Geophysical Journal International

Geophys. J. Int. (2020) **220,** 1648–1656
Advance Access publication 2019 November 19
GJI General Geophysical Methods

doi: 10.1093/gji/ggz521

Gaussian process models—II. Lessons for discrete inversion

Andrew P. Valentine and Malcolm Sambridge

Research School of Earth Sciences, The Australian National University, 142 *Mills Road, Acton, ACT* 2601*, Australia. E-mail: andrew.valentine@anu.edu.au*

Accepted 2019 November 16. Received 2019 October 23; in original form 2019 May 30

SUMMARY

By starting from a general framework for probabilistic continuous inversion (developed in Part I) and introducing discrete basis functions, we obtain the well-known algorithms for probabilistic least-squares inversion set out by Tarantola & Valette. In doing so, we establish a direct equivalence between the spatial covariance function that must be specified in continuous inversion, and the combination of basis functions and prior covariance matrix that must be chosen for discretized inversion. We show that the common choice of Tikhonov regularization ($\mathbf{C}_\mathbf{m}^{-1} = \sigma^2\mathbf{I}$) arises from a delta-function spatial covariance, and that this lies behind many of the artefacts commonly associated with discretized inversion. We show that other choices of spatial covariance function can be used to generate regularization matrices yielding substantially better results, and permitting localization of features even if global basis functions are used. We are also able to offer a straightforward explanation for the spectral leakage problem identified by Trampert & Snieder.

Key words: Inverse theory; Probability distributions; Statistical methods.

1 INTRODUCTION

Discrete inverse theory—particularly inversion based upon the least-squares algorithm—underpins much research in geophysics. As practitioners are well-aware, many choices must be made when setting up and solving an inverse problem, and different decisions can lead to different results (e.g. Boschi & Dziewoński 1999; Valentine & Trampert 2016). Particularly influential are choices surrounding model parametrization, and regularization, which together determine many of the characteristics of the eventual solution. In deciding to use a particular finite set of basis functions, we impose strong restrictions on the range of features that may appear within the recovered model; in selecting a regularization scheme, we bias the inversion towards solutions of a certain style. Clearly, understanding the consequences of such decisions is essential if we are to construct useful models, and interpret them correctly.

In Part I of this two-part paper (Valentine & Sambridge 2019), we set out a framework for continuous inverse theory, building on the statistical theory of Gaussian Processes (GPs). This circumvents the need to identify an appropriate set of basis functions prior to inversion. Instead, the solution is represented by a particular class of stochastic process, with prior information specified via a user-defined covariance function which encodes information about—for example—the characteristic length scales and correlations desired for the recovered model. Helpfully, this covariance function is defined relative to physical space, rather than an abstract 'model space', potentially making it easier to appreciate the relationship between prior information and results. However, the computational cost of the approach scales with the number of data, potentially limiting its application to large-scale problems.

In this paper—Part II—we demonstrate that if the GP approach is approximated by expanding quantities relative to a finite set of a basis functions, the familiar equations of probabilistic, discrete least-squares (e.g. Tarantola & Valette 1982) emerge. In particular, there is a duality between the choice of covariance function in the GP framework, and the twin choices of parametrization and regularization required within the discrete framework. Thus, for any choice of regularization and parametrization, one may compute the implied covariance function to assist with interpretation; equally, one may compute the regularization matrix required to impose a particular correlation function onto the recovered model. This indicates that it is possible to access the practical benefits associated with the GP approach, while continuing to work within a computationally efficient parametrized setting, and offers new strategies for approaching regularization in discrete settings. The equivalence between the two approaches is also helpful from a theoretical perspective, allowing analysis to be carried out in the continuous domain and then discretized.

In particular, a common challenge in the geosciences involves estimation of spectral properties of (largely) continuous functions, such as gravitational or magnetic fields, or the internal seismic velocity structure of the Earth. This is commonly approached by expressing the unknown field in terms of a set of basis functions (often spherical harmonics), using available data to constrain the expansion coefficients (e.g. Whaler & Gubbins 1981; Woodhouse & Dziewonski 1984; Reigber *et al.* 2005; Hoggard *et al.* 2016; Davies *et al.* 2019). However, a number of practical difficulties arise: it is difficult to ensure robust results in circumstances where the signal of interest may have unknown spectral bandwidth, and where observational constraints are often incompletely and unevenly distributed.

Much research has focussed on attempting to quantify and mitigate the errors and biases that may arise (e.g. Kaula 1967; Trampert & Snieder 1996; Simons *et al.* 2002; Schachtschneider *et al.* 2010); and ensuring that results are comparable across multiple data sets (e.g. Slobbe *et al.* 2012). The opportunity to analyse such issues without discretization as an initial step offers new insights, and potentially new strategies for ensuring robust results.

In order to make this paper accessible to readers who have not studied Part I, we begin by restating a number of key results and definitions from that paper. We then show that these can be used to derive the results of Tarantola & Valette (1982): in effect, the GP framework may be regarded as the limiting case of Tarantola & Valette (1982) as the model-space dimension tends to infinity. We discuss how this equivalence may be used to inform the regularization of least-squares inversion, and demonstrate that the common choice of Tikhonov regularization has undesirable, but avoidable, consequences. Finally, we consider the 'spectral leakage' problem identified by Trampert & Snieder (1996) and others, and show that this can be readily understood within our framework.

2 THEORETICAL DEVELOPMENT

We will assume that $f(x)$ is some function that is sought through inversion: although our notation suggests that this is a scalar function of a single variable, all our analysis translates straightforwardly to higher-dimensional settings. Our data set consists of N measurements, $d_{1\ldots N}$, which are assumed to be related to $f(x)$ by (*cf.* Part I eq. 10)

$$d_i = \int_{\mathcal{X}} w_i(x) f(x)\,\mathrm{d}x\,. \tag{1}$$

Our prior knowledge about $f(x)$ is that $f(x) \sim \mathcal{GP}(\mu(x), k(x,x'))$, for some mean function μ and covariance function k—that is, f is described by a GP, essentially an extension of the familiar Gaussian distribution into function spaces. Conditioning this prior upon the data allows us to obtain a posterior distribution (Part I eqs 14–15)

$$f(x)\,|\,\mathbf{d} \sim \mathcal{GP}(\tilde{\mu}(x), \tilde{k}(x,x')), \tag{2a}$$

