Geophysical Journal International

Geophys. J. Int. (2020) **220,** 1632–1647
Advance Access publication 2019 November 19
GJI General Geophysical Methods

doi: 10.1093/gji/ggz520

Gaussian process models—I. A framework for probabilistic continuous inverse theory

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 November 7; in original form 2019 May 30

SUMMARY

We develop a theoretical framework for framing and solving probabilistic linear(ized) inverse problems in function spaces. This is built on the statistical theory of Gaussian Processes, and allows results to be obtained independent of any basis, avoiding any difficulties associated with the fidelity of representation that can be achieved. We show that the results of Backus–Gilbert theory can be fully understood within our framework, although there is not an exact equivalence due to fundamental differences of philosophy between the two approaches. Nevertheless, our work can be seen to unify several strands of linear inverse theory, and connects it to a large body of work in machine learning. We illustrate the application of our theory using a simple example, involving determination of Earth's radial density structure.

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

1 INTRODUCTION

In the traditional taxonomy of inverse theory, strategies for solving inverse problems are generally divided into classes: deterministic or probabilistic; continuous or discrete. Such distinctions are both practical and philosophical—they denote our fundamental assumptions about the characteristics of the solution we wish to obtain. Deterministic approaches seek to identify a single entity that best explains observations; probabilistic approaches indicate ranges (volumes of model space) within which plausible solutions may lie. Continuous inverse theory is built on the mathematics of functional analysis, treating the solution as an arbitrary function; discrete approaches introduce a finite set of basis functions, and thus cast inference within the domain of linear algebra.

Discrete, deterministic approaches have a long and distinguished history in geophysics. In particular, studies based on least-squares inversion have underpinned much seminal work in the field (Wiggins 1972; Dziewonski *et al.* 1981; Woodhouse & Dziewonski 1984), and is discussed in numerous monographs (e.g. Menke 1989; Aster *et al.* 2013). The major advantage of these methods is their computational efficiency, with costs governed by the number of discrete basis functions used. This allows researchers to limit the scale of their inference task to suit available resources, but imposes strong assumptions about the properties of the model sought: we assert that it can be well-represented using the chosen set of basis functions. In many cases, there is no convincing *a priori* justification for this, and it has been demonstrated that an inadequate choice of basis can lead to a form of aliasing, where spurious model features are introduced to compensate for those that are unrepresentable (Trampert & Snieder 1996).

Continuous, deterministic inversion strategies—for which the work of Backus & Gilbert (1967, 1968, 1970) serves as the archetype—aim to avoid such assertions. Rather than attempting to construct a comprehensive 'image' of the system of interest, the continuous approach focuses on estimating targeted properties, such as the average value of some physical quantity within a given region. The great strength of Backus & Gilbert's approach is the emphasis it places upon understanding what can, and cannot, be resolved: a user must explicitly consider the sensitivities of individual observations, and how these can best be combined to access any quantity sought. However, this can be time-consuming and computationally expensive, which has often limited the practical application of the theory—although recent work by Zaroli (2016, 2019), building on that of Pijpers & Thompson (1992), offers a route to improved tractability. In addition, because the Backus–Gilbert philosophy focuses on recovering average properties, rather than a conventional 'model', results may not be suitable for use as inputs into further analysis: for example, it is typically not meaningful to calculate synthetic data based on the results of a Backus–Gilbert inversion.

An obvious drawback of any deterministic approach—discrete or continuous—is the presumption that there is a single 'answer' that can explain observations. In many geophysical settings, it is apparent that this cannot be true: available data plainly lack the sensitivity required to properly constrain all components within the basis function expansion. This motivates strategies that seek to identify the full range of models that might be compatible with observations. This task can be approached from a variety of directions, and does not inherently require a

probabilistic framework (see, e.g. Jackson 1976; Meju & Hutton 1992). However, its implementation in any large-scale, problem-independent setting is vastly simplified by using the mathematical language of probability calculus.

Probabilistic approaches also have a venerable history in geophysics, with Harold Jeffreys playing a central role in the development of Bayesian statistics (Jeffreys 1931, 1939). Monte Carlo methods for quantifying the range of models that are compatible with data have been used for 50 yr (e.g Press 1968; Anderssen *et al.* 1972; Worthington *et al.* 1972), although their scope and utility has grown immensely since then, mirroring the explosion in computational power, and a variety of strategies and techniques have been explored and developed (see, e.g. Sambridge & Mosegaard 2002). The work of Tarantola (2005) has also been hugely influential in this sphere. In particular, Tarantola & Valette (1982) provided geophysicists with a probabilistic framing of the familiar least-squares algorithm, enabling first-order uncertainty quantification at low cost.

However, most of the mainstream literature on probabilistic methods in geophysics is fundamentally discrete in character: inversion or sampling are carried out relative to a particular set of basis functions. There are hints—especially in Tarantola & Nercessian (1984) and Tarantola (2005)—at how continuous inverse theory might be cast in a probabilistic guise, but we are not aware of these ever having been developed more fully. This paper therefore attempts to lay out a framework for inference building on the statistical theory of stochastic processes: probability distributions defined in function spaces. More specifically—because, like Tarantola & Valette (1982), we wish to exploit the analytic tractability of the normal distribution—we focus on inference using Gaussian Processes (GPs) as the underlying model (e.g. Rasmussen & Williams 2006).

Our exploration of models built around GPs is also motivated by their central importance in the field of machine learning. It turns out that many of the tools developed and applied in this area—such as neural networks or support vector machines—have fundamental theoretical connections to GPs. Given the significant, and growing, interest in exploiting machine learning techniques to address geoscientific questions (e.g. Valentine & Kalnins 2016; Bergen *et al.* 2019), we believe it is valuable to explore how such methods fit into the context of 'traditional' geophysical inverse theory. We hope this will assist with the development and interpretation of new approaches, and—perhaps—help researchers focus their attention in fruitful directions.

This paper is Part I of a two-part series. It is predominantly theoretical in character, providing a brief introduction to the theory of GPs, and showing how this can be used to frame inversion in a probabilistic, continuous setting. We begin by sketching the 'standard' application of GPs, for interpolation between direct observations of a function; we then extend this to the case encountered in inverse problems, where only derived properties of the function can be observed, and discuss concepts such as resolution and information gain. To illustrate the application of the theory, we provide an example where it is deployed to allow inference of Earth's radial density structure. In Part II (Valentine & Sambridge 2019), we show that the GP approach reduces to that of Tarantola & Valette (1982) once functions are approximated within a finite-dimensional basis, and discuss some implications this brings for discretized inversion approaches.

2 THE GAUSSIAN PROCESS

We briefly sketch the theory of Gaussian processes here; for a full account, the reader is referred to standard texts such as Rasmussen & Williams (2006) or Murphy (2012). Readers familiar with the field of geostatistics may wish to note that the technique of kriging (Krige 1951; Dubrule 2018) is essentially a special case of the GP theory. A GP is a form of stochastic process, and can be regarded as a way of defining a Gaussian distribution over functions. If our knowledge of some function $f(x)$ is to be described by a GP, we denote this by writing

$$f(x) \sim \mathcal{GP}\left(\mu(x), k(x, x')\right) . \tag{1}$$

To simplify the analysis and notation within this paper, we will generally treat $f(x)$ as if it is a scalar function of a single scalar variable, but all our results extend straightforwardly to functions with higher-dimensional domains and/or ranges (in which case, the underlying model might also be referred to as a 'Gaussian random field'). As is prefigured by our notation, a GP is fully specified by two quantities: a mean function $\mu(x)$, and a symmetric two-point covariance function, $k(x, x') = k(x', x)$. Then, the defining property of a GP is that for *any* collection of sample points $\{x_1, x_2, \ldots, x_N\}$, our knowledge of f at these points is described by an N-dimensional multivariate normal distribution, such that

$$\begin{pmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ f(x_N) \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu(x_1) \\ \mu(x_2) \\ \vdots \\ \mu(x_N) \end{pmatrix}, \begin{pmatrix} k(x_1, x_1) & k(x_1, x_2) & \cdots & k(x_1, x_N) \\ k(x_2, x_1) & k(x_2, x_2) & \cdots & k(x_2, x_N) \\ \vdots & \vdots & \ddots & \vdots \\ k(x_N, x_1) & k(x_N, x_2) & \cdots & k(x_N, x_N) \end{pmatrix} \right) , \tag{2}$$

where $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the usual multivariate normal distribution. In other words, the mean and covariance functions provide a recipe for constructing an appropriate mean vector and covariance matrix to characterize our knowledge of the function, irrespective of the particular sampling points chosen. For any specific realization of these quantities, it is straightforward to generate samples from this normal distribution using standard computational tools. (Throughout this paper, $\mathcal{GP}(\cdot, \cdot)$ is used to denote a distribution defined in function-space, whereas $\mathcal{N}(\cdot, \cdot)$ indicates a normal distribution defined in the conventional manner, within a finite-dimensional vector space.)

