Applied Mathematics, 2021, 12, 157-170
https://www.scirp.org/journal/am
ISSN Online: 2152-7393
ISSN Print: 2152-7385

Tuning of Prior Covariance in Generalized Least Squares

William Menke

Lamont-Doherty Earth Observatory of Columbia University, New York, USA
Email: menke@ldeo.columbia.edu

How to cite this paper: Menke, W. (2021) Tuning of Prior Covariance in Generalized Least Squares. *Applied Mathematics*, **12**, 157-170.
https://doi.org/10.4236/am.2021.123011

Received: February 5, 2021
Accepted: March 14, 2021
Published: March 17, 2021

Abstract

Generalized Least Squares (least squares with prior information) requires the correct assignment of two prior covariance matrices: one associated with the uncertainty of measurements; the other with the uncertainty of prior information. These assignments often are very subjective, especially when correlations among data or among prior information are believed to occur. However, in cases in which the general form of these matrices can be anticipated up to a set of poorly-known parameters, the data and prior information may be used to better-determine (or "tune") the parameters in a manner that is faithful to the underlying Bayesian foundation of GLS. We identify an objective function, the minimization of which leads to the best-estimate of the parameters and provide explicit and computationally-efficient formula for calculating the derivatives needed to implement the minimization with a gradient descent method. Furthermore, the problem is organized so that the minimization need be performed only over the space of covariance parameters, and not over the combined space of model and covariance parameters. We show that the use of trade-off curves to select the relative weight given to observations and prior information is not a form of tuning, because it does not, in general maximize the posterior probability of the model parameters, and can lead to a different weighting than the procedure described here. We also provide several examples that demonstrate the viability, and discuss both the advantages and limitations of the method.

Keywords

Bayesian Inference, Covariance, Error, Generalized Least Squares, Gradient Descent, Interpolation, Regularization, Trade-Off Curve, Variance

1. Introduction

Generalized Least Squares (GLS, also called least-squared with prior information) is a tool for statistical inference [1]-[6] that is widely used in geotomography [7]-[12] and geophysical inversion [13] [14], as well as other areas of the physical sciences and engineering. One of the attractive features of GLS that makes it especially useful in the imaging of multidimensional fields (for example, density, velocity, viscosity) is its ability to implement, in a natural and versatile way, prior information of the behavior of the field. Widely-used types of prior information include the field being smooth, as quantified by its low-order derivatives [15], having a specified power spectral density or autocovariance [7] [15], and satisfying a specified partial differential equation (such as the geostrophic flow equation [16] or the diffusion equation [4]). The word "regularization" sometimes is used to describe the effect of prior information on the solution process [17].

We review the Generalized Least Squares (GLS) method here, following the notation in [6], in order to provide context and to establish nomenclature. In GLS, observations (or data) and prior information (or inferences) are combined to arrive at a best-estimate of initially-unknown model parameters (which might, for example, represent a field sampled on a regular grid). The data are assumed to satisfy the linear equation $\boldsymbol{Gm}=\boldsymbol{d}$, where $\boldsymbol{d}\in\mathbb{R}^N$ is a vector of data, $\boldsymbol{m}\in\mathbb{R}^M$ is a vector of model parameters, and $\boldsymbol{G}$ is a known "kernel" matrix associated with the data. Prior information is assumed to satisfy a linear equation $\boldsymbol{Hm}=\boldsymbol{h}$, where $\boldsymbol{h}\in\mathbb{R}^K$ is a vector of prior values and $\boldsymbol{H}$ is a kernel matrix associated with the prior information. GLS problems are assumed to be over-determined, with $N+K>M$. For observed data $\boldsymbol{d}^{obs}$, known prior information $\boldsymbol{h}^{pri}$ and a specified model $\boldsymbol{m}$, the prediction error is $\boldsymbol{e}\equiv\boldsymbol{d}^{obs}-\boldsymbol{Gm}$ and prior information error is $\boldsymbol{\ell}\equiv\boldsymbol{h}^{pri}-\boldsymbol{Hm}$. These errors are assumed to be Normally-distributed with zero mean and prior covariance $\boldsymbol{C}_d$ and $\boldsymbol{C}_h$, respectively. Then, the normalized errors $\tilde{\boldsymbol{e}}\equiv\boldsymbol{C}_d^{-1/2}\boldsymbol{e}$ and $\tilde{\boldsymbol{\ell}}\equiv\boldsymbol{C}_h^{-1/2}\boldsymbol{\ell}$ are independent and identically-distributed Normal random variables with zero mean and unit variance. Bayes theorem can be used to show that the best estimate $\boldsymbol{m}^{est}$ of the solution is the one that minimizes the generalized error $\Phi\equiv E+L$, with $E\equiv\tilde{\boldsymbol{e}}^{\mathrm{T}}\tilde{\boldsymbol{e}}$ and $L\equiv\tilde{\boldsymbol{\ell}}^{\mathrm{T}}\tilde{\boldsymbol{\ell}}$ [1] [2] [5]. The solution can be expressed in a variety of equivalent forms, among which is the widely-used version [6]:

$$\boldsymbol{m}^{est}=\boldsymbol{Z}^{-1}\left(\boldsymbol{G}^{\mathrm{T}}\boldsymbol{C}_d^{-1}\boldsymbol{d}^{obs}+\boldsymbol{H}^{\mathrm{T}}\boldsymbol{C}_h^{-1}\boldsymbol{h}^{pri}\right)\ \text{with}\ \boldsymbol{Z}\equiv\boldsymbol{G}^{\mathrm{T}}\boldsymbol{C}_d^{-1}\boldsymbol{G}+\boldsymbol{H}^{\mathrm{T}}\boldsymbol{C}_h^{-1}\boldsymbol{H} \tag{1}$$

The assumption of linear kernels $\boldsymbol{G}$ and $\boldsymbol{H}$ is a very restrictive one. In the well-studied nonlinear generalization [1] [6], the products $\boldsymbol{Gm}$ and $\boldsymbol{Hm}$ are replaced with vector functions $\boldsymbol{g}(\boldsymbol{m})$ and $\boldsymbol{h}(\boldsymbol{m})$. Then, a common solution method is to linearize the data and prior information equations around a trial solution $\boldsymbol{m}^{(0)}$:

$$\begin{gathered}\boldsymbol{G}^{(0)}\Delta\boldsymbol{m}=\Delta\boldsymbol{d}\ \text{ with }\ G_{ij}^{(0)}\equiv\left.\frac{\partial g_i}{\partial m_j}\right|_{\boldsymbol{m}^{(0)}}\ \text{ and }\ \Delta\boldsymbol{d}\equiv\boldsymbol{d}^{obs}-\boldsymbol{g}(\boldsymbol{m})\\ \boldsymbol{H}^{(0)}\Delta\boldsymbol{m}=\Delta\boldsymbol{h}\ \text{ with }\ H_{ij}^{(0)}\equiv\left.\frac{\partial h_i}{\partial m_j}\right|_{\boldsymbol{m}^{(0)}}\ \text{ and }\ \Delta\boldsymbol{h}\equiv\boldsymbol{h}^{pri}-\boldsymbol{h}(\boldsymbol{m})\end{gathered} \tag{2}$$

and $\Delta\boldsymbol{m}=\boldsymbol{m}-\boldsymbol{m}^{(0)}$. The solution is then found by iterative application of (1) applied to (2); that is, by the Gauss-Newton's method [3]. Alternatively, a gradient-descent method [18] can be used that employs:

$$\nabla_{\boldsymbol{m}}\Phi\big|_{\boldsymbol{m}^{(0)}}=-2\boldsymbol{G}^{(0)\mathrm{T}}\boldsymbol{C}_d^{-1}\left(\boldsymbol{d}^{obs}-\boldsymbol{G}\boldsymbol{m}^{(0)}\right)-2\boldsymbol{H}^{(0)\mathrm{T}}\boldsymbol{C}_h^{-1}\left(\boldsymbol{h}^{pri}-\boldsymbol{H}\boldsymbol{m}^{(0)}\right) \tag{3}$$

The latter approach is preferred for very large M, since the convergence rate of gradient descent is independent of its dimension [18], whereas the effort required to solve the $M\times M$ system (1) by a direct method scales as M^3 [19].