where

$$\tilde{\mu}(x) = \mu(x) + \hat{\mathbf{w}}^{\mathbf{T}}(x)\left(\hat{\mathbf{W}} + \mathbf{C_d}\right)^{-1}\left(\hat{\mathbf{d}} - \hat{\boldsymbol{\omega}}\right) \tag{2b}$$

$$\tilde{k}(x,x') = k(x,x') - \hat{\mathbf{w}}^{\mathbf{T}}(x)\left(\hat{\mathbf{W}} + \mathbf{C_d}\right)^{-1}\hat{\mathbf{w}}(x')\,, \tag{2c}$$

with tildes used to denote posterior quantities, and (Part I ,eq. 13)

$$[\hat{\boldsymbol{\omega}}]_i = \int_{\mathcal{X}} w_i(u)\mu(u)\,\mathrm{d}u \tag{3a}$$

$$[\hat{\mathbf{w}}(x)]_i = \int_{\mathcal{X}} w_i(u)k(u,x)\,\mathrm{d}u \tag{3b}$$

$$\left[\hat{\mathbf{W}}\right]_{ij} = \iint_{\mathcal{X}^2} w_i(u)k(u,v)w_j(v)\,\mathrm{d}u\,\mathrm{d}v\,. \tag{3c}$$

The 'hats' on these quantities are used to emphasize that they depend upon the details of the observations made, as the w_i may vary according to—for example—the locations at which data is collected. For a full discussion and derivation of the above results, the reader is referred to Part I.

2.1 Expansion in a finite basis

We now introduce a set of M basis functions, $\phi_i(x)$, $i = 1\ldots M$, defined appropriately for the space of interest. For notational convenience, we collect these into a single vector function, $\boldsymbol{\Phi}(x)$, such that $[\boldsymbol{\Phi}(x)]_i = \phi_i(x)$. The Gram matrix for this basis is denoted by $\boldsymbol{\Gamma}$, defined such that

$$\Gamma_{ij} = \int \phi_i(x)\phi_j(x)\,\mathrm{d}x\,, \tag{4}$$

and we assume that the inverse of this matrix can be stably found. Any function $\zeta(x)$ can be expressed as an expansion relative to this basis,

$$\zeta(x) \approx \sum_i \zeta_i \phi_i(x), \tag{5}$$

where the approximation arises because $\zeta(x)$ will—in general—contain features that cannot be expressed within the finite-dimensional basis. The coefficients ζ_i can be computed as

$$\zeta_i = \sum_j \left[\boldsymbol{\Gamma}^{-1}\right]_{ij} \int \zeta(x)\phi_j(x)\,\mathrm{d}x\,. \tag{6}$$

Similarly, a function of two spatial variables, $\xi(x, x')$ can be expressed using a double expansion,

$$\xi(x,x') \approx \sum_{ij} \Xi_{ij}\phi_i(x)\phi_j(x'), \tag{7}$$

where

$$\Xi_{ij} = \sum_{kl} \left[\boldsymbol{\Gamma}^{-1}\right]_{ik}\left[\boldsymbol{\Gamma}^{-1}\right]_{jl} \iint \xi(x,x')\phi_k(x)\phi_l(x')\,\mathrm{d}x\,\mathrm{d}x'\,. \tag{8}$$

For simplicity, and because it is overwhelmingly the case encountered in practice, we will henceforth assume that the basis functions used are orthonormal. This implies that $\boldsymbol{\Gamma} = \mathbf{I}$, and hence the inverse Gram matrix can be dropped from the expressions for the expansion coefficients.

2.1.1 *Exact expansion*

Since linear transformations preserve Gaussian statistics, it is apparent that expressing the GP solution in the basis should result in a Gaussian distribution of model coefficients. We therefore write $f(x)\,|\,\mathbf{d} \sim \boldsymbol{\Phi}^{\mathbf{T}}(x)\cdot\mathcal{N}(\tilde{\mathbf{m}}, \tilde{\mathbf{C}})$. Applying eq. (6), we see that

$$\tilde{m}_i = m_i + \sum_j \psi_{ij}\left[\left(\hat{\mathbf{W}} + \mathbf{C_d}\right)^{-1}\left(\hat{\mathbf{d}} - \hat{\boldsymbol{\omega}}\right)\right]_j \tag{9a}$$

$$\tilde{C}_{ij} = C_{ij} - \sum_{kl} \psi_{ik}\psi_{jl}\left[\left(\hat{\mathbf{W}} + \mathbf{C_d}\right)^{-1}\right]_{kl}, \tag{9b}$$

where m_i and C_{ij} are the expansion coefficients for the prior and covariance functions, respectively, and where ψ_{ij} represents the expansion coefficients of $\hat{\mathbf{w}}(x)$ within the basis,

$$m_i = \int \mu(x)\phi_i(x)\,\mathrm{d}x \tag{10a}$$

$$C_{ij} = \iint \phi_i(x)k(x,x')\phi_j(x')\,\mathrm{d}x\,\mathrm{d}x' \tag{10b}$$

$$\psi_{ij} = \iint \phi_i(x)k(x,u)w_j(u)\,\mathrm{d}x\,\mathrm{d}u\,. \tag{10c}$$

Using these expressions, the exact expansion of the GP solution within any set of basis functions may be found. Note that our use of the word 'exact' here does not preclude some truncation error: the GP solution may include features that are not representable using the finite set of basis functions. However, the coefficients are exact in the sense that they accurately represent that part of the GP solution that *is* representable: they will not change if the dimension of the basis set is altered.

2.1.2 Approximate expansion

To proceed further, we wish to expand the functions $k(x, x')$ and $w_i(x)$ in terms of the finite basis. In general, these will have some component that cannot be expressed within the basis, and we therefore write

$$k(x, x') = \sum_{ij} C_{ij}\phi_i(x)\phi_j(x') + \epsilon_k(x, x') \tag{11a}$$

$$w_i(x) = \sum_{j} G_{ij}\phi_j(x) + \epsilon_w^{(i)}(x) \tag{11b}$$

with the ϵ terms representing the part of the function that lies outside the basis. It is then possible to show that

$$\psi_{ij} = \left[\mathbf{CG^T} + \mathbf{P}\right]_{ij} \tag{12a}$$

$$\left[\hat{\mathbf{W}}\right]_{ij} = \left[\mathbf{GCG^T} + \mathbf{GP} + (\mathbf{GP})^\mathbf{T} + \mathbf{Q}\right]_{ij}, \tag{12b}$$

where we have introduced $\mathbf{P}$ and $\mathbf{Q}$ having elements

$$P_{ij} = \iint \phi_i(u)\epsilon_k(u, v)\epsilon_w^{(j)}(v)\,\mathrm{d}u\,\mathrm{d}v \tag{12c}$$

$$Q_{ij} = \iint \epsilon_w^{(i)}(u)\epsilon_k(u, v)\epsilon_w^{(j)}(v)\,\mathrm{d}u\,\mathrm{d}v \tag{12d}$$

to represent the 'error terms'. We emphasize that these quantities are not necessarily 'small': their significance depends entirely on the details of the problem under consideration.