2.1 Gaussian processes for function prediction

Since eq. (2) holds for any collection of points, we can contemplate a situation where we have directly observed the value of $f(x)$ at certain points, $\{\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_N\}$, and obtained measurements $\{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_N\}$; we will use the vectors $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$ to collectively denote these quantities. We assume that our measurement of $\hat{\mathbf{y}}$ is subject to observational noise, characterized by a zero-mean Gaussian with covariance matrix $\mathbf{C_d}$. Our knowledge about the function at some additional, general point x is then given by

$$\begin{pmatrix} f(x) \\ \hat{\mathbf{y}} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu(x) \\ \hat{\boldsymbol{\mu}} \end{pmatrix}, \begin{pmatrix} k(x,x) & \hat{\mathbf{k}}^{\mathbf{T}}(x) \\ \hat{\mathbf{k}}(x) & \hat{\mathbf{K}} + \mathbf{C_d} \end{pmatrix} \right) \tag{3}$$

where we have introduced vector and matrix quantities such that

$$[\hat{\boldsymbol{\mu}}]_i = \mu(\hat{x}_i) \tag{4a}$$

$$\left[\hat{\mathbf{k}}(x)\right]_i = k(\hat{x}_i, x) \tag{4b}$$

$$\left[\hat{\mathbf{K}}\right]_{ij} = k(\hat{x}_i, \hat{x}_j) \tag{4c}$$

Now, we can apply a standard result (see, e.g. Petersen & Pedersen 2015) for the conditioning of a multidimensional Gaussian to obtain

$$f(x) \,|\, \hat{\mathbf{y}} \sim \mathcal{N}\left(\mu(x) + \hat{\mathbf{k}}^{\mathbf{T}}(x) \left(\hat{\mathbf{K}} + \mathbf{C_d}\right)^{-1} (\hat{\mathbf{y}} - \hat{\boldsymbol{\mu}}) \,, k(x,x) - \hat{\mathbf{k}}^{\mathbf{T}}(x) \left(\hat{\mathbf{K}} + \mathbf{C_d}\right)^{-1} \hat{\mathbf{k}}(x) \right) . \tag{5a}$$

Thus, given some general knowledge about the properties of f, and some specific measurements of its values at a discrete set of points, we are able to predict its value at a general point. For completeness, we note that if predictions are simultaneously sought for multiple points $\{\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_N\}$, we have

$$\begin{pmatrix} f(\tilde{x}_1) \\ f(\tilde{x}_2) \\ \vdots \\ f(\tilde{x}_N) \end{pmatrix} \,|\, \hat{\mathbf{y}} = \mathcal{N}\left(\tilde{\boldsymbol{\mu}} + \boldsymbol{\kappa}^{\mathbf{T}} \left(\hat{\mathbf{K}} + \mathbf{C_d}\right)^{-1} (\hat{\mathbf{y}} - \hat{\boldsymbol{\mu}}) \,, \tilde{\mathbf{K}} - \boldsymbol{\kappa}^{\mathbf{T}} \left(\hat{\mathbf{K}} + \mathbf{C_d}\right)^{-1} \boldsymbol{\kappa} \right) , \tag{5b}$$

where $\tilde{\boldsymbol{\mu}}$ and $\tilde{\mathbf{K}}$ are defined as in eq. (4), except that evaluation points are instead drawn from $\tilde{\mathbf{x}}$; $\boldsymbol{\kappa}$ has elements given by

$$[\boldsymbol{\kappa}]_{ij} = k(\hat{x}_i, \tilde{x}_j) \,. \tag{6}$$

These two equations allow us to quantify our knowledge about the anticipated behaviour of the function at unseen points. In particular, the first allows us to identify volumes of space within which the function is likely to lie (as in Fig. 1, central column), while the second allows the generation of random realizations of functions that are compatible with both assumptions and observations (Fig. 1, right-hand column). It may be helpful to note that our knowledge of the function continues to be described by a Gaussian process after conditioning; the operation can be seen as an update to the mean and covariance functions in the light of the data.

2.2 Mean and covariance functions

It should be apparent from the foregoing discussion that the mean and covariance functions are central to the properties of any GP. In particular, the covariance function controls how the pointwise information contained within our observations can be extended into the surrounding region, and it therefore plays a key role in determining the characteristics of the interpolated function and its uncertainty. By writing $f(x) \sim \mathcal{GP}(\mu(x), k(x,x'))$, we are asserting that

$$\mathbb{E}\left[f(x)\right] = \mu(x) \tag{7a}$$

$$\begin{aligned} \mathbb{C}\left[f(x), f(x')\right] &= \mathbb{E}\left[\left(f(x) - \mathbb{E}\left[f(x)\right]\right)\left(f(x') - \mathbb{E}\left[f(x')\right]\right)\right] \\ &= k(x,x'), \end{aligned} \tag{7b}$$

where $\mathbb{E}[\,\cdot\,]$ and $\mathbb{C}[\,\cdot\,,\,\cdot\,]$ denote the statistical expectation and covariance, respectively. Of course, in most practical circumstances where we seek to reconstruct an unknown function, we are unlikely to know these quantities precisely. Instead, we must typically assume that the mean and covariance functions take a particular form, based on our knowledge or presumptions about the system of interest.

It is usually straightforward to identify a sensible choice of mean function. In many cases, this will be a constant function (possibly zero), or perhaps a simple relationship such as a linear trend that can be readily estimated from the data, and in practice this choice has relatively minor impact on the overall performance of the GP approach. Much more significant is the covariance function, which is usually assembled from one or more 'building blocks', chosen from a suite of standard options (see, e.g. Abrahamsen 1997; Rasmussen & Williams 2006). In this paper, we consider only 'stationary' covariance functions, which depend only on the distance between the points x and x', and not on their values. We denote this distance by $d(x,x')$, and note that this may take different forms depending on the situation of interest, with $d(x,x') = |x - x'|$ representing a common simple choice. Examples of stationary covariance functions then include the delta-function covariance,

$$k(x,x') = \sigma_1^2 \delta(d(x,x')) \tag{8a}$$

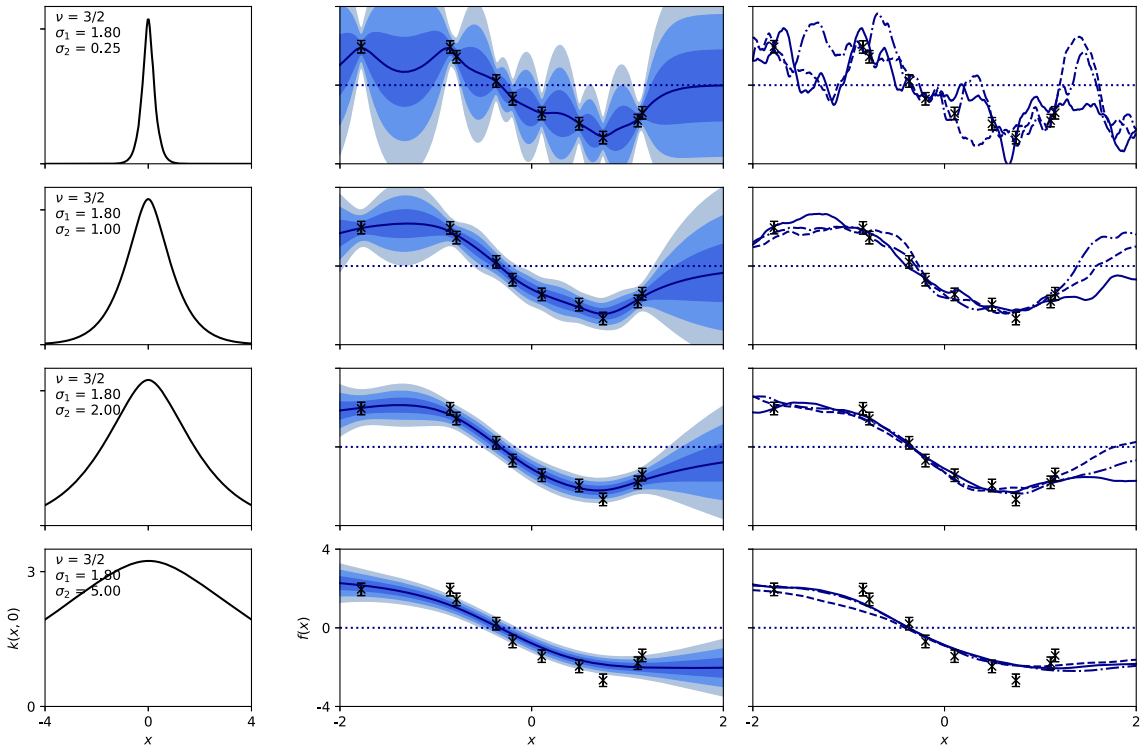

Figure 1. Gaussian process interpolations of a data set with varying length-scale parameter. Each row depicts a GP conditioned on the same data set, using a Matérn-$\frac{3}{2}$ covariance function (a transect through which, $k(x, 0)$, is shown in the left-hand column). The only difference between the four cases is the value of the parameter σ_2 (see values, inset), controlling the characteristic length-scale of the GP. In the middle column, we show (i) the prior mean of the GP (dotted line); (ii) the posterior mean of the GP (solid line); (iii) 1σ, 2σ and 3σ posterior uncertainty bands (shaded regions). Right-hand column shows three sample functions drawn at random from the posterior distribution. Data uncertainties are as depicted by error bars, with all measurements assumed independent.

the squared-exponential covariance (*cf.* Tarantola & Nercessian 1984),

$$k(x, x') = \sigma_1^2 \exp\left[-\frac{d\,(x, x')^2}{2\sigma_2^2}\right] \tag{8b}$$

and the Matérn family of covariance functions, which take the form

$$k(x, x') = \sigma_1^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}\, d(x, x')}{\sigma_2}\right)^{\nu} K_\nu \left(\frac{\sqrt{2\nu}\, d(x, x')}{\sigma_2}\right) \tag{8c}$$

where K_ν represents a modified Bessel function and $\Gamma(\nu)$ is a Gamma function.