We now discuss issues related to the covariance matrices that appear in GLS. The data covariance $\boldsymbol{C}_d$ quantifies the uncertainty of the observations and the information covariance $\boldsymbol{C}_h$ quantifies the uncertainty of the prior information. Prior knowledge of the inherent accuracy of the measurement technique is needed to assign $\boldsymbol{C}_d$, and prior knowledge of the physically-plausible solutions, perhaps stemming from and understanding of the underlying physics, is needed to assign $\boldsymbol{C}_h$. These assignments are often very subjective, especially when correlations are believed to occur (that is, $\boldsymbol{C}_d$ and $\boldsymbol{C}_h$ have non-zero off-diagonal elements). For example, one geotomographic study [7] reconstructs a two-dimensional field using a $\boldsymbol{C}_h$ that represents autocovariance of the field and that is dependent upon a scale length q. The value of q is chosen on the basis of broad physical arguments that, while plausible, leaves considerable room for subjectivity.

The matrices $\boldsymbol{C}_d$ and $\boldsymbol{C}_h$ together contain $\frac{1}{2}N(N+1)+\frac{1}{2}K(K+1)$ elements, many more than the $(N+K)$ constraints imposed by the data $\boldsymbol{d}$ and prior information $\boldsymbol{h}$.Consequently, insufficient information is available to uniquely solve for all the elements of $\boldsymbol{C}_d$ and $\boldsymbol{C}_h$. However, it sometimes may be possible to parameterize $\boldsymbol{C}_d(\boldsymbol{q})$ and/or $\boldsymbol{C}_h(\boldsymbol{q})$ in terms of $\boldsymbol{q}\in\mathbb{R}^J$, and ask whether an initial estimate of $\boldsymbol{q}$ can be improved. As long as $(M+J)<(N+K)$, adequate information may be available to determine a best estimate $\boldsymbol{q}^{est}$. We refer to the process of determining $\boldsymbol{q}^{est}$ as "tuning", since in typical practice it requires that the covariances be close to their true values.

As an example of a parametrized covariance, we consider the case where the model parameters represent a sampled version of a continuous function $m(x)$, where $x\in\mathbb{R}$ is an independent variable; that is, $m_n=m(x_n)$, with $x_n\equiv n\Delta x$ and Δx the sampling interval. The prior information that $m(x)$ is approximately oscillatory with wavenumber q can be modeled by:

$$\boldsymbol{H}=\boldsymbol{I} \text{ and } \boldsymbol{h}^{pri}=0 \text{ and } \left[\boldsymbol{C}_h\right]_{nm}=\sigma_h^2\cos\left(q\left|x_n-x_m\right|\right) \tag{4}$$

In this case, $\boldsymbol{C}_h$ approximates the autocovariance of $m(x)$, which is assumed to be stationary. The goal of tuning is to provides a best-estimate q^{est}, as well of best estimated $\boldsymbol{m}^{est}$ of the model parameters. This problem is further developed in Example 4, below.

Although the GLS formulation is widely used in geotomography and geophysical imaging, the tuning of variance is typically implemented in a very limited fashion, through the use of trade-off curves [7]-[12]. In this procedure, a scalar parameter q controls the relative size of $\mathbf{C}_d$ and $\mathbf{C}_h$, that is, $\mathbf{C}_h(q) = q\mathbf{C}_h^{(0)}$, where $\mathbf{C}_h^{(0)}$ is specified [20]. The GLS problem is then solved for a suite of qs, the functions $E(q)$ and $L(q)$ are tabulated and the resulting trade-off curve $E(L)$ is used to identify a solution $\mathbf{m}(q_0)$ that has acceptably low E and L (for example, Figure 1 of [20]). As we will show below, this ad hoc procedure is not a consistent extension of GLS, because it results in a different q than the one implied by Bayes' principle. A more consistent approach is to apply Bayes theorem directly to estimate both the model parameters $\mathbf{m}$ and the covariance parameters $\mathbf{q}$. Such an approach has been implemented in the context of ordinary least squares [21] and the Markov chain Monte Carlo (MCMC) inversion method [22] (which is a computationally-intensive alternative to GLS). An important and novel result of this paper is a computationally-efficient procedure for tuning GLS in a Bayes-consistent manner.

2. Bayesian Extenion of GLS

The general process of using Bayes' theorem to construct a posterior probability density function (p.d.f.) that depends on unknown parameters and of estimating those parameters though the maximization of probability is very well understood [23]. In the current case, the p.d.f. has M model parameters and J covariance parameters, so the maximization process (implemented, say, with a gradient ascent method) must search an $(M+J)$-dimensional space. Our main purpose here is to show that the process can be organized in a way that makes use of the GLS solution (1) and thus reduce the dimensionality of the searched space to J.

The GLS solution (1) yields the $\mathbf{m}$ that minimizes the generalized error $\Phi(\mathbf{m})$, or equivalently, the $\mathbf{m}$ that maximizes the Normal posterior probability density function (p.d.f.) $p(\mathbf{m} \mid \mathbf{d}^{obs}, \mathbf{h}^{pri})$:

$$\begin{aligned} &\mathbf{m}^{est} = \underset{\mathbf{m}}{\arg\max}\, p(\mathbf{m} \mid \mathbf{d}^{obs}, \mathbf{h}^{pri}) \\ &\text{with } p(\mathbf{m} \mid \mathbf{d}^{obs}, \mathbf{h}^{pri}) \propto p(\mathbf{d}^{obs} \mid \mathbf{m})\, p(\mathbf{h}^{pri} \mid \mathbf{m}) \end{aligned} \tag{5}$$

Here, Bayes theorem [23] is used to related the Normal posterior p.d.f. $p(\mathbf{m} \mid \mathbf{d}^{obs}, \mathbf{h}^{pri})$ to the Normal likelihood $p(\mathbf{d}^{obs} \mid \mathbf{m})$ and the Normal prior $p(\mathbf{h}^{pri} \mid \mathbf{m})$. When poorly known parameters $\mathbf{q}$ are added to the problem, they must be treated as additional random variables [22]. Writing $\mathbf{q} \equiv \left[\mathbf{q}^{(d)}, \mathbf{q}^{(h)}\right]^{\mathrm{T}}$, with $\mathbf{q}^{(d)}$ appearing in the likelihood and $\mathbf{q}^{(h)}$ appear in the prior, we have:

$$\begin{aligned} &\mathbf{m}^{est}, \mathbf{q}^{est} = \underset{\mathbf{m}, \mathbf{q}^{(d)}, \mathbf{q}^{(h)}}{\arg\max}\, p(\mathbf{m}, \mathbf{q} \mid \mathbf{d}^{obs}, \mathbf{h}^{pri}) \\ &\text{with } p(\mathbf{m}, \mathbf{q} \mid \mathbf{d}^{obs}, \mathbf{h}^{pri}) \propto p(\mathbf{d}^{obs} \mid \mathbf{m}, \mathbf{q}^{(d)})\, p(\mathbf{h}^{pri} \mid \mathbf{m}, \mathbf{q}^{(h)})\, p(\mathbf{q}^{(d)})\, p(\mathbf{q}^{(h)}) \end{aligned} \tag{6}$$