If, nevertheless, $\mathbf{P}$ and $\mathbf{Q}$ are deemed negligible in a particular case—in other words, we think that the basis is adequate for representing all quantities—we obtain

$$\tilde{\mathbf{m}}' = \mathbf{m} + \mathbf{CG^T}\left(\mathbf{GCG^T} + \mathbf{C_d}\right)^{-1}\left(\hat{\mathbf{d}} - \mathbf{Gm}\right) \tag{13a}$$

$$\tilde{\mathbf{C}}' = \mathbf{C} - \mathbf{CG^T}\left(\mathbf{GCG^T} + \mathbf{C_d}\right)^{-1}\mathbf{GC}, \tag{13b}$$

where we have used primed quantities are to distinguish between approximate and exact theories. Eq. (13) will no doubt appear familiar to many readers: it has the form of the well-known result from Tarantola & Valette (1982) for least-squares inversion 'in the data space', for the case where the prior on model coefficients is given by $\mathcal{N}(\mathbf{m}, \mathbf{C})$. As those authors show, application of the Woodbury matrix identity allows these expressions to be transformed into 'the model space',

$$\tilde{\mathbf{m}}' = \mathbf{m} + \left(\mathbf{G^T C_d^{-1} G} + \mathbf{C^{-1}}\right)^{-1}\mathbf{G^T C_d^{-1}}\left(\hat{\mathbf{d}} - \mathbf{Gm}\right) \tag{14a}$$

$$\tilde{\mathbf{C}}' = \left(\mathbf{G^T C_d^{-1} G} + \mathbf{C^{-1}}\right)^{-1}. \tag{14b}$$

This form is often more amenable to computations, since the matrix to be inverted is typically smaller.

3 DISCUSSION

Some consequences of the analysis thus far deserve to be highlighted:

(i) The work of Tarantola & Valette (1982) can be viewed as the finite-dimensional, discretized, analogue of the probabilistic, continuous inverse theory developed in Part I;

(ii) Any continuous inversion can be represented *approximately* within a finite-dimensional basis, by adopting $\mathbf{m}$ and $\mathbf{C}$ according to eq. (10) and using the results of Tarantola & Valette (1982);

(iii) Any discrete inversion can be replicated *exactly* within the continuous framework of Part I, by choosing $\mu(x) = \mathbf{\Phi^T}(x)\mathbf{m}$ and $k(x, x') = \mathbf{\Phi^T}(x)\mathbf{C}\mathbf{\Phi}(x')$.

If a continuous inversion is approximated within a finite basis using the least-squares approach, the result will be biased through the omission of the terms involving $\mathbf{P}$ and $\mathbf{Q}$. This is essentially the root of the problem of 'spectral leakage', discussed by Trampert & Snieder (1996), and we will consider this in more detail in due course.

3.1 Choosing discrete prior covariance matrices

However, we first consider the duality between the (discrete) prior model covariance matrix $\mathbf{C}$, and the (continuous) prior covariance function, $k(x, x')$. A perennial practical question concerns the 'correct' way to choose $\mathbf{C}$ in any discrete inversion: often, our prior knowledge about the problem is not in a form conducive to specifying this covariance matrix. The same problem, couched in different language, arises in deterministic discrete inversion: how does one regularize the inversion? Various schemes have been proposed, including analysis of trade-off curves (e.g. Hansen & O'Leary 1993) and treating regularization as a hierarchical Bayesian problem (Valentine & Sambridge 2018); nevertheless, issues surrounding regularization continue to create difficulties for inversion and the interpretation of results.

One challenge in specifying the prior covariance matrix is its abstract nature: since it exists in model-space, it is difficult to develop any intuition for how different choices will impact results. However, the results of this paper show that it can be easily transformed into physical space, using $k(x, x') = \mathbf{\Phi^T}(x)\mathbf{C}\mathbf{\Phi}(x')$. We suggest that intuition should be much more straightforward in this domain. Loosely, the covariance function describes how information about $f(x)$ at one point constrains our knowledge of surrounding points: thus, the suitability of a given covariance function can be assessed based on knowledge of the physical properties of the system of interest, and the length-scales over which correlations are to be expected. The problem can also be approached from the opposite direction: if a particular spatial covariance function is desired, it is straightforward to use eq. (10) to compute the corresponding model covariance matrix. Of course, it will not usually be possible to exactly reproduce the desired covariance within the finite basis—but this may be an acceptable price to pay for the computational efficiency of eq. (14).

This highlights an important distinction: we will refer to a *desired* covariance function, from which a covariance matrix is constructed following eq. (10b), and an *implied* covariance function, constructed from the covariance matrix using $k(x, x') = \mathbf{\Phi^T}(x)\mathbf{C}\mathbf{\Phi}(x')$. In general, the implied covariance function will not be identical to the desired covariance function—and as we shall see, this is the root of many of the 'challenges' associated with discrete inversion.

3.1.1 Tikhonov regularization

In particular, it is instructive to consider the case where the desired covariance function is given by $k(x, x') = \sigma_1^2\delta(x - x')$. A delta-function covariance implies that we do not expect any spatial correlations within the recovered model, and it can be arbitrarily rough—effectively, our prior is for a white-noise style function. The equivalent covariance matrix is given by

$$C_{ij} = \sigma_1^2 \iint \phi_i(x)\delta(x - x')\phi_j(x')\,\mathrm{d}x\,\mathrm{d}x' = \sigma_1^2\Gamma_{ij}. \tag{15}$$

In any orthonormal basis, we will obtain $\mathbf{C} = \sigma_1^2\mathbf{I}$—a choice often known as Tikhonov regularization, or 'ridge regression'. Thus, when we adopt Tikhonov regularization, we are effectively attempting to impose delta-function spatial covariances upon our model.

