s}^{(k)}$ represents the k -th column of the identity matrix (say a column k corresponding to position x_k). The equation $\mathbf{g}^{(k)} = \mathbf{G}^{-\mathbf{g}}\mathbf{s}^{(k)}$ can be understood both as the k -th column of the generalized inverse and as the data predicted by a model that is a single spike at position x_k . We have:

$$\mathbf{g}^{(k)} = \mathbf{G}^{-\mathbf{g}}\mathbf{s}^{(k)} = \mathbf{B}^{-1}\mathbf{s}^{(k)} = [\mathbf{I} + \gamma^2\mathbf{H}_0^T\mathbf{H}]^{-1}\mathbf{s}^{(k)} \tag{5.5.1}$$

Moving the matrix inverse to the l.h.s. of the equation yields:

$$\gamma^2 \mathbf{H}_0^T \mathbf{H}_0 \mathbf{g}^{(k)} + \mathbf{g}^{(k)} = \mathbf{s}^{(k)} \quad (5.5.2)$$

The second-derivative matrix \mathbf{H}_0 is symmetric, so that $\mathbf{H}_0^T \mathbf{H}_0 = \mathbf{H}_0 \mathbf{H}_0$; that is, a second-derivative operator applied twice to yield the fourth-derivative operator. Except for the first and last row, where edge effects are important, the matrix equation is the discrete analog to the differential equation:

$$\gamma^2 \frac{d^4 g^{(k)}}{dx^4} + g^{(k)}(x) = \Delta x \delta(x - x_k) \quad (5.5.3)$$

The factor of Δx has been added so that the area under $\Delta x \delta(x - x_k)$ is the same as the area under $\mathbf{s}^{(k)}$. This well-known differential equation has solution (Hetenyi 1979; see also Menke and Abbott 1989; Smith and Wessel 1990):

$$g^{(k)}(x) = V \exp(-|x - x_k|/a) \{ \cos(|x - x_k|/a) + \sin(|x - x_k|/a) \} \quad (5.5.4)$$

with

$$V = \frac{\Delta x a^3}{8\gamma^2} \text{ and } a = (2\gamma)^{1/2} \quad (5.5.5)$$

This differential equation arises in a civil engineering context, where it is used to describe the deflection $g^{(k)}(x)$ of an elastic beam of flexural rigidity γ^2 floating on a fluid foundation, due to a point load at x_k (Hetenyi 1979). The beam will take on a shape that exactly mimics the load only in the case when it has no rigidity; that is, $\gamma^2 = 0$. For any finite rigidity, the beam will take on a shape that is a smoothed version of the load, where the amount of smoothing increases with γ^2 . In our example, the model is analogous to the deflection of the beam and the data to the load; that is, the data are smoothed to produce the model. The parameter $a = (2\gamma)^{1/2}$ gives the scale length over which the smoothing occurs. The function $g^{(k)}(x)$ is analogous to the k -th row of the generalized inverse, so its (k, j) element is just the function evaluated at the x -position corresponding to the j -th position, or:

$$G_{kj}^{-g} = g^{(k)}(x_j) = V \exp(-|x_j - x_k|/a) \{ \cos(|x_j - x_k|/a) + \sin(|x_j - x_k|/a) \}$$

Since, in this example, $\mathbf{C}_m = \sigma_d^2 \mathbf{G}^{-g}$ and $\mathbf{R}^G = \mathbf{G}^{-g}$, we have found expressions for the covariance and resolution, as well. The variance of an estimated model parameter is $\sigma_m^2 = \sigma_d^2 V = \sigma_d^2 \Delta x (8\gamma)^{-1/2}$. Note that the variance of a model parameter declines as the smoothing is increased, but the number of highly correlated neighboring model parameters increases, being proportional to $a/\Delta x = (2\gamma)^{1/2}/\Delta x$. Owing to the trigonometric functions, \mathbf{R}^G has negative (but small) side lobes. Note that the resolution matrix \mathbf{R}^G is symmetric—a result guaranteed by the fact that both \mathbf{G} and \mathbf{H}_0 correspond to convolutions. The spread of the resolution is proportional to $a = (2\gamma)^{1/2}$, a measure of width of the main diagonal of \mathbf{R}^G .

The size of covariance $\sigma_d^2 \Delta x (8\gamma)^{-1/2}$ and the spread of resolution $(2\gamma)^{1/2}$ define a trade-off curve in the tunable parameter γ . When γ is small, the spread is small (good), but the variance is large (bad). When γ is large, the spread is large (bad), but the variance is small (good). While size of covariance and spread of resolution cannot be controlled independently, the parameter γ can be chosen to fix the best combination that is optimum for a given purpose.