It will be noted that in all of the above examples, the covariance function depends on one or more additional 'hyperparameters'. The parameter σ_1 represents an amplitude scaling for the covariance function, and governs width of the prior distribution for $f(x)$: in the absence of any additional data, it is apparent from eq. (1) that $f(x) = \mathcal{N}(\mu(x), \sigma_1^2)$. The parameter σ_2 controls the 'width' of the covariance function, and determines the characteristic length-scale present within the GP. Finally, for the Matérn family, the order parameter ν governs the form of the covariance function. In fact, it turns out that this also controls the differentiability of the GP (and thus of samples from it): we must have $\nu > n$ if the process is to be n-times differentiable. Commonly, half-integer values are chosen for ν, since this allows considerable simplification of the covariance expression; however, this is not a necessity. It can also be shown that the squared-exponential covariance represents the limiting case of the Matérn family as $\nu \to \infty$.

To illustrate the effects of these parameters, Fig. 1 shows interpolation of a single data set using a Matérn-$\frac{3}{2}$ covariance function (i.e. a Matérn covariance with $\nu = \frac{3}{2}$), for four different values of the length-scale parameter σ_2. When this is small, measurements can only constrain the function in the immediate vicinity of the observation point, and uncertainty grows rapidly away from these. Reflecting this, individual samples from the GP display variability on short length-scales. As σ_2 is increased, the samples provide constraint across a wider region, and the short-range function variability disappears.

Similarly, Fig. 2 illustrates the influence of the order parameter, ν, using the same data set, with the other two hyperparameters (σ_1, σ_2) chosen to be 'optimal' in a sense to be discussed below. It is readily apparent that increasing ν leads to a GP that represents 'smoother' functions: for low values of ν, samples from the GP display significant roughness at short length-scales (indeed, at length-scales significantly below that of the imposed correlation structure). It is also evident that the influence of variations in ν diminishes as ν increases: $\nu = \frac{5}{2}$ and $\nu = \infty$ lead to rather similar inferences, whereas $\nu = \frac{1}{2}$ and $\nu = \frac{3}{2}$ are very different.

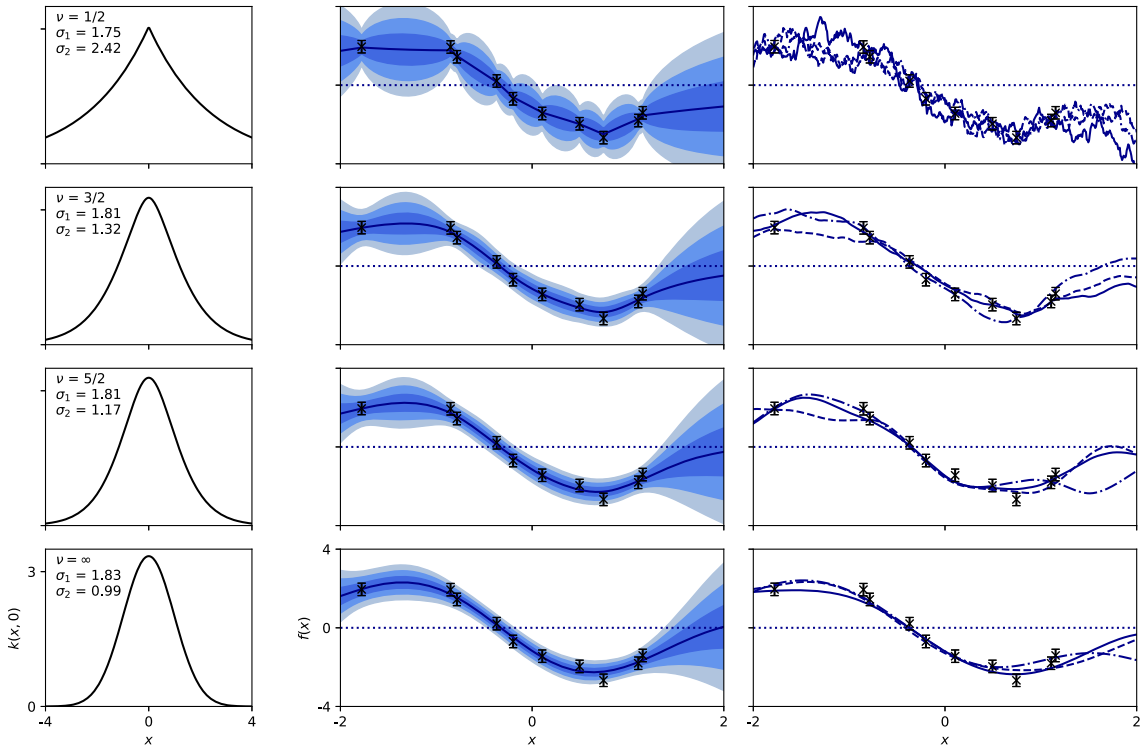

Figure 2. Gaussian process interpolations; varying form of covariance function. As Fig. 1, using a Matérn covariance function and varying the 'order parameter' ν. Other hyperparameters (σ_1, σ_2) are chosen to be optimal according to the theory described in Section 2.2.1. The 'Matérn-∞' covariance in the bottom row is identical to a squared-exponential covariance. Note that—depending on the choice of ν—samples from the GP may exhibit short-wavelength variability that is not apparent from the depiction of the overall distribution (middle column).

2.2.1 Covariance hyperparameter selection

From Figs 1 and 2, it is apparent that the covariance function is of prime importance in determining the properties of the GP. As such, one might question how it should be selected in any given case. A hint towards the answer can be seen from Fig. 1: when the imposed length-scale σ_2 is small, the GP displays a level of variability far above that mandated by the data. On the other hand, when σ_2 is large, the observed data set lies in increasingly low-probability regions from the GP perspective. By balancing these two effects, it is possible to identify optimal values for the hyperparameters—or, more generally, to treat them within a hierarchical Bayesian framework and 'integrate out' their influence.

To achieve this, we use eq. (2). Using $\boldsymbol{\sigma}$ to denote a given set of hyperparameter values, we have

$$\mathbb{P}(\hat{\mathbf{y}} \,|\, \boldsymbol{\sigma}) = \frac{1}{\sqrt{(2\pi)^N |\hat{\mathbf{K}}_\sigma + \mathbf{C_d}|}} \exp\left[-\frac{1}{2} (\hat{\mathbf{y}} - \hat{\boldsymbol{\mu}})^{\mathbf{T}} \left(\hat{\mathbf{K}}_\sigma + \mathbf{C_d} \right)^{-1} (\hat{\mathbf{y}} - \hat{\boldsymbol{\mu}}) \right] \tag{9a}$$

with $\hat{\mathbf{K}}_\sigma$ used to emphasize that the matrix $\hat{\mathbf{K}}$ has been computed for that particular choice of $\boldsymbol{\sigma}$. Then, by Bayes' theorem,

$$\mathbb{P}(\boldsymbol{\sigma} \,|\, \hat{\mathbf{y}}) \propto \mathbb{P}(\hat{\mathbf{y}} \,|\, \boldsymbol{\sigma}) \mathbb{P}(\boldsymbol{\sigma}) \tag{9b}$$

where $\mathbb{P}(\boldsymbol{\sigma})$ is a hyperprior describing our assumptions about $\boldsymbol{\sigma}$. It is therefore straightforward to 'map out' the probability associated with different values of the hyperparameters, and to identify those that are most-probable in light of the available data (see also, e.g. Simons & Olhede 2013). In principle, one may also integrate over hyperparameter space to obtain function predictions that account for plausible variability in $\boldsymbol{\sigma}$,

$$\mathbb{P}(f(x) \,|\, \hat{\mathbf{y}}) = \int \mathbb{P}(f(x) \,|\, \hat{\mathbf{y}}, \boldsymbol{\sigma}) \mathbb{P}(\boldsymbol{\sigma} \,|\, \hat{\mathbf{y}}) \, \mathrm{d}\boldsymbol{\sigma} \,. \tag{9c}$$

Clearly, whether or not this integral can be evaluated in a tractable fashion is problem-dependent, and in the remainder of this paper we consider only the case where a single set of maximum-likelihood hyperparameters are used. We note that—particularly given the results of Part II—this hyperparameter estimation procedure is essentially equivalent to the situation discussed at length in Valentine & Sambridge (2018), where we developed a hierarchical Bayesian approach to the determination of regularization matrices in linearized least-squares inversion.