Two issues may be identified. First, one may reasonably question whether a white-noise prior is a reasonable description for the systems we typically seek to image: few physical processes display such extreme disorder. To give a concrete example, consider a seismic tomography experiment, where we wish to estimate the shear-wave speed at every point within the Earth's interior. If, through some artifice, we came to know the precise value of this quantity at one particular location, how would this affect our state of knowledge elsewhere? It is implausible to suggest that this information would not influence our beliefs elsewhere: in the absence of strong evidence to the contrary, we would undoubtedly assume that this value was also indicative of wave speed in the surrounding region of space. This implies that we expect the model to exhibit correlations over some length-scale, although there may not be consensus over exactly how this should be quantified. While it may appear superficially attractive to address this by imposing no correlation, it turns out—as we shall see—that this creates more problems than it solves.

This arises from the second issue: it is readily-apparent that even if a delta-function covariance is desired, it is impossible to represent its infinite sharpness accurately using a finite set of basis functions. To assess the effects of this, we can compute the implied covariance function, by forming $\mathbf{\Phi}^{\mathbf{T}}(x)\mathbf{\Phi}(x')$. This is illustrated in Figs 1(a)–c) for a basis comprised of normalized Legendre polynomials, complete to degree 50. As might be anticipated, we see a Gibbs phenomenon-like 'ringing' effect, with weak correlations and anti-correlations being imposed throughout the domain. We also see that the value of $k(x, x)$—which corresponds to the width of our prior distribution at the point x—is not constant, and that the pattern of imposed correlations also varies with x. None of these features are a manifestation of true prior information: all emerge solely because we are seeking to impose correlations that are incompatible with our basis.

3.1.2 *Matérn regularization*

As an alternative, Figs 1(d)–(f) shows similar plots for the case where the desired covariance takes the form of a Matérn-$\frac{3}{2}$ function, as introduced in Part I,

$$k(x, x') = \sigma_1^2 \left(1 + \frac{\sqrt{3}|x - x'|}{\sigma_2}\right) \exp\left(-\frac{\sqrt{3}|x - x'|}{\sigma_2}\right). \qquad (16)$$

We choose $\sigma_1 = 1$, and specify a characteristic length-scale of correlation $\sigma_2 = 0.1$. Expanded in the discrete basis of Legendre polynomials, this results in a covariance matrix with off-diagonal structure (Fig. 1d), with the implied covariance function almost exactly reproducing that which is desired (Figs 1e and f).

This example is intended to illustrate two points. First, we see that the correlation structure actually imposed upon a discrete inversion may differ significantly from that which we set out to impose: one must consider whether the desired correlation function can be accurately represented using the chosen set of basis functions. Notwithstanding this, we also see that it is possible to impose a localized correlation structure even in cases where the model is constructed using basis functions that have global support. This is important: for many geoscience applications, global basis functions such as spherical harmonics are mathematically and computationally convenient. However, it is usually the case that our observational constraints are of a fundamentally local character: learning that seismic wave speed is (say) higher than average beneath Australia should not alter our knowledge about seismic wave speed beneath Europe. A common difficulty encountered when performing inversion using global basis functions is that spurious model features get introduced into regions with poor data coverage, as they allow improved data fit elsewhere. Our results indicate that this effect can be suppressed by an appropriate choice of correlation function. Of course, the choice must also be guided by the detailed requirements of a particular application, and we do not claim that the Matérn family is necessarily superior to any other possible correlation function. The key issue is one of representability: can the chosen covariance function be expressed accurately within the model basis? The fact that the Matérn functions are generally smooth, and have finite support, are helpful here—but many other classes of function can be found that share these properties.

Inevitably, every method comes with drawbacks. The major barrier to building covariance matrices from correlation functions is that, in general, the elements of $\mathbf{C}$ must be evaluated by numerical integration. This may amount to a significant computational cost. Moreover, if we wish to exploit the model-space formulation of least-squares, we must invert the resulting covariance matrix—again, an expensive task in high-dimensional settings. However, we suspect that there may be routes around these challenges: given the relatively limited range of basis functions commonly used in practice, it may prove feasible to develop approximations or heuristics that allow inverse covariance matrices to be constructed directly for a given style of desired covariance function. For example, one might envisage a computational library that constructs the Matérn-derived $\mathbf{C}_\mathbf{m}^{-1}$ for an arbitrary length-scale σ_2, by interpolation from a small set of pre-computed inverse matrices—much as numerical integration libraries typically use pre-computed sets of quadrature weights. Clearly, this is an area where further investigation is necessary.

If we are to impose a finite length-scale of correlation upon our models, how should this be chosen? One route is to treat this as a regularization hyperparameter within the framework set out in Valentine & Sambridge (2018), and seek the scale-length that is most compatible with the observational data. This procedure can be framed probabilistically, making it possible to integrate over all possible correlation lengths. However, without an efficient method for obtaining the inverse covariance matrix directly, this is likely to be prohibitively expensive for all but the simplest problems. Nevertheless, it may often be possible to choose a reasonable value based on physical grounds. For example, in a tomography experiment, it may be reasonable to impose a length-scale based on the frequency content of the seismic data set: it is evident that one cannot usefully constrain structural features that are small compared to the wavelength of seismic waveforms used. It is also worth noting that choices regarding length-scale may be usefully related to those regarding the dimension of the basis—for any given basis set, there will be a length-scale below which implied and desired covariance functions begin to diverge. We suggest this may be a reasonable choice to impose in the absence of any alternative rationale. Conversely, if a certain length-scale is chosen *a priori*, this can be used to determine the dimension of basis to use: for example, the vanishing diagonal elements seen in Fig. 1(d) indicate that the highest-degree Legendre polynomials are probably unnecessary in the inversions presented. We highlight that any ill-conditioning in the covariance matrix (preventing its inversion) can be readily addressed by identifying and discarding 'unnecessary' components within the basis. In principle, more complex correlation functions characterized by