Here, we have assumed that $\mathbf{q}$ and $\mathbf{m}$ are not correlated with one another. The maximization with respect to the two variables can be performed as a sequence of two single-variable maximizations:

$$\boldsymbol{m}(\boldsymbol{q}_0) : \underset{\boldsymbol{m}}{\arg\max}\; p\left(\boldsymbol{m}, \boldsymbol{q}_0 \mid \boldsymbol{d}^{obs}, \boldsymbol{h}^{pri}\right) (\text{at fixed } \boldsymbol{q}_0) \tag{7a}$$

$$\boldsymbol{q}^{est} : \underset{\boldsymbol{q}_0}{\arg\max}\; p\left(\boldsymbol{m}(\boldsymbol{q}_0), \boldsymbol{q}_0 \mid \boldsymbol{d}^{obs}, \boldsymbol{h}^{pri}\right) \tag{7b}$$

$$\boldsymbol{m}^{est} = \boldsymbol{m}\left(\boldsymbol{q}^{est}\right) \tag{7c}$$

In the special case of the uniform prior $p\left(\boldsymbol{q}_{(d)}\right) p\left(\boldsymbol{q}_{(h)}\right) \propto \text{constant}$, the maximization in (7a) is the GPR solution at fixed $\boldsymbol{q}_0$. For the Normal p.d.f.:

$$\begin{aligned} & p\left(\boldsymbol{m}(\boldsymbol{q}_0), \boldsymbol{q}_0 \mid \boldsymbol{d}^{obs}, \boldsymbol{h}^{pri}\right) \\ & = (2\pi)^{-\frac{1}{2}(N+K)} \left(\det \boldsymbol{C}_d\right)^{-\frac{1}{2}} \left(\det \boldsymbol{C}_k\right)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}E\right)\left(-\frac{1}{2}L\right) \end{aligned} \tag{8}$$

the maximization (7b) is equivalent to the minimization of an objective function $\Psi(\boldsymbol{q})$, defined as:

$$\Psi \equiv -2\left[\ln p + \frac{1}{2}(N+K)\ln(2\pi)\right] = \ln\left(\det \boldsymbol{C}_d\right) + \ln\left(\det \boldsymbol{C}_h\right) + E + L \tag{9}$$

The quantity $\ln\left(\det \boldsymbol{C}_d\right)$ is best computed by finding the Choleski decomposition $\boldsymbol{C}_d = \boldsymbol{D}\boldsymbol{D}^{\mathrm{T}}$, the algorithm [24] for which is implemented in many software environments, including MATLAB® and PYTHON/linalg. Then, $\ln\left(\det \boldsymbol{C}_d\right) = 2\sum_n \ln\left(D_{nn}\right)$ (and similarly for $\ln\left(\det \boldsymbol{C}_h\right)$).The nonlinear optimization problem of minimizing $\Psi(\boldsymbol{q})$ can be implemented using a gradient descent method, provided that the derivative $\partial\Psi/\partial q_m$ can be calculated [18]. In the next section, we derive analytic formula for this and related derivatives.

3. Solution Method and Formula for Derivatives

The process of simultaneously estimating the covariance parameters $\boldsymbol{q}^{est}$ and model parameters $\boldsymbol{m}^{est}$ consists of six steps. First, the analytic form of the covariance matrices $\boldsymbol{C}_d(\boldsymbol{q})$ and $\boldsymbol{C}_h(\boldsymbol{q})$ are specified, and their derivatives $\partial\boldsymbol{C}_d/\partial q_m$ and $\partial\boldsymbol{C}_h/\partial q_m$ are computed analytically. Second, an initial estimate $\boldsymbol{q}^{(0)}$ is identified. Third, the covariance matrices $\boldsymbol{C}_d\left(\boldsymbol{q}^{(0)}\right)$ and $\boldsymbol{C}_h\left(\boldsymbol{q}^{(0)}\right)$ are inserted into (1), yielding model parameters $\boldsymbol{m}\left(\boldsymbol{q}^{(0)}\right)$. Fourth, using formulas developed below, the value of the derivative $\partial\Psi/\partial q_m$ is calculated at $\boldsymbol{q}^{(0)}$. Fifth, a gradient descent method employing $\partial\Psi/\partial q_m$ is used to iteratively perturb $\boldsymbol{q}^{(0)}$ towards the minimum of Ψ at $\boldsymbol{q}^{est}$ (and in process, repeating steps three through five many times). Sixth, the estimated model parameters are computed as $\boldsymbol{m}^{est} = \boldsymbol{m}\left(\boldsymbol{q}^{est}\right)$. This process is depicted in **Figure 1**.

Our derivation of $\partial\Psi/\partial q_m$ uses three matrix derivatives, $\partial\boldsymbol{M}^{-1}/\partial q$, $\partial\boldsymbol{M}^{-1/2}/\partial q$ and $\partial\ln\left(\det\boldsymbol{M}\right)/\partial q$ that may be unfamiliar to some readers, so we derive them here for completeness. Let $\boldsymbol{M}(q)$ be asquare, invertible, differentiable matrix. Differentiating $\boldsymbol{M}^{-1}\boldsymbol{M} = \boldsymbol{I}$ yields $\left[\partial\boldsymbol{M}^{-1}/\partial q_m\right]\boldsymbol{M} + \boldsymbol{M}^{-1}\left[\partial\boldsymbol{M}/\partial q_m\right] = \boldsymbol{0}$, which can be rearranged into ([25], their (36)):

$$\frac{\partial\boldsymbol{M}^{-1}}{\partial q_m} = -\boldsymbol{M}^{-1}\left[\frac{\partial\boldsymbol{M}}{\partial q_m}\right]\boldsymbol{M}^{-1} \tag{10}$$

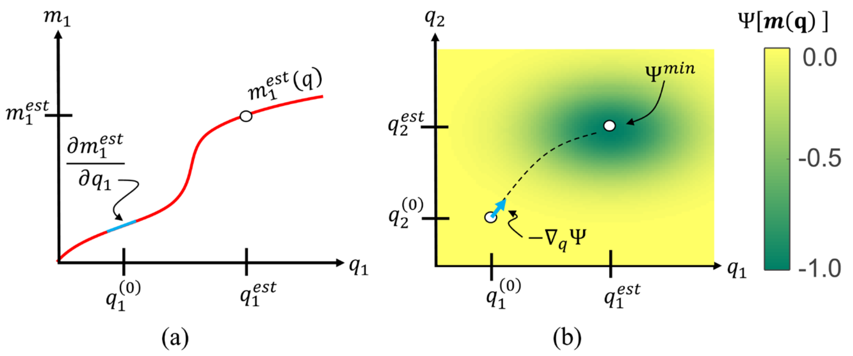


Figure 1. Schematic depiction of solution process. (a) The GLS solution $\boldsymbol{m}^{est}$ (red curve) is considered a function of the covariance parameters $\boldsymbol{q}$ and its derivative $\partial \boldsymbol{m}^{est}/\partial q_n$ (blue line) at a point $\boldsymbol{q}^{(0)}$ is computed by analytic differentiation of GLS equation (1); (b) The objective function Ψ (colors) is considered a function of $\boldsymbol{q}$. The results of (a) are used to compute its gradient $\nabla_q \Psi$ at the point $\boldsymbol{q}^{(0)}$. The gradient descent method is used to iteratively perturb this point anti-parallel to the gradient until it reaches the minimum $\Psi^{\min}$ of the objective function, resulting in the best-estimate $\boldsymbol{q}^{est}$. This value is then used to determine a best-estimate of the model parameters $\boldsymbol{m}^{est}$, as depicted in (a).