To illustrate the feasibility of this approach, Fig. 3 shows slices through the distribution $\mathbb{P}(\sigma_1, \sigma_2, \nu \,|\, \hat{\mathbf{y}})$ associated with the examples shown in Figs 1 and 2. We assume that any positive value of each parameter is *a priori* equally probable: that is, we adopt a uniform improper hyperprior. It is apparent that the data strongly prefer a large value of ν: of the examples presented in Fig. 2, the $\nu = \infty$ case is clearly

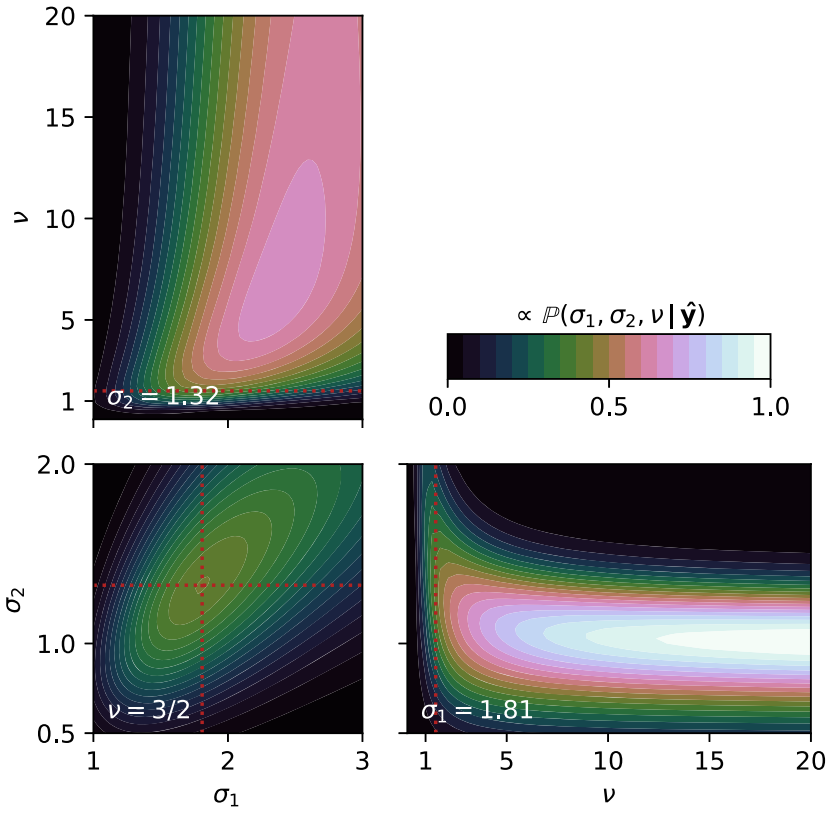

Figure 3. Data support for hyperparameter values, Matérn covariance family. Slices through the (unnormalized) distribution $\mathbb{P}(\sigma_1, \sigma_2, \nu \,|\, \hat{\mathbf{y}})$ for the examples presented in Figs 1 and 2. One hyperparameter is held constant within each slice (see value, inset); dashed lines depict intersections between panels. It is apparent that a well-constrained set of maximum-likelihood hyperparameters may be identified.

most appropriate. It is also evident that for any choice of ν, well-constrained optimal values for σ_1 and σ_2 can be identified (as used in the construction of Fig. 2). In principle, this strategy can be extended to arbitrarily complex covariance functions, allowing the GP approach to adapt to suit a wide variety of functions (e.g. Wilson & Adams 2013)—although in all cases, an overarching Gaussian structure is inevitably assumed.

3 GAUSSIAN PROCESS REGRESSION WITH INDIRECT OBSERVATIONS

So far, we have assumed that we have the ability to directly observe the value of the function of interest at a certain number of points. However, in many problems—particularly those relevant to geophysical inversion—this is not possible, and we only have the ability to measure certain averages of the function. We suppose that the ith piece of data is related to the underlying function by

$$d_i = \int_{\mathcal{X}} w_i(x) f(x) \,\mathrm{d}x \tag{10}$$

for some weighting function $w_i(x)$ and appropriately chosen domain of integration $\mathcal{X}$. We remark that in the case where $w_i(x) = \delta(x - \hat{x}_i)$, the observation d_i is a direct measurement of $f(\hat{x}_i)$, and this formulation reduces to the case that has already been considered.

Integration is a linear transformation, and linear transformations preserve Gaussian statistics. Hence, if we have $f(x) \sim \mathcal{GP}(\mu(x), k(x, x'))$, then we can make a statistical prediction about the value of the ith datum,

$$d_i \sim \mathcal{N}\left(\int_{\mathcal{X}} w_i(x)\mu(x) \,\mathrm{d}x, \iint_{\mathcal{X}^2} w_i(x) k(x, x') w_i(x') \,\mathrm{d}x \,\mathrm{d}x' \right). \tag{11}$$

Moreover, if $\mathcal{L}$ represents a linear transformation, then $\begin{pmatrix} \mathcal{I} & \mathcal{L} \end{pmatrix}$ is also a linear transformation, where $\mathcal{I}$ denotes the identity operator. Therefore, the underlying function and our N observables can be regarded as jointly forming a Gaussian process in a higher-dimensional space, such that

$$\begin{pmatrix} f(x) \\ d_1 \\ \vdots \\ d_N \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu(x) \\ \hat{\boldsymbol{\omega}} \end{pmatrix}, \begin{pmatrix} k(x, x) & \hat{\mathbf{w}}^{\mathbf{T}}(x) \\ \hat{\mathbf{w}}(x) & \hat{\mathbf{W}} + \mathbf{C_d} \end{pmatrix} \right), \tag{12}$$

where

$$[\hat{\boldsymbol{\omega}}]_i = \int_{\mathcal{X}} w_i(u)\mu(u)\,\mathrm{d}u \tag{13a}$$

$$[\hat{\mathbf{w}}(x)]_i = \int_{\mathcal{X}} w_i(u)k(u,x)\,\mathrm{d}u \tag{13b}$$

$$\left[\hat{\mathbf{W}}\right]_{ij} = \iint_{\mathcal{X}^2} w_i(u)k(u,v)w_j(v)\,\mathrm{d}u\,\mathrm{d}v\,. \tag{13c}$$

Thus, our knowledge of $f(x)$ subsequent to making our observations can be expressed

$$f(x)\,|\,\hat{\mathbf{d}} \sim \mathcal{GP}\left(\mu(x) + \hat{\mathbf{w}}^{\mathbf{T}}(x)\left(\hat{\mathbf{W}} + \mathbf{C_d}\right)^{-1}\left(\hat{\mathbf{d}} - \hat{\boldsymbol{\omega}}\right), k(x,x) - \hat{\mathbf{w}}^{\mathbf{T}}(x)\left(\hat{\mathbf{W}} + \mathbf{C_d}\right)^{-1}\hat{\mathbf{w}}(x)\right)\,. \tag{14}$$

This is the core result of the present paper. We emphasize that the result of our inference procedure remains a Gaussian Process, and therefore represents a posterior distribution in function space. For convenience, we introduce the following notation for the posterior mean and covariance functions

$$\tilde{\mu}(x) = \mu(x) + \hat{\mathbf{w}}^{\mathbf{T}}(x)\left(\hat{\mathbf{W}} + \mathbf{C_d}\right)^{-1}\left(\hat{\mathbf{d}} - \hat{\boldsymbol{\omega}}\right) \tag{15a}$$

$$\tilde{k}(x,x') = k(x,x') - \hat{\mathbf{w}}^{\mathbf{T}}(x)\left(\hat{\mathbf{W}} + \mathbf{C_d}\right)^{-1}\hat{\mathbf{w}}(x'), \tag{15b}$$

where we have followed the convention of Tarantola (2005) that tildes denote posterior quantities. We also introduce the vector-valued function

$$\mathbf{a}(x) = \left(\hat{\mathbf{W}} + \mathbf{C_d}\right)^{-1}\hat{\mathbf{w}}(x) \tag{16}$$

to represent the 'classical' inverse operator, allowing us (in particular) to write $\tilde{\mu}(x) = \mu(x) + \mathbf{a}^{\mathbf{T}}(x)(\hat{\mathbf{d}} - \hat{\boldsymbol{\omega}})$.