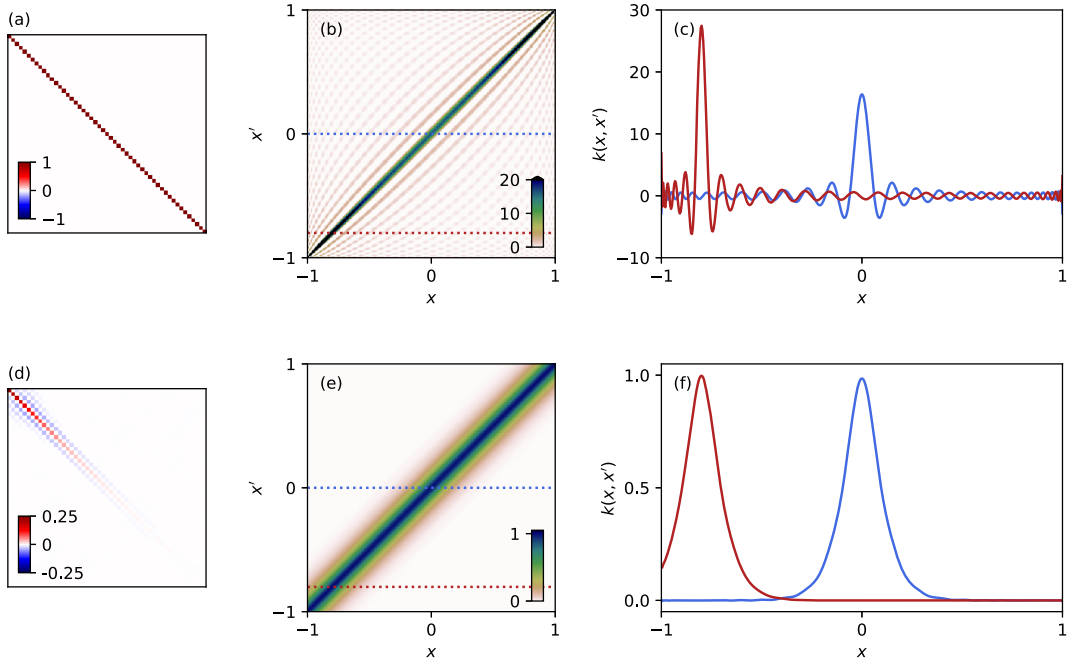

Figure 1. Simple covariance matrices imply complex covariance functions. Choosing the prior covariance matrix, $\mathbf{C}$, to be an identity matrix (a), and adopting a finite basis comprised of normalized Legendre polynomials complete to degree 50, is equivalent to the spatial covariance function $k(x, x')$ shown in (b). This has a complex structure; two slices through this function (dotted lines) are illustrated in (c). Notable (and undesirable) features include a Gibbs phenomenon-like ringing of variable wavelength, and variation in peak amplitude. As an alternative, (d-f) show corresponding plots for a covariance matrix constructed based on a Matérn-$\frac{3}{2}$ function (d). Using this within an inversion imposes a simple spatial covariance function (e–f) upon the recovered model, without any of the undesirable features.

multiple or even spatially-variable length-scales may be used, but these lie beyond the scope of this paper.

3.1.3 A simple example

To illustrate the practical benefits associated with adopting non-delta-like covariance functions, we perform a simple inversion experiment. To begin with, we draw samples from the function

$$f(x) = \begin{cases} 0 & -1 \leq x < \frac{1}{2} \\ \cos(x) & \frac{1}{2} \leq x < 0 \\ 1 + x - x^2 & 0 \leq x \leq 1 \end{cases} \tag{17}$$

at 20 randomly distributed points x_i in the ranges $-0.7 \leq x \leq 0.1$ and $0.5 \leq x \leq 1$ (see Fig. 2a). To each sample, we add Gaussian random noise, using a zero-mean distribution with standard deviation $\sigma = 0.1$. We then attempt to recover the underlying function by performing a least-squares inversion (following Tarantola & Valette 1982) of this data set, using a basis comprised (again) of normalized Legendre polynomials to degree-50. First, we regularize this using Tikhonov regularization, $\mathbf{C} = \sigma^2\mathbf{I}$, choosing σ using the approach set out in Valentine & Sambridge (2018). Results are shown in Fig. 2(b), and display reasonable performance in areas where there is a high density of samples, but also considerable oscillation in areas where there is little data: it is readily apparent that the results are contaminated by the 'ringing' effects identified in Figs 1(b) and (c). We then perform a second inversion, identical to the first except for the regularization: we add a term to the inverse prior covariance matrix to penalize steep gradients in the recovered model, with relative weight determined based on Valentine & Sambridge (2018). Results for this are shown in Fig. 2(c), and while they remain imperfect, especially—unsurprisingly—in poorly sampled areas, the recovered function contains much less spurious structure, and overall performance appears superior. Broadly similar results are seen in a third inversion, which is regularized using the Matérn-derived regularization operator shown in Figs 1(d)–(f).

In these three experiments, we are effectively performing regression: our data amounts to direct observations of the target function. This is not the usual scenario in geophysical inverse problems, where indirect measurements are more common. To better-represent this, we therefore construct a second data set, consisting of the average values of $f(x)$ between each pair that can be constructed from the sampling points, x_i. In other words, we suppose we have 190 'measurements',

$$d_k = \frac{1}{x_j^{(k)} - x_i^{(k)}} \int_{x_i^{(k)}}^{x_j^{(k)}} f(x)\,\mathrm{d}x\,, \tag{18}$$

where we assume $x_i < x_j$. In the notation of eq. (1), this corresponds to a weighting function

$$w_k(x) = \frac{1}{x_j^{(k)} - x_i^{(k)}} H\left(x - x_i^{(k)}\right) H\left(x_j^{(k)} - x\right), \tag{19}$$

where $H(x)$ is a Heaviside step function; to illustrate the density of information about $f(x)$ provided by this data set, Fig. 2(e) shows $\sum_k w_k(x)$. Again, we add Gaussian random noise to this data, and perform one inversion using each style of regularization (with the weight of the regularization term determined following Valentine & Sambridge (2018)). Results may be seen in Figs 2(f)–(h), and again—at least, from a visual perspective—the gradient-based and