Similarly, differentiating $\boldsymbol{M}^{-1/2}\boldsymbol{M}^{-1/2} = \boldsymbol{M}^{-1}$ and applying (10), yields the Sylvester equation:

$$\frac{\partial \boldsymbol{M}^{-1/2}}{\partial q_m}\boldsymbol{M}^{-1/2} + \boldsymbol{M}^{-1/2}\frac{\partial \boldsymbol{M}^{-1/2}}{\partial q_m} = \frac{\partial \boldsymbol{M}^{-1}}{\partial q_m} = -\boldsymbol{M}^{-1}\left[\frac{\partial \boldsymbol{M}}{\partial q_m}\right]\boldsymbol{M}^{-1} \qquad (11)$$

We have not been able to determine a source for this equation, but in all likelihood, it has been derived previously. In practice, (11) is not significantly harder to compute than (10), because efficient algorithms for solving Sylvester equations [26] and for computing a symmetric (principal) square root [27], are widely available and implemented in many software environments, including MATLAB® and PYTHON/linalg. The derivative of $\ln\left(\det \boldsymbol{C}_d\right)$ is derived starting with Jacobi's formula [12]:

$$\frac{\partial \det\left(\boldsymbol{M}\right)}{\partial q} = \operatorname{tr}\left(\operatorname{adj}\left(\boldsymbol{M}\right)\frac{\partial \boldsymbol{M}}{\partial q}\right) = \operatorname{tr}\left(\det\left(\boldsymbol{M}\right)\boldsymbol{M}^{-1}\frac{\partial \boldsymbol{M}}{\partial q}\right) = \det\left(\boldsymbol{M}\right)\operatorname{tr}\left(\boldsymbol{M}^{-1}\frac{\partial \boldsymbol{M}}{\partial q}\right) \qquad (12)$$

where $\operatorname{adj}(.)$ is the adjugate and $\operatorname{tr}(.)$ is the trace, applying Laplace's identify [28] $\operatorname{adj}\left(\boldsymbol{C}_d\right) = \det\left(\boldsymbol{C}_d\right)\boldsymbol{C}_d^{-1}$ and the rule $\operatorname{tr}\left(c\boldsymbol{M}\right) = c\operatorname{tr}\left(\boldsymbol{M}\right)$ (where c is a scalar and $\boldsymbol{M}$ is a matrix) [29]. Finally, the determinant is moved to the left-hand side and the well-known relationship $\partial \ln\left(f\right)/\partial q = f^{-1}\left(\partial f/\partial q\right)$, for a differentiable function $f\left(q\right)$, is applied, yielding ([25], their (38)):

$$\frac{\partial \ln\left(\det \boldsymbol{M}\right)}{\partial q} = \frac{1}{\det\left(\boldsymbol{M}\right)}\frac{\partial \det\left(\boldsymbol{M}\right)}{\partial q} = \operatorname{tr}\left(\boldsymbol{M}^{-1}\frac{\partial \boldsymbol{M}}{\partial q}\right) \qquad (13)$$

We begin the main derivation by considering the case in which data variance $\boldsymbol{C}_d\left(\boldsymbol{q}\right)$ depends on a parameter vector $\boldsymbol{q}$, and the information variance $\boldsymbol{C}_h$ is constant. The derivative of the GLS solution can be found by applying the chain rule applied to (1):

$$\begin{aligned}\frac{\partial \boldsymbol{m}^{est}}{\partial q_m} &= \frac{\partial \boldsymbol{Z}^{-1}}{\partial q_m}\boldsymbol{G}^{\mathrm{T}}\boldsymbol{C}_d^{-1}\boldsymbol{d}^{obs} + \boldsymbol{Z}^{-1}\boldsymbol{G}^{\mathrm{T}}\frac{\partial \boldsymbol{C}_d^{-1}}{\partial q_m}\boldsymbol{d}^{obs} + \frac{\partial \boldsymbol{Z}^{-1}}{\partial q_m}\boldsymbol{H}^{\mathrm{T}}\boldsymbol{C}_h^{-1}\boldsymbol{h}^{pri} \\ &= \boldsymbol{Z}^{-1}\left(\boldsymbol{G}^{\mathrm{T}}\frac{\partial \boldsymbol{C}_d^{-1}}{\partial q_m}\boldsymbol{d}^{obs} - \frac{\partial \boldsymbol{Z}}{\partial q_m}\boldsymbol{m}^{est}\right) \\ &\text{with } \frac{\partial \boldsymbol{Z}}{\partial q_m} = \boldsymbol{G}^{\mathrm{T}}\frac{\partial \boldsymbol{C}_d^{-1}}{\partial q_m}\boldsymbol{G} \text{ and } \frac{\partial \boldsymbol{C}_d^{-1}}{\partial q_m} = -\boldsymbol{C}_d^{-1}\frac{\partial \boldsymbol{C}_d}{\partial q_m}\boldsymbol{C}_d^{-1}\end{aligned} \tag{14}$$

Note that we have used (10). The derivative of the normalized prediction error is $\tilde{\boldsymbol{e}} \equiv \boldsymbol{C}_d^{-1/2}\left(\boldsymbol{d}^{obs} - \boldsymbol{G}\boldsymbol{m}^{est}\right)$ and total error $E \equiv \tilde{\boldsymbol{e}}^{\mathrm{T}}\tilde{\boldsymbol{e}}$ are:

$$\begin{aligned}&\frac{\partial \tilde{\boldsymbol{e}}}{\partial q_m} = -\boldsymbol{C}_d^{-1/2}\boldsymbol{G}\frac{\partial \boldsymbol{m}^{est}}{\partial q_m} + \frac{\partial \boldsymbol{C}_d^{-1/2}}{\partial q_m}\left(\boldsymbol{d}^{obs} - \boldsymbol{G}\boldsymbol{m}^{est}\right) \text{ and } \frac{\partial E}{\partial q_m} = 2\tilde{\boldsymbol{e}}^{\mathrm{T}}\frac{\partial \tilde{\boldsymbol{e}}}{\partial q_m} \\ &\text{with } \frac{\partial \boldsymbol{C}_h^{-1/2}}{\partial q_m}\boldsymbol{C}_h^{-1/2} + \boldsymbol{C}_h^{-1/2}\frac{\partial \boldsymbol{C}_h^{-1/2}}{\partial q_m} = -\boldsymbol{C}_h^{-1}\frac{\partial \boldsymbol{C}_h}{\partial q_m}\boldsymbol{C}_h^{-1}\end{aligned} \tag{15}$$