6 Spin-offs of Generalized Least Squares

The GLS method remains one of the most useful and versatile techniques of data analysis, yet at the same time one that spawned a host of alternatives methodologies, some of which currently have only niche applications, while others are being more widely applied. In aggregate, they represent an evolving mindset regarding how data analysis problems are approached. We review some of these developments briefly here.

6.1 Sparsity

Prior information of the *sparseness* of the solution, meaning that all but a few of m_i 's are zero, is finding broad application (Candes et al. 2008). However, implementing the idea of sparsity requires a modification to the GLS method. Consider the generalized error in (3.2.1) in the special case of $\mathbf{H} = \mathbf{I}$, $\mathbf{h}^{\text{pri}} = 0$ and $\mathbf{C}_h^{-1} = \lambda \mathbf{I}$:

$$\Phi_{\text{GLS}} = [\mathbf{d}^{\text{obs}} - \mathbf{G}\mathbf{m}]^T \mathbf{C}_d^{-1} [\mathbf{d}^{\text{obs}} - \mathbf{G}\mathbf{m}] + \lambda \sum_{i=1}^M m_i^2 \quad (6.1.1)$$

The minimization of this generalized error selects for smallness (meaning that most m_i 's have small amplitude) but not for sparseness. However, the modified error:

$$\Phi_0 = [\mathbf{d}^{\text{obs}} - \mathbf{G}\mathbf{m}]^T \mathbf{C}_d^{-1} [\mathbf{d}^{\text{obs}} - \mathbf{G}\mathbf{m}] + \lambda \sum_{i=1}^M P(m_i) \quad (6.1.2)$$

will select for sparseness when the function P is defined as returning zero when $m_i = 0$ and unity otherwise. In other words, the summation is just a count of the number of nonzero model parameters. Unfortunately, the only currently known technique for minimizing Φ_0 involves an exhaustive search through all 2^M possible combinations of nonzero m_i 's, which is practical only when M is very small (see Menke and Levin 2003, for an example, involving seismic anisotropy). A better-behaved approximation to (6.1.2) is (Figueiredo et al. 2007):

$$\Phi_1 = [\mathbf{d}^{\text{obs}} - \mathbf{G}\mathbf{m}]^T \mathbf{C}_d^{-1} [\mathbf{d}^{\text{obs}} - \mathbf{G}\mathbf{m}] + \lambda \sum_{i=1}^M |m_i| \quad (6.1.3)$$

The function $P(m_i)$ in (6.1.2) is constant for increasing m_i , whereas the function m_i^2 in (6.1.1) increases very rapidly with it. The function $|m_i|$ in (6.1.3) has an intermediate behavior; it grows only slowly with m_i . Thus, $|m_i|$ can be used as an approximation—or proxy—for $P(m_i)$. Its use allows the minimization problem (6.1.3) to be recast as a more tractable linear programming problem (Cuer and Bayer 1980; Boyd and Vandenberghe 2004), for which solution methods are widely available. We will discuss an important application of sparse solutions in the next section.

6.2 Over-Parameterization and Basis Pursuit

Consider a continuous model parameter $m(x)$ that has been parameterized as a time series \mathbf{m} ; that is, by its values at M equally spaced values of x . Other parameterizations are possible; $m(x)$ can also be represented by a Fourier series (say, with coefficients $\mathbf{m}^{(1)}$), by orthogonal polynomials (say, with coefficients $\mathbf{m}^{(2)}$), by overlapping step functions (say, with coefficients $\mathbf{m}^{(3)}$) and in many other ways. In each case, we can write down a linear

relationship between the coefficients of a particular parameterization (say the k -th parameterization) and the time series:

$$\mathbf{m} = \mathbf{D}^{(k)} \mathbf{m}^{(k)} \quad (6.2.1)$$

The matrix $\mathbf{D}^{(k)}$ can be thought of as a *dictionary* of the component shapes (or *basis functions*) in the k -th parameterization. Each of its columns gives the pattern of one shape as a function of x . The data Eq. (2.1.1) then become:

$$\mathbf{d} = \mathbf{G}\mathbf{m} \text{ becomes } \mathbf{d} = \left[\mathbf{GD}^{(k)} \right] \mathbf{m}^{(k)} = \mathbf{G}^{(k)} \mathbf{m}^{(k)} \quad (6.2.2)$$

Here, $\mathbf{G}^{(k)}$ is an abbreviation for $\mathbf{GD}^{(k)}$.