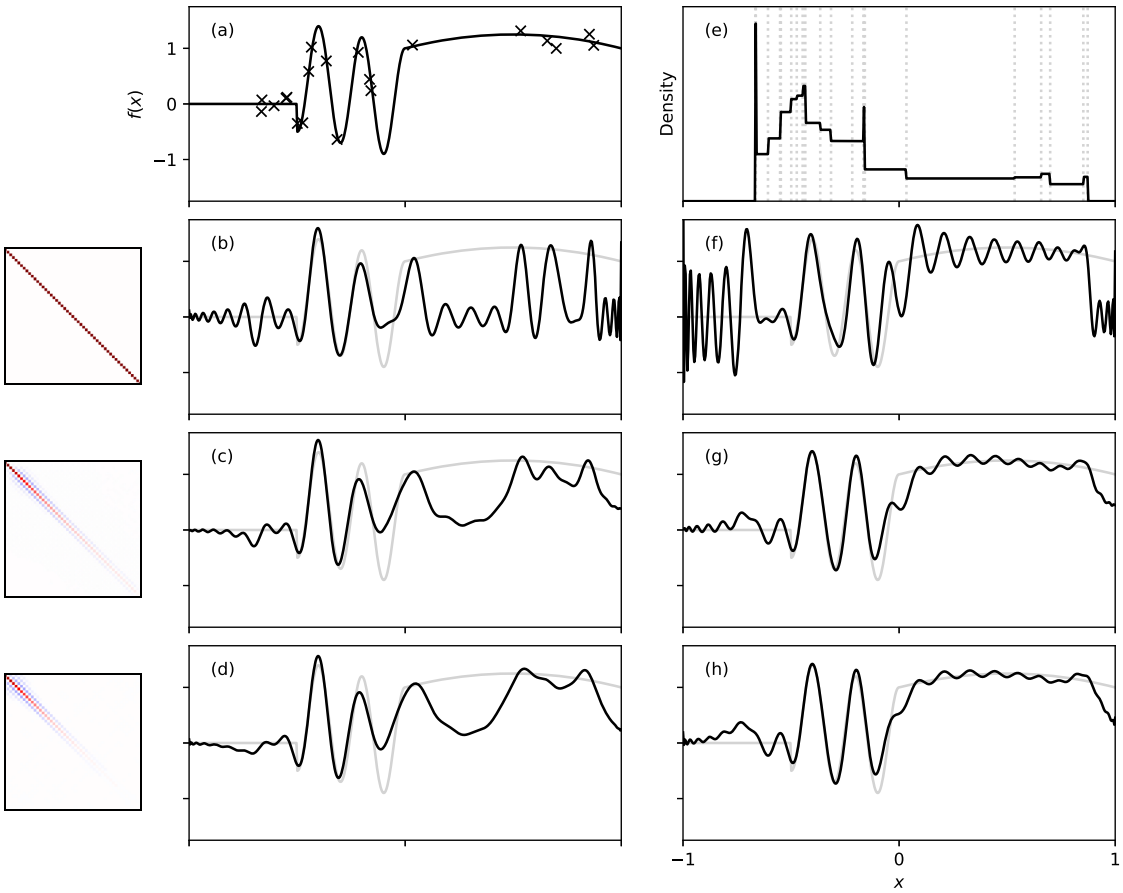

Figure 2. Localized structure with global basis functions. We generate a set of noisy samples from a known function (a), and fit a model expressed in Legendre polynomials complete to degree-51 to these. When Tikhonov regularization is used (b), the model performs poorly in regions of limited data coverage; compare Figs 1(a)–(c). In (c) we add an additional constraint to keep the first derivative of the recovered function small, while in (d) we use a Matérn-derived covariance as in Figs 1(d)–(f). As a second experiment, we compute the average value of $f(x)$ between every pair of sample-points from (a); the resulting density of information about f is shown in (e). Adding noise to this data set of indirect measurements, we again perform inversion using Tikhonov (f), Tikhonov with first derivative damping (g) and the Matérn-derived (h) regularization matrices. Tikhonov regularization clearly introduces many artefacts, which can be readily related to the implied covariance function depicted in Fig. 1(b), whereas the Matérn-derived regularization gives an excellent reproduction of the original function. Requiring the first derivative to be small yields results that are broadly similar to the Matérn case, but with slightly stronger artefacts: see Fig. 3 for the equivalent correlation functions.

Matérn-derived approaches give similar results, vastly outperforming Tikhonov regularization, with a much-reduced incidence of artefacts in the recovered model.

To better-understand these similarities, Fig. 3 shows the spatial covariance function associated with the gradient-based regularization operator. We see that this has a general form that is close to that of the Matérn covariance function, with good localization of information; however, there remains some residual 'ringing' of the kind encountered in the Tikhonov case. By requiring that the gradient of the recovered function be 'small', we are implicitly stipulating that the function cannot change too dramatically from one location to the next—in other words, that the function values must display correlations over at least some length scale. The notion that an inversion should result in a 'smooth' solution can therefore be approached from either perspective. Whether one viewpoint is preferable to the other will be application-dependent, but we observe that in many cases there is little objective justification for a decision to penalize certain orders of derivative and not others. In the typical case where prior knowledge is relatively loosely defined, we suggest that

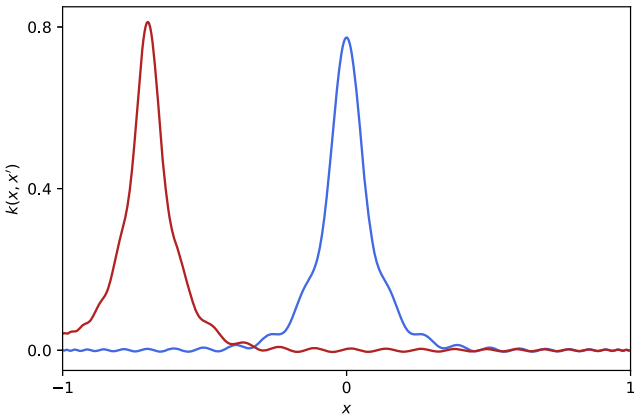

Figure 3. Spatial correlation functions corresponding to a Tikhonov regularization operator incorporating first-derivative constraints, as used in Figs 2(c) and (g).

introducing regularization by reference to a desired spatial correlation function and correlation length-scale may be an attractive philosophy. It may also ease the process of ensuring comparability between differently parametrized inversions. Regardless of the viewpoint used to define the regularization operator, we believe that knowledge of the implied spatial correlation function is of great assistance in understanding how regularization choices are likely to be impacting results.

3.2 Spectral Leakage

Trampert & Snieder (1996) consider the concept of spectral leakage in discrete inverse problems, adopting the probabilistic least-squares formulation of Tarantola & Valette (1982) and a Tikhonov-style covariance matrix ($\mathbf{C} = \sigma_1^2\mathbf{I}$). Working with spherical harmonic basis functions, they compute synthetic data for a model containing structure at degree L. They then invert this synthetic data, using a parametrization that only allows structure up to degree $L_m < L$, and find significant errors in the recovered model. This is due to spectral leakage—an effect similar to aliasing, whereby the unrepresentable structure between degrees L_m and L gets mapped into results. Given that real physical systems, such as the Earth, will almost inevitably contain features at a finer scale than any basis used during inversion, this is problematic. Another manifestation of the same fundamental problem occurs if one compares two inversions, one obtained using a basis complete to L_1, and a second using basis functions complete to $L_2 > L_1$: while one might naïvely expect the model coefficients to agree below L_1, this is often not the case.