Here, the Sylvester equation arises from (11). An alternate way of differentiating E that does not require solving a Sylvester equation is:

$$\frac{\partial E}{\partial q_m} = \frac{\partial}{\partial q_m}\left(\boldsymbol{e}^{\mathrm{T}}\boldsymbol{C}_d^{-1}\boldsymbol{e}\right) = -\left(\frac{\partial \boldsymbol{m}^{est}}{\partial q_m}\right)^{\mathrm{T}}\boldsymbol{G}^{\mathrm{T}}\boldsymbol{C}_d^{-1}\boldsymbol{e} + \boldsymbol{e}^{\mathrm{T}}\frac{\partial \boldsymbol{C}_d^{-1}}{\partial q_m}\boldsymbol{e} - \boldsymbol{e}^{\mathrm{T}}\boldsymbol{C}_d^{-1}\boldsymbol{G}\frac{\partial \boldsymbol{m}^{est}}{\partial q_m} \tag{16}$$

The derivative of the normalized error in prior information $\tilde{\boldsymbol{\ell}} = \boldsymbol{C}_h^{-1/2}\left(\boldsymbol{h} - \boldsymbol{H}\boldsymbol{m}^{est}\right)$ and total error $L \equiv \tilde{\boldsymbol{\ell}}^{\mathrm{T}}\tilde{\boldsymbol{\ell}}$ are:

$$\frac{\partial \tilde{\boldsymbol{\ell}}}{\partial q_m} = -\boldsymbol{C}_h^{-1/2}\boldsymbol{H}\frac{\partial \boldsymbol{m}^{est}}{\partial q_m} \text{ and } \frac{\partial L}{\partial q_m} = 2\tilde{\boldsymbol{\ell}}^{\mathrm{T}}\frac{\partial \tilde{\boldsymbol{\ell}}}{\partial q_m} \tag{17}$$

Finally, since $\Psi = \ln\left(\det \boldsymbol{C}_d\right) + \ln\left(\det \boldsymbol{C}_h\right) + E + L$, we have:

$$\frac{\partial \Psi}{\partial q_m} = \frac{\partial \ln\left(\det \boldsymbol{C}_d\right)}{\partial q_m} + \frac{\partial E}{\partial q_m} + \frac{\partial L}{\partial q_m} = \mathrm{tr}\left(\boldsymbol{C}_d^{-1}\frac{\partial \boldsymbol{C}_d}{\partial q_m}\right) + \frac{\partial E}{\partial q_m} + \frac{\partial L}{\partial q_m} \tag{18}$$

Note that we have applied (13).

Finally, we consider the case in which the information variance $\boldsymbol{C}_h(\boldsymbol{q})$ depends on parameters $\boldsymbol{q}$, and $\boldsymbol{C}_d$ is constant. Since the data and prior information play completely symmetric roles in (1), the derivatives can be obtained by interchanging the roles of $\boldsymbol{C}_d$ and $\boldsymbol{C}_h$, $\boldsymbol{G}$ and $\boldsymbol{H}$**,** $\boldsymbol{d}^{obs}$ and $\boldsymbol{h}^{pri}$, $\tilde{\boldsymbol{e}}$ and $\tilde{\boldsymbol{\ell}}$ and E and L, in the equations above, yielding:

$$\begin{aligned}&\frac{\partial \boldsymbol{m}^{est}}{\partial q_m} = \boldsymbol{Z}^{-1}\left(\boldsymbol{H}^{\mathrm{T}}\frac{\partial \boldsymbol{C}_h^{-1}}{\partial q_m}\boldsymbol{h}^{pri} - \frac{\partial \boldsymbol{Z}}{\partial q_m}\boldsymbol{m}^{est}\right) \\ &\text{with } \frac{\partial \boldsymbol{Z}}{\partial q_m} = \boldsymbol{H}^{\mathrm{T}}\frac{\partial \boldsymbol{C}_h^{-1}}{\partial q_m}\boldsymbol{H} \text{ and } \frac{\partial \boldsymbol{C}_h^{-1}}{\partial q_m} = -\boldsymbol{C}_h^{-1}\frac{\partial \boldsymbol{C}_h}{\partial q_m}\boldsymbol{C}_h^{-1} \\ &\frac{\partial \tilde{\boldsymbol{e}}}{\partial q_m} = -\boldsymbol{C}_d^{-1/2}\boldsymbol{G}\frac{\partial \boldsymbol{m}^{est}}{\partial q_m} \text{ and } \frac{\partial E}{\partial q_m} = 2\tilde{\boldsymbol{e}}^{\mathrm{T}}\frac{\partial \tilde{\boldsymbol{e}}}{\partial q_m} \\ &\frac{\partial \tilde{\boldsymbol{\ell}}}{\partial q_m} = -\boldsymbol{C}_h^{-1/2}\boldsymbol{H}\frac{\partial \boldsymbol{m}^{est}}{\partial q_m} + \frac{\partial \boldsymbol{C}_h^{-1/2}}{\partial q_m}\left(\boldsymbol{h}^{pri} - \boldsymbol{H}\boldsymbol{m}^{est}\right)\end{aligned}$$

$$\frac{\partial L}{\partial q_m}=2\tilde{\ell}^{\mathrm{T}}\frac{\partial \tilde{\ell}}{\partial q_m}=-\left(\frac{\partial \boldsymbol{m}^{est}}{\partial q_m}\right)^{\mathrm{T}}\boldsymbol{H}^{\mathrm{T}}\boldsymbol{C}_h^{-1}\ell+\ell^{\mathrm{T}}\frac{\partial \boldsymbol{C}_h^{-1}}{\partial q_m}\ell-\ell^{\mathrm{T}}\boldsymbol{C}_h^{-1}\boldsymbol{H}\frac{\partial \boldsymbol{m}^{est}}{\partial q_m}$$

$$\frac{\partial \boldsymbol{C}_h^{-1/2}}{\partial q_m}\boldsymbol{C}_h^{-1/2}+\boldsymbol{C}_h^{-1/2}\frac{\partial \boldsymbol{C}_h^{-1/2}}{\partial q_m}=-\boldsymbol{C}_h^{-1}\frac{\partial \boldsymbol{C}_h}{\partial q_m}\boldsymbol{C}_h^{-1}$$

$$\frac{\partial \ln\left(\det \boldsymbol{C}_h\right)}{\partial q_m}=\operatorname{tr}\left(\boldsymbol{C}_h^{-1}\frac{\partial \boldsymbol{C}_h}{\partial q_m}\right)$$

$$\frac{\partial \Psi}{\partial q_m}=\operatorname{tr}\left(\boldsymbol{C}_h^{-1}\frac{\partial \boldsymbol{C}_h}{\partial q_m}\right)+\frac{\partial E}{\partial q_m}+\frac{\partial L}{\partial q_m} \tag{19}$$

These formulas have been checked numerically.