3.1 Discussion

Eq. (14) provides a route to probabilistic inference in function spaces (albeit for the restricted class of functions that are representable by Gaussian processes). We emphasize that—as with other continuous inversion approaches—this theory avoids the need to introduce any basis functions for the model space, removing the risk of systematic errors arising from an inadequate choice of parametrization. Typically, data kernels w_i, and their integrals, can be calculated to arbitrary numerical accuracy if appropriate computational techniques are used. Typically, these are based on some sort of adaptive scheme, designed to ensure that errors remain below some user-determined value. Thus, the GP framework offers a route by which users can be confident that their results are accurate, subject only to the assumptions inherent to the GP (and, of course, whatever physical assumptions may be inherent in the definition of the kernels).

Readers may recognize certain similarities between eq. (14) and the well-known results for probabilistic discretized inversion set out in Tarantola & Valette (1982). As we show in Part II of this paper, the two are precisely equivalent as the dimension of the discretized model space tends to infinity. We also note that the work of Tarantola & Nercessian (1984) is effectively based on an application of eq. (14), in the specific case where $k(x,x')$ has the form of a squared-exponential, and that this formulation has subsequently been used by others (Montagner & Nataf 1988; Montagner & Tanimoto 1990, 1991).

3.1.1 Resolution

To understand the resolution properties of GP-driven inference, we can consider applying eq. (14) to synthetic data, computed according to eq. (10) for some known model $f_0(x)$. Considering the posterior mean only, we would then have

$$\tilde{\mu}(x) - \mu(x) = \iint_{\mathcal{X}^2} k(x,u)\left\{\sum_{ij} w_i(u)\left[\left(\hat{\mathbf{W}} + \mathbf{C_d}\right)^{-1}\right]_{ij} w_j(v)\right\}\left(f_0(v) - \mu(v)\right)\,\mathrm{d}u\,\mathrm{d}v, \tag{17}$$

where we have exploited the symmetry of $k(x,x')$. This expression has the form

$$\tilde{\mu}(x) - \mu(x) = \int_{\mathcal{X}} R(x,u)\left[f_0(u) - \mu(u)\right]\,\mathrm{d}u, \tag{18}$$

where

$$R(x,x') = \int_{\mathcal{X}} k(x,u)\left\{\sum_{ij} w_i(u)\left[\left(\hat{\mathbf{W}} + \mathbf{C_d}\right)^{-1}\right]_{ij} w_j(x')\right\}\,\mathrm{d}u\,. \tag{19}$$

We refer to $R(x,x')$ as the 'resolution kernel', and it can be regarded as describing the filter through which inference results have been obtained; in order to obtain perfect recovery of the underlying function, we would require $R(x,x') = \delta(x - x')$. We can express the posterior covariance

function in terms of the resolution kernel,

$$\begin{aligned}\tilde{k}(x,x') &= k(x,x') - \int_{\mathcal{X}} R(x,u)k(u,x')\,\mathrm{d}u \\ &= \int_{\mathcal{X}} [\delta(x-u) - R(x,u)]\,k(u,x')\,\mathrm{d}u \end{aligned} \tag{20}$$

mirroring a well-known result from discrete inverse theory (Tarantola 2005). It is worth noting that this resolution analysis is valid for an entirely general choice of $f_0(x)$: we have not required this function to satisfy the covariance structure that is imposed upon recovered models. Clearly, any mismatch here will remove the possibility of perfect recovery.

The posterior covariance function can also be expressed in the form (see the Appendix)

$$\tilde{k}(x,x') = \iint_{\mathcal{X}^2} [\delta(x-u) - R(x,u)]\,k(u,v)\left[\delta(x'-v) - R(x,v)\right]\,\mathrm{d}u\,\mathrm{d}v + \mathbf{a}^{\mathbf{T}}(x)\mathbf{C_d}\mathbf{a}(x') \tag{21}$$

which bears a useful interpretation. The first term in this expression represents the part of the prior distribution that can never be constrained, due to the limited resolution of the imaging process: it can recognized as the covariance associated with the quantity $\int_{\mathcal{X}}[\delta(x-u) - R(x,u)]\,f(u)\,\mathrm{d}u$ when $f(x) \sim \mathcal{GP}(\mu(x), k(x,x'))$. The second term is simply the action of the (classical) imaging operator $\mathbf{a}$ upon the data covariance, and represents the propagation of observational uncertainty into function space. Of course, the two terms are not wholly independent: written out in full, the quantities $k(x,x')$ and $\mathbf{C_d}$ appear in both. Nevertheless, the partitioning can provide an indication of the relative contributions of prior uncertainty and data uncertainty to the posterior, and particularly how this varies as a function of x.

3.1.2 Information gain

The concept of resolution is a useful tool during the interpretation of results, but it is fundamentally deterministic in character: it considers a single model within the prior and posterior distributions, and does not provide a straightforward quantification of what has been learnt from imaging. To measure this, it is useful to compute the information gain from prior to posterior, which essentially measures the difference between the two distributions. A popular measure for quantifying this is the Kullback-Leibler divergence (Kullback & Leibler 1951). For a prior distribution $p(x)$ and a posterior distribution $\tilde{p}(x)$, this is defined as

$$D_{\mathrm{KL}}(\tilde{p}\|p) = \int \tilde{p}(x)\log\frac{\tilde{p}(x)}{p(x)}\,\mathrm{d}x \tag{22}$$

with integration over the full support of the distributions. If both distributions are Gaussian in form, this integral can be performed analytically: in the specific case of 1-D distributions, we obtain

$$D_{\mathrm{KL}}(\tilde{p}\|p) = \frac{1}{2}\left\{\frac{(\mu-\tilde{\mu})^2}{\sigma^2} + \frac{\tilde{\sigma}^2}{\sigma^2} - \log\frac{\tilde{\sigma}^2}{\sigma^2} - 1\right\}, \tag{23}$$

where μ and σ denote the means and standard deviations of the two distributions. It is straightforward to apply this to compute the information gain in any inference problem, as a function of the spatial variable x; note that $\sigma^2(x) = k(x,x)$. In interpreting this expression, it may be helpful to note that

$$\frac{\tilde{\sigma}^2(x)}{\sigma^2(x)} \equiv \frac{\tilde{k}(x,x)}{k(x,x)} = 1 - \frac{\int_{\mathcal{X}} R(x,u)k(u,x)\,\mathrm{d}u}{k(x,x)}, \tag{24}$$

essentially a measure of how the limited resolution of the imaging setup impacts on features with the characteristics expected in results; clearly, $(\mu(x)-\tilde{\mu}(x))^2/\sigma^2(x)$ is a measure of the significance of any model update relative to the prior. The units of information gain depend on the base of the logarithm in eq. (23); where a natural logarithm is adopted, the quantity is measured in 'nats' (*cf.* 'bits' in base-2).

3.1.3 Average properties of f(x)

For completeness, it is useful to also state the following result, which follows straightforwardly from our earlier discussion: if we wish to estimate some average property of $f(x)$, rather than its point value, we can use

$$\int_{\mathcal{X}} f(x)v(x)\,\mathrm{d}x \,|\, \hat{\mathbf{d}} \sim \mathcal{N}\left(\nu + \hat{\mathbf{v}}^{\mathbf{T}}\left(\hat{\mathbf{W}} + \mathbf{C_d}\right)^{-1}\left(\hat{\mathbf{d}} - \hat{\boldsymbol{\omega}}\right), \kappa - \hat{\mathbf{v}}^{\mathbf{T}}\left(\hat{\mathbf{W}} + \mathbf{C_d}\right)^{-1}\hat{\mathbf{v}}\right) \tag{25a}$$

with

$$\nu = \int_{\mathcal{X}} \mu(x)v(x)\,\mathrm{d}x \tag{25b}$$

$$\kappa = \iint_{\mathcal{X}^2} v(x)k(x,x')v(x')\,\mathrm{d}x\,\mathrm{d}x' \tag{25c}$$

$$[\hat{\mathbf{v}}]_i = \int_{\mathcal{X}} [\hat{\mathbf{w}}(x)]_i \, v(x) \, \mathrm{d}x$$
$$= \iint_{\mathcal{X}^2} w_i(u) k(u, x) v(x) \, \mathrm{d}u \, \mathrm{d}x . \tag{25d}$$

If required, it is straightforward to quantify the information gain associated with this inference by using eq. (23).

3.1.4 The gradient of f

Finally, we may sometimes wish to characterize the rate of change of the underlying function, as distinct from its value. Again, this is straightforward, by using the result that, for $f(x) \sim \mathcal{GP}(\mu(x), k(x, x'))$

$$\frac{\mathrm{d}f}{\mathrm{d}x} \sim \mathcal{GP}\left(\frac{\mathrm{d}\mu}{\mathrm{d}x}, \frac{\mathrm{d}^2 k(x, x')}{\mathrm{d}x \mathrm{d}x'}\right) , \tag{26}$$

with higher derivatives following directly. This can be used in conjunction with the posterior process, eq. (14), and it is easy to verify that only the derivatives of the prior covariance function $k(x, x')$ are required. As already noted, the differentiability properties of the covariance function are central to the differentiability of the GP.