The choice of parameterization has long been an important subject in least squares theory. Traditionally, models have been either even-parameterized (able to uniquely represent any possible \mathbf{m}) or under-parameterized (able to represent only a subset of possible \mathbf{m} 's). Under-parameterizations are used both to reduce computational burden and to impose prior information (e.g., impose smoothness by omitting high-wavenumber coefficients from a Fourier parameterization). However, recent work has demonstrated that over-parameterizations have extremely important uses, as well.

An over-parameterization can be assembled by combining several even-parameterizations. Suppose, for example, that $k = 1$ represents a Fourier parameterization and that $k = 2$ represents a step function parameterization. Then, a new parameterization \mathbf{m}^* , with a $M^* = 2M$ coefficients, is:

$$\mathbf{m}^* = \begin{bmatrix} \mathbf{m}^{(1)} \\ \mathbf{m}^{(2)} \end{bmatrix} \text{ implying } \mathbf{D}^* = \begin{bmatrix} \mathbf{D}^{(1)} & \mathbf{D}^{(2)} \end{bmatrix} \text{ and } \mathbf{G}^* = \mathbf{GD}^* \quad (6.2.3)$$

Superficially, the redundancy of an over-parameterization would appear to be a disadvantage. It adds more model parameters without increasing the ability of the parameterization to represent the model. However, it can be quite useful when the choice of basis functions is guided by prior information about the character of \mathbf{m} . For instance, suppose one believes that \mathbf{m} should consist of the superposition of a single sinusoid and a single-step function. The \mathbf{m}^* parameterization requires just two nonzero coefficients, whereas either the $\mathbf{m}^{(1)}$ and $\mathbf{m}^{(2)}$, used individually, require M . The problem of finding a sparse solution to an over-parameterized model, which is associated with the phrase *basis pursuit* (Chen et al. 1998), connects the ideas of sparseness developed in Sect. 6.1 with those of pattern detection. Note that the parameter λ in:

$$\Phi_1 = [\mathbf{d}^{\text{obs}} - \mathbf{G}^* \mathbf{m}^*]^T \mathbf{C}_d^{-1} [\mathbf{d}^{\text{obs}} - \mathbf{G}^* \mathbf{m}^*] + \gamma \sum_{i=1}^{M^*} |m_i^*| \quad (6.1.3)$$

controls the relative weight given to prediction error and sparseness. Applications of the technique include detecting sharp boundaries in tomographic images (Gholami and Siahkoobi 2010) and estimating compact earthquake slip patterns (Evans and Meade 2012).

7 Conclusions

Generalized least squares has proven to be an extremely powerful tool for solving inverse problems, that is, for gaining knowledge about the world. The concepts of resolution and

variance, so useful for understanding the behavior of inverse problems in general, are applicable to GLS, but with some caveats. Resolution is computed via the usual formula; however, the quantity that is being resolved is not the model itself (as it is in simpler inverse problems), but its deviation from the prior model; that is, the model implied by the prior information. This is true irrespective of whether the problem is exactly linear or approximately linearized. The formula for covariance contains a term not present in the simple least squares case, which is proportional to the uncertainty of the prior information. Thus, the covariance of the model depends on both the covariance of the data and the covariance of the prior information. Although both formulas superficially require large matrices to be inverted, the calculations can be organized to allow individual rows and columns of both to be computed without the need for matrix inversion. In practice, a few representative rows are usually all that is needed, for it is impractical to completely analyze very large matrices, anyway.

Acknowledgments This research was supported by the US National Science Foundation under grants OCE-0426369 and EAR 11-47742.