4. Examples with Discussion

In the first example, we examine the simplistic case in which the parameter q represents an overall scaling of variance; that is $\boldsymbol{C}_d(q)=q\boldsymbol{C}_d^{(0)}$ and $\boldsymbol{C}_h(q)=q\boldsymbol{C}_h^{(0)}$, with specified $\boldsymbol{C}_d^{(0)}$ and $\boldsymbol{C}_h^{(0)}$. The solution $\boldsymbol{m}^{est}$ is independent of q, as can be verified by substitution into (1). The parameter q can then be found by direct minimization of (9), which simplifies to:

$$\Psi=\ln\left(q^N\det \boldsymbol{C}_d^{(0)}\right)+\ln\left(q^K\det \boldsymbol{C}_k^{(0)}\right)+q^{-1}E_0+q^{-1}L_0 \tag{20}$$

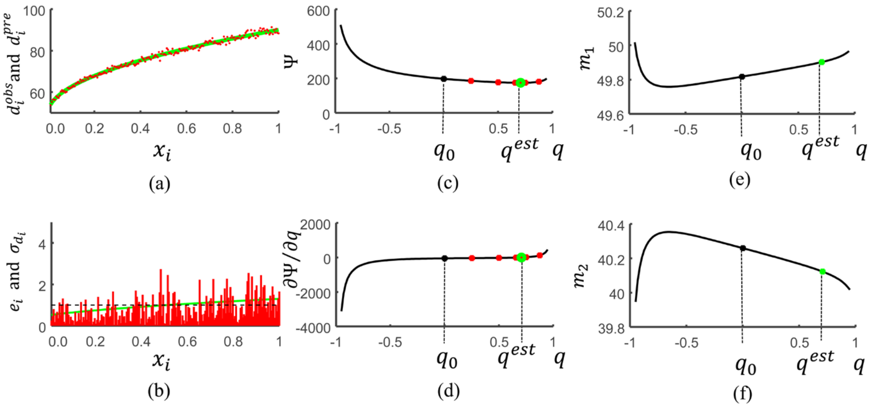
Here, we have used the rule $\det(q\boldsymbol{M})=q^N\det(\boldsymbol{M})$ [25], valid for any $N\times N$ matrix $\boldsymbol{M}$, and have defined $E_0\equiv E(q=1)$ and $L_0\equiv L(q=1)$. The minimum occurs when:

$$\frac{\partial \Psi}{\partial q}=0=(N+K)q^{-1}-(E_0+L_0)q^{-2}\ \text{ or }\ q=\frac{E_0+L_0}{N+K} \tag{21}$$

This is a generalization of the well-known maximum likelihood estimate of the sample variance [30]. As long as (E_0+L_0) exists, the minimization in (21) is well-behaved and the overall scaling q is uniquely determined.

In the second example, we examine another simplistic case in which a parameter q represents the relative weighting of variance; that is $\boldsymbol{C}_d^{-1}(q)=q\boldsymbol{I}$ and $\boldsymbol{C}_h^{-1}(q)=(1-q)\boldsymbol{I}$ .We consider the problem of estimating the mean m_1 of data given observations $\boldsymbol{d}=\boldsymbol{1}$ and prior information $\boldsymbol{h}=\boldsymbol{0}$ (where $\boldsymbol{0}$ and $\boldsymbol{1}$ are vectors of zeros and ones, respectively), when $N=K$, $M=1$ and $\boldsymbol{G}=\boldsymbol{H}=\boldsymbol{1}$. Applying (1), we find that $m^{est}=q$. Then, the objective function is $\Psi=\ln\left(q^N\right)+\ln\left((1-q)^N\right)+Nq(1-q)$ and its derivative is $\partial\Psi/\partial q=N\left[-q^{-1}+(1-q)^{-1}+(1-q)-q\right]$. The solution to $\partial\Psi/\partial q=0$ is $q^{est}=1/2$, as can be verified by direct substitution. Thus, the solution splits the difference between the observations and the prior values, and yields prior variances $\boldsymbol{C}_d$ and $\boldsymbol{C}_h$ that are equal. While simplistic, this problem illustrates that, at least in some cases, GLS is capable of uniquely determining the relative sizes of $\boldsymbol{C}_d$ and $\boldsymbol{C}_h$. Because trade-off curves, as defined in the Introduction, are based on the behavior of E and L, and not the complete objective function Ψ, the weighting parameter q_0 estimated from them in general will be different from q^{est}.Consequently, the trade-off curve procedure is not consistent with the Bayesian framework upon which GLS rests.

Our third example demonstrates the tuning of data covariance $\boldsymbol{C}_d(\boldsymbol{q})$. In many cases, observational error increases during the course of an experiment, due to degradation of equipment or to worsening environmental conditions. The example demonstrates that the method is capable of accurately quantifying the fractional rate of increase p of the variance σ_{d_n}, which is assumed to vary with position x_n. In our simulation, we consider $N = 201$ synthetic data, evenly-spaced on the interval $0 \le x_i \le 1$, which scatter around the curve $d_i = m_1 + m_2 x_i^{1/2}$ (**Figure 2**). The covariance of the data is modeled as $[\boldsymbol{C}_d]_{mn} = \sigma_{d_n}^2 \delta_{mn}$, where $\sigma_{d_n} = (1)^2 (1 + q(2x_n - 1))$ and δ_{mn} is the Kronecker delta; that is, the data are uncorrelated and their variance increases linearly with *x*. The derivative of the covariance is $(\partial/\partial q)[\boldsymbol{C}_d]_{mn} = (1)^2 (2x_n - 1)\delta_{mn}$. We have included prior information with $\boldsymbol{H} = \boldsymbol{I}$ and $\boldsymbol{h}^{pri} = 0$, which implements the notion that the model parameters are small. The corresponding covariance is chosen to be large, $\boldsymbol{C}_h = (1000)^2 \boldsymbol{I}$, indicating that this information is weak. The goal is to tune the rate of increase of variance and to arrive at a best-esti- mate of the two model parameters. The starting value is taken to be $q_0 = 0$, which corresponds to uniform variance. It is successively improved by a gradient descent method that minimizes Ψ, yielding an estimated value $q^{est} \approx 0.709$.This estimate differs from the true value $q^{true} = 0.700$ by about 1%. The estimated solution $\boldsymbol{m}^{est}$ differs from $\boldsymbol{m}(q = 0)$ by a few tenths of a percent, which may be significant in some applications.



(a) (c) (e)

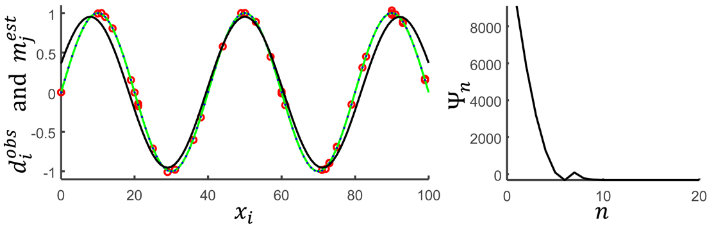
(b) (d) (f)

Figure 2. Example of tuning $\boldsymbol{C}_d(q)$. (a) Plot of synthetic data (red dots) and predicted data (green curve); (b) The starting value $q_0 = 0$ corresponds to uniform variance (black curve). The estimate q^{est} corresponds to increasing variance (green curve); (c) Generalized error $\Phi(q)$ (black curve). The starting value q_0 (black circle) is successively improved (red circles) by a gradient descent method, yielding an estimate q^{est} (green circle); (d) The gradient $\partial\Phi/\partial q$, computed using the formulas developed in the text; (e) The first model parameter $m_1(q)$, highlighting the initial value (black circle) and estimated value (green circle) (f) Same as (e), except for the second model parameter $m_2(q)$.