3.2 Connections to Backus–Gilbert inversion

It is convenient at this point to compare the GP inference procedure to the seminal work of Backus & Gilbert (1968), who also consider inference for $f(x)$ in problems of the form

$$d_i = \int_{\mathcal{X}} w_i(x) f(x) \, \mathrm{d}x \, . \tag{27}$$

For any set of weights a_i, this implies that

$$\sum_i a_i d_i = \int_{\mathcal{X}} \left[\sum_i a_i w_i(x)\right] f(x) \, \mathrm{d}x \tag{28}$$

and thus the distribution

$$\psi(x, \mathbf{a}) = \sum_i a_i w_i(x) \tag{29}$$

represents an average property of $f(x)$ that may be estimated from a linear combination of the available data. By tuning the a_i, it is possible to adjust the implied averaging kernel, and thus one may frame an optimization problem to identify the a_i that most-closely represent a particular desired average. This is the essence of the Backus–Gilbert approach. In practice, it may sometimes be desirable to add additional constraints, such as a requirement that $\int \psi(x, \mathbf{a}) \, \mathrm{d}x = 1$, to ensure the average is physically interpretable (e.g. Zaroli 2016).

In order to identify the weights $\mathbf{a}$ that bring ψ as close as possible to target averaging function, we must first define the measure by which we will assess 'closeness'. A variety of choices are possible, depending on the situation of interest and decisions here may have significant impact on results and their interpretation. For present purposes, we focus on one choice: we suppose that the size of a function u is to be quantified by the measure

$$\mathcal{L}[u] = \left[\iint u(x) k(x, x') u(x') \, \mathrm{d}x \, \mathrm{d}x'\right]^{\frac{1}{2}} , \tag{30}$$

where $k(x, x')$ is some symmetric, positive-definite function. The similarity of any two functions u and v can then be assessed using a distance measure $\mathcal{D}[u, v] = \mathcal{L}[u - v]$.

If we wish to construct a Backus–Gilbert estimate for the value of our function $f(x)$ at some specific point x_0, we wish to make $\psi(x, \mathbf{a})$ an approximation to $\delta(x - x_0)$. We therefore seek the minimum of $\mathcal{D}[\psi(x, \mathbf{a}), \delta(x - x_0)]$ with respect to the parameters $\mathbf{a}$. In fact, it is convenient (and equivalent) to minimize the square of this quantity, and differentiating with respect to a general element of $\mathbf{a}$ yields

$$\frac{\partial}{\partial a_\lambda} \mathcal{D}^2[\psi(x, \mathbf{a}), \delta(x - x_0)] = 2 \iint_{\mathcal{X}^2} \left[\sum_i a_i w_i(x) - \delta(x - x_0)\right] k(x, x') w_\lambda(x') \, \mathrm{d}x \, \mathrm{d}x'$$
$$= 2 \left[\sum_i \hat{W}_{\lambda i} a_i - \hat{w}_\lambda(x_0)\right] \tag{31}$$

with $\hat{W}$ and $\hat{w}(x)$ defined as in eq. (13). For $\mathcal{D}[\psi, \delta]$ to be minimal, this derivative must be zero for all elements of $\mathbf{a}$. This indicates that the optimal choice (according to our definition of $\mathcal{L}$) is

$$\mathbf{a_0} = \hat{\mathbf{W}}^{-1} \mathbf{w}(x_0) \tag{32}$$

and hence we obtain the Backus–Gilbert estimator $f(x_0) \approx \mathbf{w}^{\mathbf{T}}(x_0)\hat{\mathbf{W}}^{-1}\mathbf{d}$. Clearly, this requires $\hat{\mathbf{W}}$ to be invertible—or, equivalently, the function $\mathcal{D}[\psi, \delta]$ to have a unique global minimum—which cannot be guaranteed. One may therefore choose to regularize the optimization by preferring small $\mathbf{a}$, and instead minimize

$$\chi^2(x_0, \mathbf{a}) = \mathcal{D}^2[\psi(x, \mathbf{a}), \delta(x - x_0)] + \mathbf{a}^{\mathbf{T}}\mathbf{C}_{\mathbf{d}}\mathbf{a}. \tag{33}$$

This leads to a Backus–Gilbert estimator $f(x_0) \approx \mathbf{w}^{\mathbf{T}}(x_0)\left(\hat{\mathbf{W}} + \mathbf{C}_{\mathbf{d}}\right)^{-1}\mathbf{d}$. This is precisely the posterior mean of our Gaussian Process solution, eq. (15a), in the case of a zero prior mean; other choices of prior mean function can straightforwardly be incorporated into the foregoing analysis as a correction to the data applied before inversion.

This does not imply that the GP framework is equivalent to Backus–Gilbert theory: the two are built upon very different axioms and philosophies. However, given the above choices, both provide numerically consistent estimates for the value of $f(x_0)$. Recognizing this similarity may prove useful for interpretation of results, allowing analysis using one framework to be related to the other. Moreover, if eq. (33) is written out in full, it can be recognized as being closely related to eq. (21). In fact, noting that $\psi(x, \mathbf{a_0}) = R(x_0, x)$ leads us to

$$\chi^2(x_0, \mathbf{a_0}) = \tilde{k}(x_0, x_0), \tag{34}$$

where $\tilde{k}$ is the posterior covariance function of our GP solution. This suggests that the GP inference procedure is, in a certain sense, constructed to minimize the point-wise uncertainty associated with results.

The Backus–Gilbert approach is not restricted to the construction of point estimators: indeed, one of its strengths is that it allows any average property of the underlying system to be the target of an inference procedure, and in many problems the measurement kernels $w_i(x)$ are such that bulk properties can be constrained with far greater accuracy than point values. In order to construct an estimator for the average $\int_{\mathcal{X}} f(x)v(x)\,\mathrm{d}x$, we would minimize $\mathcal{D}[\psi, v]$. It is straightforward to show that doing so, using the function size measure $\mathcal{L}$ defined above with a regularization term as in eq. (33), leads to the mean of eq. (25a). Again, a clear consistency can be identified between the deterministic and probabilistic formulations of continuous inverse theory.

3.3 Weakly non-linear data–model relationships

So far, we have focused only on cases where the data is linearly related to the underlying model, according to eq. (10). In many geophysical problems, this is not the situation; typically, the 'weighting function' $w_i(x)$ can only be computed within a specific model realization. For example, in a typical seismic problem, observations may be treated as representative of the average structure along the ray path between source and receiver; however, the path can only be determined if a particular structure is assumed. In other words, a more accurate statement of the data–model relationship is

$$d_i = \int_{\mathcal{X}} w_i[x, f]f(x)\,\mathrm{d}x\,. \tag{35}$$

In the most general case, it is difficult to proceed beyond this point except through sampling: the distribution of data will not be Gaussian, even if our knowledge of f is assumed to be represented by a GP, and hence there are few analytical results that can be exploited.

If, however, the non-linearity is weak, approximate results may be obtained by adopting the following iterative scheme:

(i) Define a prior distribution as usual, $f(x) \sim \mathcal{GP}(\mu(x), k(x, x'))$;
(ii) Compute the w_i using the mean model, $w_i[x, \mu]$;
(iii) Perform inference according to eq. (14), obtaining posterior mean $\tilde{\mu}$;
(iv) Recompute the w_i for this posterior mean, $w_i[x, \tilde{\mu}]$;
(v) Compute a new estimate of the posterior distribution according to eq. (14) using these updated w_i;
(vi) Repeat steps (iv)–(v) until convergence.

Note that the prior remains $f(x) \sim \mathcal{GP}(\mu(x), k(x, x'))$ throughout: in effect, every 'iteration' constitutes the inference being performed afresh, using a refined estimate of the w_i each time.