From the perspective of this paper, these issues all stem from the fact that a delta-like desired covariance function is unrepresentable in the finite basis. As we have seen, the choice $\mathbf{C} = \sigma_1^2\mathbf{I}$ ends up imposing relatively strong correlations upon the recovered model, characterized by the shortest length-scale present in the basis set. If a different basis is used (e.g. by increasing the maximum spherical harmonic degree), the continued choice $\mathbf{C} = \sigma_1^2\mathbf{I}$ will result in an implied covariance function with different properties, and thus the recovered models will also differ.

To illustrate this, Fig. 4 shows two implied spatial covariance functions, both corresponding to the choice $\mathbf{C_m} = \mathbf{I}$, but constructed using different basis sets: one with spherical harmonics complete to degree $L = 8$, and the other with spherical harmonics complete to degree $L = 16$. The difference in spatial form can be readily appreciated; in particular, the dominant wavelength associated with periodic correlations changes as the basis set is expanded. Thus, the manner in which the inversion 'fills in gaps' between data will also change, despite the practitioner not (apparently) having altered their approach to regularization. Coupled with an irregular data distribution, and a non-trivial data-model relationship, it is evident that inversion results obtained in the two cases may differ considerably.

The easiest route to avoiding significant spectral leakage is therefore to ensure that the desired spatial covariance function is representable within the chosen set of basis functions—or, equivalently, ensuring that the implied spatial covariance function does not change dramatically as the dimension of the basis set is increased. For most 'nice' functions, this is likely to simply be a matter of ensuring that the minimum length-scales within covariance and basis are consistent. Any part of the desired covariance function that lies outside the basis will manifest itself through non-zero matrices $\mathbf{P}$ and $\mathbf{Q}$, and (at least in principle) these can be used to obtain a quantitative understanding of the resulting spectral leakage.

We note that Trampert & Snieder (1996) derive a 'spectral leakage correction', designed to ameliorate the problem. However, this requires inversion in the data-space—and thus, comes at similar computational cost to that of the GP framework set out in Part I. In cases where spectral leakage is a concern, it is therefore also feasible to use our continuous approach. Where expression relative to a basis is nevertheless desirable (e.g. to estimate power spectra), this can be achieved through the use of eq. (9). In doing so, one can ensure that no discretization error has been introduced.

4 CONCLUDING REMARKS

According to the 'folk wisdom' of inversion using the least-squares algorithm, use of global basis functions (such as spherical harmonics) risks models containing unconstrained artefacts in regions of poor data coverage, while use of local basis functions (e.g. grid cells) inevitably results in very 'patchy' models. One solution to this problem is to adopt alternative inversion schemes, such as the Backus–Gilbert approach adopted by Zaroli (2016); another strategy involves the imposition of strong regularization and smoothing constraints. The results of this paper indicate that the key requirement is to ensure that the implied correlation function associated with the regularization is appropriately localized, so that information obtained in one region of the model does not unduly influence results elsewhere.

Four points deserve to be highlighted. First, from a theoretical perspective, the results set out in Section 2 provide an attractive route to deriving the results of Tarantola & Valette (1982). While discrete basis functions tend to simplify computations, they often complicate analysis, and the approximation implicit in discretization can often be a source of difficulty. The connections developed in this paper enable analysis in the continuous domain with discretization as a final step. We suggest this may greatly simplify theoretical studies in linearized inversion.

Secondly, any regularization scheme adopted in a least-squares inversion can straightforwardly and cheaply be represented as an implied covariance function. We suggest that plotting this function, or transects through it, should become routine whenever results are to be analysed or presented. It is far more readily understood than a model covariance matrix, and one may straightforwardly appreciate where regularization choices might be impacting results.

Thirdly, we encourage the community to move away from Tikhonov regularization, particularly in conjunction with global basis functions, and to instead explore the potential to construct covariance matrices from a chosen desired covariance function. We believe this enables a more principled approach to the incorporation of prior information, and we have demonstrated clear practical benefits associated with the imposition of local correlations. For this to be possible, certain practical hurdles must be overcome: in particular, computing large-scale covariance matrices by numerical integration of a covariance function is computationally expensive. Whether or not this is a barrier to adoption is obviously situation-dependent; however, it may prove possible to develop algorithms that reduce such costs.

The final message we wish to convey is more general. The inspiration for this work came from the machine learning literature: how could we understand and apply the theory of GPs in a geophysical context? Eventually, this developed into the probabilistic, continuous inverse theory set out in Part I, and the links to Tarantola & Valette (1982) emerged. The resulting insights offer an opportunity to transform our approach to regularization in least-squares