The fourth example demonstrates tuning of information covariance $\boldsymbol{C}_h(\boldsymbol{q})$. In many instances, one may need to "reconstruct" or "interpolate" a function on the basis of unevenly and sparsely sampled data. In this case, prior information on the autocovariance of the function can enable a smooth interpolation. Furthermore, it can enforce a covariance structure that may be required, say, by the underlying physics of the problem. In our example, we suppose that the function is known to be oscillatory on physical grounds, but that the wavenumber of those oscillations is known only imprecisely. The goal is to tune prior knowledge of wavenumber to arrive at a best-estimate of the reconstructed function. In our simulation, a total of $M = 101$ model parameters m_j are uniformly spaced on the interval $0 \le x \le 100$ and representing a sampled version of a continuous, sinusoidal function $m(x)$ with wavenumber $p^{true} = 0.1571$ (Figure 3). Synthetic data d_i^{obs} with uncorrelated error with variance $\sigma_d^2 = (0.01)^2$ are available for $N = 40$ randomly-chosen points $x_{j(i)}$, where the index function $j(i)$ aligns in x observations to model parameters. The data kernel is $G_{ij} = \delta_{i,j(i)}$. The prior information is given in (4), with autocovariance $[\boldsymbol{C}_h]_{nm} = \sigma_h^2 \cos(q|x_n - x_m|)$ and $\sigma_h^2 = (10)^2$. The derivative is $(\partial/\partial q)[\boldsymbol{C}_h]_{nm} = -\sigma_h^2 |x_n - x_m| \sin(q|x_n - x_m|)$. An initial guess $p_0 = 0.95 p^{true}$ is improved using a gradient descent method, yielding an estimated value of $p^{est} = 0.1571$ that differs from p^{true} by less than 0.01%. The reconstructed function is smooth and sinusoidal and the fit to the data is much improved.

Examples three and four were implemented in MATLAB® and executed in <5s on a notebook computer. They confirm the flexibility, speed and effectiveness of the method. An ability to tune prior information on autocovariance may be of special utility in seismic exploration applications, where three-dimensional waveform datasets are routinely interpolated.

A limitation of this overall "parametric" approach is that the solution is dependent on the choice of parameterization, which must be guided by prior knowledge of the general properties of the covariance matrices in particular problem being solved. In Example 3, we were able to recognize (say, by visually



(a) (b)

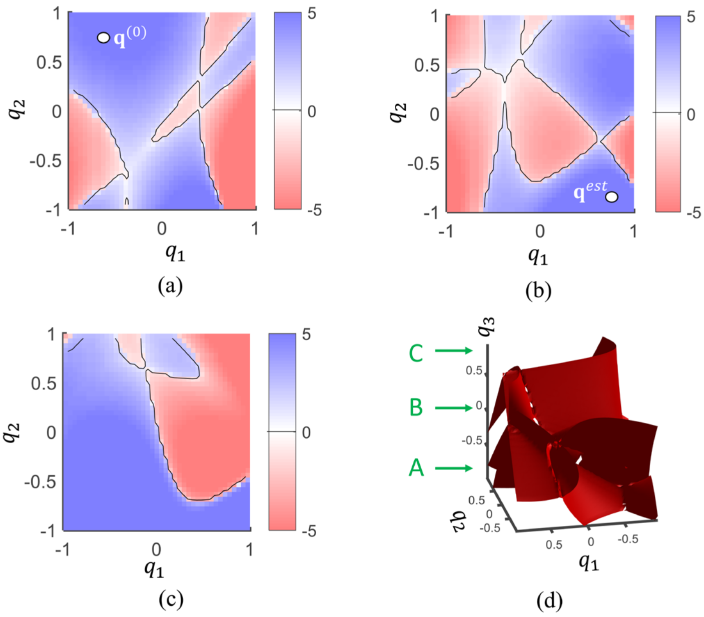
Figure 3. Example of tuning $\boldsymbol{C}_h(q)$. Sparsely-sampled synthetic data d_i^{obs} (red dots) are oscillatory. (a) A regularly-sampled version m_j^{est} is created by imposing the oscillatory covariance $[\boldsymbol{C}_h]_{nm} = \sigma_h^2 \cos(q|x_n - x_m|)$. With the starting value $q_0 = 0.9500 q^{true}$, the reconstruction poorly fits the data (black curve). Tuning leads to a better fit (green curve with dots), as well as a precise estimate of wavenumber $q_0 \approx 0.9999 q^{true}$; (b) Decrease in Ψ_n with iteration number n during the gradient descent process.

examining the data plotted in Figure 2(a)) that observational error increases with x and chose $\left[\boldsymbol{C}_d\right]_{mn} = \sigma_d^2\left(1+q\left(x_n - x_1\right)\right)\delta_{mn}$ that matched this scenario. If, instead, the degree of correlation between successive data increased with x, this pattern might be less expected, more difficult to detect, and require a different parameterization—say, $\left[\boldsymbol{C}_d\left(q\right)\right]_{nm} = \sigma_d^2 \exp\left[-\frac{1}{2}q\left(x_n + x_m\right)\left|x_n - x_m\right|\right]$.

Not every parameterization of $\boldsymbol{C}_d$ (or $\boldsymbol{C}_h$) is necessarily well-behaved. To avoid poor behavior, the parameterization must be chosen so its determinant does not have zeros at values of $\boldsymbol{q}^{est}$ that will prevent the steepest descent process from converging to the global minimum. That this choice can be problematical is illustrated by the simple Toeplitz version of $\boldsymbol{C}_d$ (with $N = 10$, $J = 9$):

$$\boldsymbol{C}_d = \begin{bmatrix} 1 & q_1 & q_2 & q_3 & \cdots & q_9 \\ q_1 & 1 & q_1 & q_2 & \cdots & q_8 \\ q_2 & q_1 & 1 & q_1 & \cdots & q_7 \\ q_3 & q_2 & q_1 & 1 & \cdots & q_6 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ q_9 & q_8 & q_7 & q_6 & \cdots & 1 \end{bmatrix} \tag{22}$$

with $\left|q_i\right| < 1$. This form is useful for quantifying correlations within a stationary sequence of data [31]. Yet as is illustrated in Figure 4, the $\mathbb{R}^J$ volume is crossed



(a)

(b)

(c)

(d)

Figure 4. The function $\det \boldsymbol{C}_d\left(\boldsymbol{q}\right) = 0$ for the case given by (22). (a) The $\left(q_1, q_2\right)$ surface for $q_3 = -0.95$ and the other qs randomly assigned; (b) Same as (a), but with $q_3 = 0.00$; (c) Same as (a), but with $q_3 = 0.95$; (d) Perspective view of the surfaces in the q_1, q_2, q_3 volume. The positions of the three slices in (a), (b) and (c) are noted on the q_3-axis (green arrows). A question posed in the text is whether, given an arbitrary point $\boldsymbol{q}^{(0)}$ and the global minimum of the objective function, say at $\boldsymbol{q}^{est}$ (and with both points satisfying $\det \boldsymbol{C}_d > 0$), a steepest-descent path necessarily exists between them.

by many $\det \boldsymbol{C}_d = 0$ surfaces that correspond to surfaces of singular objective function Ψ. Their presence suggests that the steepest descent path between a starting value $\boldsymbol{q}^{(0)}$ and the global minimum at $\boldsymbol{q}^{est}$ may be very convoluted (if, indeed, such a path exists) unless $\boldsymbol{q}^{(0)}$ is very close to $\boldsymbol{q}^{est}$.

5. Conclusion

Generalized Least Squares requires the assignment of two prior covariance matrices, the prior covariance of the data and the prior covariance of the prior information. Making these assignments is often a very subjective process. However, in cases in which the forms of these matrices can be anticipated up to a set of poorly-known parameters, information contained within the data and prior information can be used to improve knowledge of them—a process we call "tuning". Tuning can be achieved by minimizing an objective function that depends on both the generalized error and determinants of the covariance matrices to arrive at a best estimate of the parameters. Analytic and computationally-tractable formulas are derived for the derivative needed to implement the minimization via a gradient descent method. Furthermore, the problem is organized so that the minimization need be performed only over the space of covariance parameters, and not over the typically-much-larger space of model and covariance parameters. Although some care needs to be exercised as the covariance matrices are parametrized, the minimization is tractable and can lead to better estimates of the model parameters. An important outcome is this study is the recognition that the use of trade-off curves to determine relative weighting of covariance—a practice ubiquitous in the geophysical imaging—is not consistent with the underlying Bayesian framework of Generalized Least Squares. The strategy outlined here provides a consistent solution.