3.4 Other Gaussian Process approaches

Our focus in this paper has been on representing the model function as a GP, and then deriving an analytical inversion framework from this. However, this is not the only way in which GPs can be used when the goal is to solve an inverse problem. Two alternative general strategies present themselves, and it is instructive to briefly mention them in the present paper. First, one may use a GP to provide an approximate representation of the forward model, by asserting that

$$d[f] \sim \mathcal{GP}\left(\mu_d[f], k_d[f, f']\right)\,. \tag{36a}$$

For simplicity, we treat the observable d as a single scalar quantity, although extension to more complicated settings is straightforward. This falls into a more general class of approach known as 'surrogate modelling': the GP serves as a replacement for (typically) a more expensive accurate calculation. A certain number of accurate simulations are performed, to generate a 'training data set' $\mathbb{T}$, consisting of N examples of

models and their corresponding datum, $(f^{(i)}, d^{(i)})$. Then, the GP is conditioned on these, and can interpolate between them to predict d for any model, according to

$$d[f] \,|\, \mathbb{T} \sim \mathcal{N} \left(\mu_d[f] + \begin{pmatrix} k_d[f, f^{(1)}] \\ \vdots \\ k_d[f, f^{(N)}] \end{pmatrix}^{\mathbf{T}} \begin{pmatrix} k_d[f^{(1)}, f^{(1)}] & \cdots & k_d[f^{(1)}, f^{(N)}] \\ \vdots & \ddots & \vdots \\ k_d[f^{(N)}, f^{(1)}] & \cdots & k_d[f^{(N)}, f^{(N)}] \end{pmatrix}^{-1} \begin{pmatrix} d^{(1)} - \mu_d[f^{(1)}] \\ \vdots \\ d^{(N)} - \mu_d[f^{(N)}] \end{pmatrix}, \right.$$

$$\left. k_d[f, f] - \begin{pmatrix} k_d[f, f^{(1)}] \\ \vdots \\ k_d[f, f^{(N)}] \end{pmatrix}^{\mathbf{T}} \begin{pmatrix} k_d[f^{(1)}, f^{(1)}] & \cdots & k_d[f^{(1)}, f^{(N)}] \\ \vdots & \ddots & \vdots \\ k_d[f^{(N)}, f^{(1)}] & \cdots & k_d[f^{(N)}, f^{(N)}] \end{pmatrix}^{-1} \begin{pmatrix} k_d[f, f^{(1)}] \\ \vdots \\ k_d[f, f^{(N)}] \end{pmatrix} \right) . \tag{36b}$$

For a suitable choice of covariance function, it is also possible to obtain Fréchet derivative of d with respect to f. Thus, the GP can be used within a sampling- or optimization-based framework that seeks to use $d[f]$ to explain observations. Since the inverse matrix in eq. (36b) depends only on the training data, it can be precomputed and stored, allowing $d[f]$ to be computed cheaply and straightforwardly. This may be several orders of magnitude faster than evaluation of the accurate numerical forward model, enabling use of inversion approaches that would otherwise be computationally intractable. Of course, this comes with potential for degradation in results, since the GP interpolation may omit facets of the data–model relationship that are significant for the scenario of interest. To mitigate this, it is possible to use hybrid approaches, whereby the GP is used to identify promising candidate models for which accurate simulations are performed, and then these are used to refine the GP predictions (e.g. Wang *et al.* 2016).

Surrogate-based approaches (including frameworks built around tools other than GPs) have found application across many areas of the physical sciences and engineering, and are increasingly being seen in the geosciences (e.g. Quiepo *et al.* 2005; Sóbester & Forrester 2007; Shahriari *et al.* 2016). They are also closely related to popular techniques such as the Neighbourhood Algorithm (Sambridge 1999a, b), which in effect builds a surrogate model for a Likelihood calculation, based on a nearest-neighbour interpolation scheme. Although surrogate modelling can be implemented using GPs, it is a fundamentally different approach from that set out in this paper. In particular, it is not—of itself—a technique for solving inverse problems. Instead, it is a tool that can be used within an inversion framework to reduce computational costs. It is built on some fundamentally different assumptions: chiefly, that $f(x)$ is a deterministic entity, not a probabilistic one, and that all knowledge of the physical relationship between data and model must be 'learned' from examples rather than imposed *a priori* through a relationship such as eq. (10).

The second alternative GP-based framework starts from an assertion that

$$f(x, d) \sim \mathcal{GP}\left(\mu_f(x, d), k_f(x, d; x', d')\right). \tag{37}$$

This implies that the function f depends on the data—a significant change from all earlier discussion, which can be rationalized by regarding $f(x, d)$ as representing 'our knowledge' about the target function, rather than the target function itself. Again, a training data set $\mathbb{T}$ can be used to condition the GP on known function values corresponding to particular choices of x and d. The resulting expression is rather unwieldy, and for present purposes can be summarized as taking the form

$$f(x, d) \sim \mathcal{N}\left(\mu_f(x, d) + \mathbf{k}^{\mathbf{T}}(x, d, \mathbb{T})\mathbf{K}^{-1}(\mathbb{T})\left[\mathbf{f}(\mathbb{T}) - \boldsymbol{\mu}_{\mathbf{f}}(\mathbb{T})\right], k_f(x, d; x, d) - \mathbf{k}^{\mathbf{T}}(x, d, \mathbb{T})\mathbf{K}^{-1}(\mathbb{T})\mathbf{k}(x, d, \mathbb{T})\right). \tag{38}$$

The quantities in this expression take similar form to their counterparts elsewhere in this paper, and it is not necessary to explicitly define them here. Again, the GP can be understood as an interpolator: results are based on any training data that is 'close' to the desired point in (x, d)-space, as measured by the covariance function. This is rather similar in philosophy to the 'prior sampling' approach discussed at length by Käufl *et al.* (2016), where a training data set is used to construct approximations to posterior marginal distributions implemented using a particular class of neural network.

Some key distinctions between this approach and the method introduced in this paper deserve to be highlighted. Perhaps the most important is the manner in which the observations enter the inverse problem: in eq. (38), they are introduced via the covariance function, whereas eq. (14) arises from directly conditioning the joint data–model distribution upon the observations. This points to another difference: eq. (38) treats the observations as a deterministic entity, and it is not immediately obvious how data noise should be handled in such a framework; our approach regards the data as inherently Gaussian distributed.

Both alternative frameworks offer certain advantages: in particular, they involve no assumption of linearity, and so can be straightforwardly applied in a range of problems. On the other hand, this comes at a cost: our approach embeds physical knowledge into the GP covariance (via integrals involving both k and w), whereas the alternatives do not. Instead, they require the data–model relationship to be entirely learned from examples. Whether or not this is achievable with satisfactory accuracy in any given case is likely to be problem dependent.

4 A SIMPLE EXAMPLE: EARTH'S RADIAL DENSITY STRUCTURE

To illustrate the practical application of our approach, we consider the classic problem of constraining Earth's radial density structure. We assume three things are known about the planet: its mass, M; its moment of inertia, I; and the typical density of near-surface rocks,

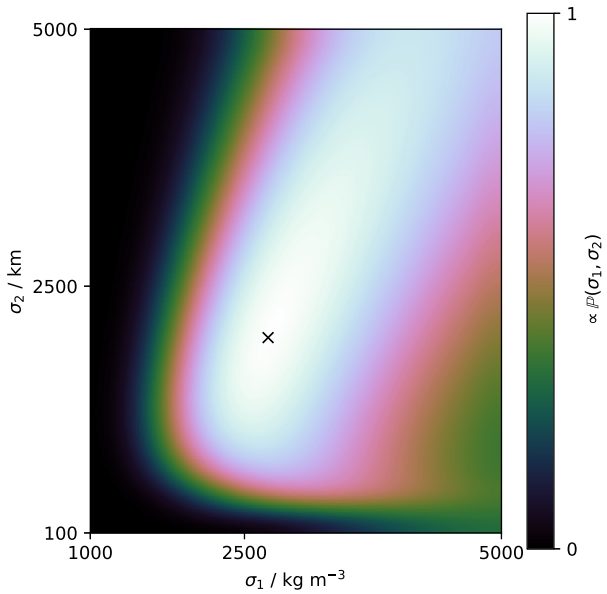

Figure 4. Hyperparameter selection. The (unnormalized) probability distribution $\mathbb{P}(\sigma_1, \sigma_2 \,|\, \mathbf{d})$ for the radial density structure problem, assuming a Matérn-$\frac{3}{2}$ covariance function. A cross marks the location of the most-probable set of hyperparameters, which are then used for the construction of Fig. 5.

ρ_{surf}—specifically, the average density of the upper 25 km of the crust. We use $\rho(r)$ to denote the density function, with $r = a$ at the Earth's surface, and assume that the planet is spherically symmetric. The geophysical parameters are taken to be

$$M = 4\pi \int_0^a \rho(r) r^2 \,\mathrm{d}r = (5.9733 \pm 0.0090) \times 10^{24}\,\mathrm{kg} \tag{39a}$$

$$I = \frac{8\pi}{3} \int_0^a \rho(r) r^4 \,\mathrm{d}r = (8.018 \pm 0.012) \times 10^{37}\mathrm{m}^2\,\mathrm{kg}, \tag{39b}$$

where numerical values have been taken from Chambat & Valette (2001). For the sake of simplicity in this illustrative example, we choose to ignore the fact that these are not independently determined, and thus that their uncertainties will be correlated. The near-surface density is given by

$$\rho_{\text{surf}} = \frac{1}{a - a_0} \int_{a_0}^a \rho(r) \,\mathrm{d}r = (2800 \pm 200)\,\mathrm{kg\,m^{-3}}, \tag{39c}$$

where $a_0 = a - 25\,\mathrm{km}$, and where the numerical value is loosely based on the work of Christensen & Mooney (1995).