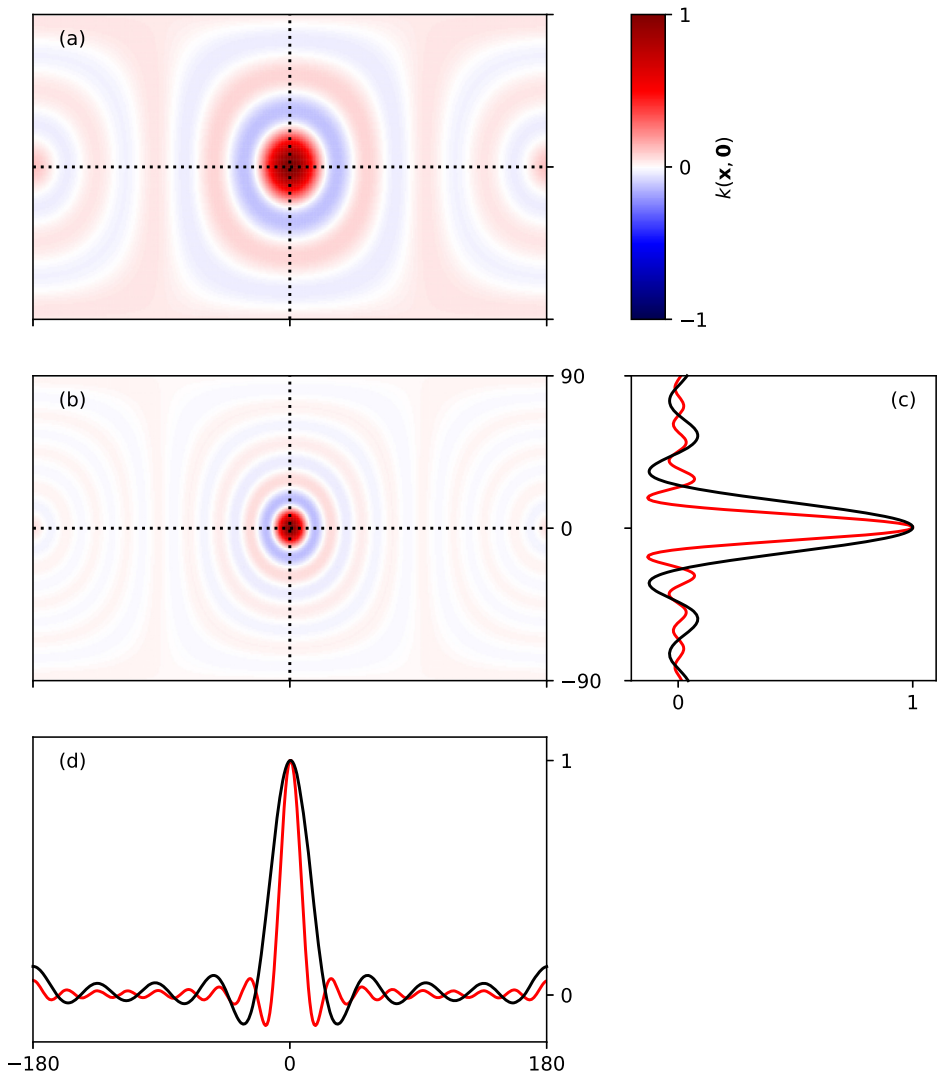

Figure 4. Spectral leakage: a covariance-function perspective. Spatial covariance functions $k(\mathbf{x}, \mathbf{0})$ corresponding to $\mathbf{C} = \mathbf{I}$, expanded within a spherical-harmonic basis complete to (a) $L = 8$ and (b) $L = 16$. In (c) and (d), we show cross-sections through both 2-D functions. The difference in dominant spatial wavelength can be clearly appreciated: in changing the basis set, we have implicitly changed our assumptions about the properties of the underlying function.

inversion—a technique that has been central to geophysical investigations for over half a century, and about which an immense body of literature has been assembled. We do not believe this is a unique case: there is much to be gained by attempting to join the dots that connect geophysical inverse theory to the mathematical and statistical results that underpin the broad field of data science.

ACKNOWLEDGEMENTS

We gratefully acknowledge the thoughtful and insightful comments received from Frederik Simons, Jeannot Trampert and an anonymous reviewer. Discussions with David Al-Attar and Paul Käufl also influenced this work. APV is supported by the Australian Research Council through a Discovery Early Career Research Award (DE180100040).

REFERENCES

Boschi, L. & Dziewoński, A., 1999. High- and low-resolution images of the Earth's mantle: implications of different approaches to tomographic modeling, *J. geophys. Res.,* **104,** 25 567–25 594.

Davies, R., Valentine, A., Kramer, S., Rawlinson, N., Hoggard, M., Eakin, C. & Wilson, C., 2019. Earth's multi-scale topographic response to global mantle flow, *Nat. Geosci.,* **12,** 845–850.

Hansen, P. & O'Leary, D., 1993. The use of the L-curve in the regularization of discrete ill-posed problems, *SIAM J. Scient. Comput.,* **14,** 1487–1503.

Hoggard, M., White, N. & Al-Attar, D., 2016. Global dynamic topography observations reveal limited influence of large-scale mantle flow, *Nat. Geosci.,* **9,** 456–463.

Kaula, W., 1967. Theory of statistical analysis of data distributed over a sphere, *Rev. Geophys.,* **5,** 83–107.

Reigber, C., Schmidt, R., Flechtner, F., König, R., Meyer, U., Neumayer, K.-H., Schwintzer, P. & Zhu, S., 2005. An Earth gravity field model compete to degree and order 150 from GRACE: EIGEN-GRACE02S, *J. Geodyn.,* **39,** 1–10.

Schachtschneider, R., Holschneider, M. & Mandea, M., 2010. Error distribution in regional inversion of potential field data, *Geophys. J. Int.,* **181,** 1428–1440.

Simons, F., van der Hilst, R., Montagner, J.-P. & Zielhuis, A., 2002. Multimode Rayleigh wave inversion for heterogeneity and azimuthal anisotropy of the Australian upper mantle, *Geophys. J. Int.,* **151,** 738–754.

Slobbe, D., Simons, F. & Klees, R., 2012. The spherical Slepian basis as a means to obtain spectral consistency between mean sea level and the geoid, *J. Geod.,* **86,** 609–628.

Tarantola, A. & Valette, B., 1982. Generalized nonlinear inverse problems solved using the least squares criterion, *Rev. Geophys. Space Phys.,* **20,** 219–232.

Trampert, J. & Snieder, R., 1996. Model estimations biased by truncated expansions: possible artifacts in seismic tomography, *Science,* **271,** 1257–1260.

Valentine, A. & Sambridge, M., 2018. Optimal regularization for a class of linear inverse problem, *Geophys. J. Int.,* **215,** 1003–1021.

Valentine, A. & Trampert, J., 2016. The impact of approximations and arbitrary choices on geophysical images, *Geophys. J. Int.,* **204,** 59–73.

Valentine, A.P. & Sambridge, M., 2019. Gaussian process models—I. A framework for probabilistic continuous inverse theory, *Geophys. J. Int.,* in press, doi: 10.1093/gji/ggz520.

Whaler, K. & Gubbins, D., 1981. Spherical harmonic analysis of the geomagnetic field: an example of a linear inverse problem, *Geophys. J. Int.,* **65,** 645–693.

Woodhouse, J. & Dziewonski, A., 1984. Mapping the upper mantle: three-dimensional modelling of Earth structure by inversion of seismic waveforms., *J. geophys. Res.,* **89,** 5953–5986.

Zaroli, C., 2016. Global seismic tomography using Backus-Gilbert inversion, *Geophys. J. Int.,* **207,** 876–888.