Acknowledgements

The author thanks Roger Creel for helpful discussion.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Tarantola, A. and Valette, B. (1982) Generalized Non-Linear Inverse Problems Solved Using the Least Squares Criterion. *Reviews of Geophysics and Space Physics*, **20**, 219-232. https://doi.org/10.1029/RG020i002p00219

[2] Tarantola, A. and Valette, B. (1982) Inverse Problems = Quest for Information. *Journal of Geophysics*, **50**, 159-170. https://n2t.net/ark:/88439/y048722

[3] Menke, W. (2018) Geophysical Data Analysis: Discrete Inverse Theory. 4th Edition, Elsevier, 350 p.

[4] Menke, W. and Menke, J. (2016) Environmental Data Analysis with MATLAB. 2nd Edition, Elsevier, 3342 p. https://doi.org/10.1016/B978-0-12-804488-9.00001-X

[5] Tarantola, A. (2005) Inverse Problem Theory and Methods for Model Parameter Estimation. SIAM: Society for Industrial and Applied Mathematics, 342 p. https://doi.org/10.1137/1.9780898717921

[6] Menke, W. (2014) Review of the Generalized Least Squares Method. *Surveys in Geophysics*, **36**, 1-25. https://doi.org/10.1007/s10712-014-9303-1

[7] Abers, G. (1994) Three-Dimensional Inversion of Regional P and S Arrival Times in the East 723 Aleutians and Sources of Subduction Zone Gravity Highs. *Journal of Geophysical Research*, **99**, 4395-4412. https://doi.org/10.1029/93JB03107

[8] Schmandt, B. and Lin, F.-C. (2014) *P* and *S* Wave Tomography of the Mantle beneath the United States. *Geophysical Research Letters*, **41**, 6342-6349. https://doi.org/10.1002/2014GL061231

[9] Menke, W. (2005) Case Studies of Seismic Tomography and Earthquake Location in a Regional Context. Geophysical Monograph 157. American Geophysical Union, Washington DC. https://doi.org/10.1029/157GM02

[10] Nettles, M., and Dziewonski, A.M. (2008) Radially Anisotropic Shear Velocity Structure of the Upper Mantle Globally and Beneath North America. *Journal of Geophysical Research*, **113**, B02303. https://doi.org/10.1029/2006JB004819

[11] Chen, W. and Ritzwoller, M.H. (2016) Crustal and Uppermost Mantle Structure Beneath the United States. *Journal of Geophysical Research*, **121**, 4306-4342. https://doi.org/10.1002/2016JB012887

[12] Humphreys, E.D., Dueker, K.G., Schutt, D.L. and Smith, R.B. (2000) Beneath Yellowstone: Evaluating Plume and Nonplume Models Using Teleseismic Images of the Upper Mantle. *GSA Today*, **10**, 1-7. https://www.geosociety.org/gsatoday/archive/10/12/

[13] Gillet, N., Schaeffer, N. and Jault, D. (2011) Rationale and Geophysical Evidence for Quasi-Geostrophic Rapid Dynamics within the Earth's Outer Core. *Physics of the Earth and Planetary Interiors*, **187**, 380-390. https://doi.org/10.1016/j.pepi.2011.01.005

[14] Zhao, S. (2013) Lithosphere Thickness and Mantle Viscosity Estimated from Joint Inversion of GPS and GRACE-Derived Radial Deformation and Gravity Rates in North America. *Geophysical Journal International*, **194**, 1455-1472. https://doi.org/10.1093/gji/ggt212

[15] Menke, W. and Eilon, Z. (2015) Relationship between Data Smoothing and the Regularization of Inverse Problems. *Pure and Applied Geophysics*, **172**, 2711-2726. https://doi.org/10.1007/s00024-015-1059-0

[16] Voorhies, C.F. (1986) Steady Flows at the Top of Earth's Core Derived from Geomagnetic Field Models. *Journal of Geophysical Research*, **91**, 12444-12466. https://doi.org/10.1029/JB091iB12p12444

[17] Yao, Z.S. and Roberts, R.G. (1999) A Practical Regularization for Seismic Tomography. *Geophysical Journal International*, **138**, 293-299. https://doi.org/10.1046/j.1365-246X.1999.00849.x

[18] Snyman, J.A. and Wilke, D.N. (2018) Practical Mathematical Optimization—Basic Optimization Theory and Gradient-Based Algorithms. Springer Optimization and Its Applications, 2nd Edition, Springer, New York, 340 p.

[19] Hidebrand, F.B. (1987) Introduction to Numerical Analysis. 2nd Edition, Dover Publications, New York.

[20] Zaroli, C., Sambridge, M., Lévêque, J.-J., Debayle, E. and Nolet, G. (2013) An Objective Rationale for the Choice of Regularization Parameter with Application to Glob-

al Multiple-Frequency S-Wave Tomography. *Solid Earth*, **4**, 357-371. https://doi.org/10.5194/se-4-357-2013

[21] Malinverno, A. and Parker, R.L. (2006) Two Ways to Quantify Uncertainty in Geophysical Inverse Problems. *Geophysics*, **71**, W15-W27. https://doi.org/10.1190/1.2194516

[22] Malinverno, A. and Briggs, V.A. (2004) Expanded Uncertainty Quantification in Inverse Problems: Hierarchical Bayes and Empirical Bayes. *Geophysics*, **69**, 877-1103. https://doi.org/10.1190/1.1778243

[23] Box, G.E.P. and Tiao, G.C. (1992) Bayesian Inference in Statistical Analysis. Wiley, New York, 589 p. https://doi.org/10.1002/9781118033197

[24] Schmidt, E. (1973) Cholesky Factorization and Matrix Inversion, National Oceanic and Atmospheric Administration Technical Report NOS-56. US Government Printing Office, Washington DC. https://books.google.com/books?id=MiRHAQAAIAAJ

[25] Petersen, K.B. and Pedersen, M.S. (2008) The Matrix Cookbook, 71 p. https://archive.org/details/imm3274

[26] Bartels, R.H. and Stewart, G.W. (1972) Solution of the matrix equation AX + XB = C. *Communications of the ACM*, **15**, 820-826. https://doi.org/10.1145/361573.361582

[27] Higham, N.J. (1987) Computing Real Square Roots of a Real Matrix. *Linear Algebra and its Applications*, **88-89**, 405-430. https://doi.org/10.1016/0024-3795(87)90118-2

[28] Magnus, J.R. and Neudecker, H. (1999) Matrix Differential Calculus with Applications in Statistics and Econometrics, Revised Edition. John Wiley and Sons, New York, 424 p.

[29] Gantmacher, F.R. (1960) The Theory of Matrices, Volume 1. Chelsea Publishing, New York, 374 p.

[30] Fisher, R.A. (1925) Theory of Statistical Estimation. *Mathematical Proceedings of the Cambridge Philosophical Society*, **22**, 700-725. https://doi.org/10.1017/S0305004100009580

[31] Claerbout, J.F. (1985) Fundamentals of Geophysical Data Processing with Applications to Petroleum Prospecting. Blackwell Scientific Publishing, Oxford, UK, 267 p.