References

- Abers G (1994) Three-dimensional inversion of regional P and S arrival times in the East Aleutians and sources of subduction zone gravity highs. *J Geophys Res* 99:4395–4412
- Backus GE, Gilbert JF (1968) The resolving power of gross earth data. *Geophys J R Astron Soc* 16:169–205
- Backus GE, Gilbert JF (1970) Uniqueness in the inversion of gross Earth data. *Philos Trans R Soc Lond Ser A* 266:123–192
- Boyd S, Vandenberghe L (2004) *Convex optimization*. Cambridge University Press, Cambridge
- Bracewell R (1986) *The Fourier transform and its applications*. McGraw-Hill, New York
- Briggs IC (1974) Machine contouring using minimum curvature. *Geophysics* 39:39–48
- Candes EJ, Wakin MB, Boyd SP (2008) Enhancing sparsity by reweighted L_1 minimization. *J Fourier Anal Appl* 14:877–905
- Chen S, Donoho D, Saunders M (1998) Atomic decomposition by basis pursuit. *SIAM J Sci Comput* 20:33–61
- Claerbout J (1976) *Fundamentals of geophysical data processing*. McGraw-Hill, New York, p 274
- Cuer M, Bayer R (1980) FORTRAN routines for linear inverse problems. *Geophysics* 45:1706–1719
- Evans EL, Meade BJ (2012) Geodetic imaging of coseismic slip and postseismic afterslip: sparsity promoting methods applied to the great Tohoku earthquake. *Geophys Res Lett* 39. doi:[10.1029/2012GL051990](https://doi.org/10.1029/2012GL051990)
- Figueiredo M, Bioucas-Dias JM, Nowak RD (2007) Majorization–minimization algorithms for wavelet-based image restoration. *IEEE Trans Image Process* 16:2980–2991
- Friedel S (2003) Resolution, stability and efficiency of resistivity tomography estimated from a generalized inverse approach. *Geophys J Int* 153:305–316
- Gholami A, Siahkoochi H (2010) Regularization of linear and non-linear geophysical ill-posed problems with joint sparsity constraints. *Geophys J Int* 180:871–882
- Gunther T (2004). *Inversion methods and resolution analysis for the 2D/3D reconstruction of resistivity structures from DC measurements* PhD thesis, Technische Universität Bergakademie Freiberg
- Hetenyi M (1979) *Beams on elastic foundation*. University of Michigan Press, Ann Arbor, p 245
- Householder AS (1958) Unitary triangularization of a nonsymmetric matrix. *J ACM* 5:339–342
- Kalscheuer T (2008) *Improvement and assessment of two-dimensional resistivity models derived from radiomagnetotelluric and direct-current resistivity data* PhD thesis, Uppsala University, 2008
- Kalscheuer T, de los Angeles García Juanatey M, Meqbel N, Pedersen LB (2010) Non-linear model error and resolution properties from two-dimensional single and joint inversions of direct current resistivity and radiomagnetotelluric data. *Geophys J Int* 182:1174–1188
- Lawson C, Hanson R (1974) *Solving least squares problems*. Prentice-Hall, Englewood Cliffs
- Legendre AM (1805) *Nouvelles méthodes pour la détermination des orbites des comètes*. [New Methods for the Determination of the Orbits of Comets]. (in French), Paris: F. Didot
- Lerner-Lam A, Jordan TH (1987) How thick are the continents? *J Geophys Res* 92:14007–14026

- Menke W (1984) Geophysical data analysis: discrete inverse theory, 1st edn. Academic Press Inc, New York
- Menke W (2005) Case studies of seismic tomography and earthquake location in a regional context. In: Levander A, Nolet G (eds) Seismic earth: array analysis of broadband seismograms. Geophysical Monograph Series 157. American Geophysical Union, pp 7–36
- Menke W (2012) Geophysical data analysis: discrete inverse theory, MATLAB edn. Elsevier Inc, New York
- Menke W, Abbott D (1989) Geophysical theory. Columbia University Press, Columbia
- Menke W, Levin VV (2003) The cross-convolution method for interpreting SKS splitting observations, with application to one and two layer anisotropic earth models. *Geophys J Int* 154:379–392
- Menke W, Menke J (2011) Environmental data analysis with MATLAB. Elsevier Inc, New York
- Press W, Teukolsky S, Vetterling WT, Flannery BP (2007) Numerical recipes, 3rd edn. Cambridge University Press, Cambridge
- Rektorys K (1969) Survey of applicable mathematics. M.I.T Press, Cambridge
- Schoenberg IJ (1946) Contributions to the problem of approximation of equidistant data by analytic functions. *Quart Appl Math* 4:45–99, 112–141
- Smith SW (1997) The scientist and engineer's guide to digital signal processing. (eBook), DSPguide.com
- Smith W, Wessel P (1990) Gridding with continuous curvature splines in tension. *Geophysics* 55:293–305
- Tarantola A, Valette B (1982a) Generalized non-linear inverse problems solved using the least squares criterion. *Rev Geophys Space Phys* 20:219–232
- Tarantola A, Valette B (1982b) Inverse problems = quest for information. *J Geophys* 50:159–170
- Wiggins RA (1972) The general linear inverse problem: Implication of surface waves and free oscillations for Earth structure. *Rev Geophys Space Phys* 10:251–285
- Yao Z, Roberts R, Tryggvason A (1999) Calculating resolution and covariance matrices for seismic tomography with the LSQR method. *Geophys J Int* 138:886–894