For numerical reasons, it is desirable for all data to be of roughly similar magnitude, and thus, adopting $a = 6371.230\,\mathrm{km}$, we work with $I/a^2 = (1.975 \pm 0.003) \times 10^{24}\,\mathrm{kg}$ and $a^3\rho_{\text{surf}} = (7.2 \pm 0.5) \times 10^{23}\,\mathrm{kg}$; again, we neglect the uncertainty associated with the planetary radius. In the language of eq. (10), we are therefore working with the data kernels

$$w_1(r) = 4\pi r^2 \tag{40a}$$

$$w_2(r) = \frac{8\pi}{3a^2} r^4 \tag{40b}$$

$$w_3(r) = \frac{a^3}{a - a_0} H(r - a_0), \tag{40c}$$

where $H(x)$ is the Heaviside step function, and the domain of integration $\mathcal{X}$ is $[0, a]$.

Initially, we assume that the density function is everywhere characterized by a single covariance function, which we choose to be Matérn-$\frac{3}{2}$ in form. As discussed in Section 2.2, this depends upon two hyperparameters: one governing the amplitude of the prior, and the other its characteristic length-scale. Fig. 4 shows the strength of support provided by the data for different values of these hyperparameters: the most-probable choice is $\sigma_1 = 2730\,\mathrm{kg\,m^{-3}}$ and $\sigma_2 = 2000\,\mathrm{km}$. Using these values, we perform inference according to eq. (14), and obtain the results shown in Fig. 5. These are broadly consistent with the density structure of reference models such as PREM (Dziewonski & Anderson 1981), although our results are obviously much less well-constrained, due to the limited nature of the data used. By plotting information gain as a function of radius, the limited sensitivity of our data set to deep structure is readily appreciated, and by plotting $\mathbf{a}^{\mathbf{T}}(x)\mathbf{C_d}\mathbf{a}(x)/\tilde{k}(x,x)$ we see that our posterior uncertainty is attributable to the limited resolution of the data, except in the uppermost part where data uncertainties dominate.

Of course, this model assumes there are no discontinuities within the Earth. To allow for these, we can use a non-stationary covariance function, which includes knowledge of the inner/outer-core boundary (assumed to be located at $r = 1221.5\,\mathrm{km}$) and the core/mantle boundary (at $r = 3480\,\mathrm{km}$). We assume that within each unit, our function $\rho(r)$ remains characterized by a Matérn-$\frac{3}{2}$ covariance, but that there is no

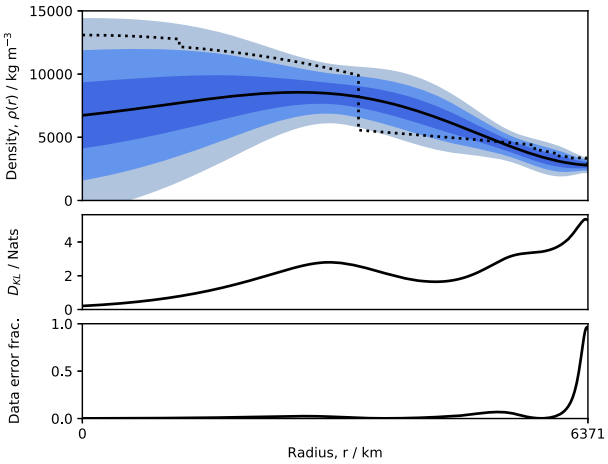

Figure 5. Earth's radial density structure. GP inference results based on three pieces of data: mass, moment of inertia, and average density of near-surface rocks, using a Matérn-$\frac{3}{2}$ covariance function with optimal hyperparameters (see Fig. 4). As in earlier figures, the solid line in the uppermost plot corresponds to the posterior mean, while shaded regions denote 1σ-, 2σ- and 3σ-uncertainty bands. For reference, the dotted line depicts the density structure of PREM (Dziewonski & Anderson 1981), although this is based on a richer data set than ours. The middle plot depicts the information gain (Kullback-Liebler divergence, eq. 23) as a function of radius, quantifying the change in constraint from prior to posterior. In the lowermost plot, we show the fraction of the posterior attributable to propagation of data uncertainty (see eq. 21and accompanying discussion), demonstrating that posterior uncertainties are largely due to an inability to resolve structure using available data.

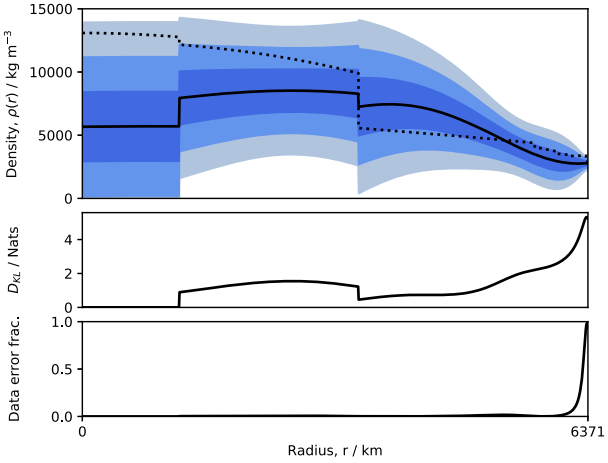

Figure 6. Earth's radial density structure, including inner and outer cores. Results of GP inference, as in Fig. 5, except that we introduce discontinuities corresponding to the inner/outer core- and core–mantle boundaries. Again, lower plots depict information gain and the fraction of uncertainty attributable to data errors.

correlation with any point outside the unit [i.e. $k(x, x') = 0$ unless x and x' both lie in the same region of the model]. We assume that the prior uncertainty hyperparameter, σ_1, is shared between all three regions, but that each has a different characteristic length-scale σ_2. Using the hyperparameter optimization procedure, we find that the data prefers $\sigma_1 = 2755\,\mathrm{kg\,m^{-3}}$, and characteristic length-scales of 2001 km (inner core), 2629 km (outer core) and 1113 km (mantle), although the length-scale in the inner-core is poorly constrained. Performing inference based upon this covariance structure, we obtain the results shown in Fig. 6. Again, results are broadly compatible with models based on more complex data sets.

To illustrate the application of eq. (25a), we also estimate the density jump at core–mantle boundary in this model. We define this jump as being the difference between the average density in the 100 km below the boundary, and the average density in the 100 km above the boundary. This corresponds to an averaging function

$$v(r) = \begin{cases} 1 & 3380 \leq r < 3480 \\ -1 & 3480 \leq r < 3580 \\ 0 & \text{otherwise} \end{cases} \tag{41}$$

for r with units of kilometres. We therefore find $\Delta\tilde{\rho}_{\mathrm{CMB}} = (1015 \pm 3656)\,\mathrm{kg\,m^{-3}}$ for the inference shown in Fig. 6, giving a 61 per cent chance that the density change at this interface is indeed positive. In our prior, this quantity was $\Delta\rho_{\mathrm{CMB}} = (0 \pm 3895)\,\mathrm{kg\,m^{-3}}$, and thus our data has provided only 0.08 nats of information about the density jump.

5 CONCLUDING REMARKS

In this paper, we have developed a framework for tackling probabilistic, continuous inverse problems. This relies on two core assumptions: first, that our prior on the target function can be expressed as a Gaussian Process; and second, that data and model are connected by a linear relationship, as in eq. (10). As discussed in Section 3.3, cases where the data–model relationship is weakly non-linear may be handled, approximately, via an iterative approach. Our approach is consistent with Backus & Gilbert's deterministic treatment of continuous inverse theory, and we have derived the conditions under which the two approaches yield equivalent results.

The great advantage of adopting a continuous—as opposed to discrete—approach to inversion is that one avoids any need to identify an appropriate model parametrization. This is particularly significant in circumstances where the system may exhibit features that are difficult to represent accurately using typical basis sets, such as discontinuities. We discuss this in more detail in Part II of this paper, where we show that the GP approach is effectively equivalent to the probabilistic least-squares framework of Tarantola & Nercessian (1984), within an infinite-dimensional model space.

A powerful feature of the GP approach is the ability to govern the general characteristics of the recovered solution by specifying a covariance function, while allowing the data to decide properties such as scale-lengths and uncertainties (as discussed in Section 2.2.1. Because all quantities are specified in physical space, rather than in an abstract 'model space', the researcher is well-placed to make realistic and meaningful choices, and to appreciate how these impact upon results. There is much scope for innovation in this area: for example, one could adapt the trans-dimensional inference philosophy (e.g. Sambridge *et al.* 2012) to determine where discontinuities should be placed within the covariance function. This is also an area where the machine learning community is undertaking much research, which the geosciences can undoubtedly exploit.

Of course, the advantages of continuous inversion are not without challenges. In particular, the GP framework requires inversion of a matrix with dimension governed by the number of data, and this may impose practical restrictions upon the scale of problem that can realistically be tackled. This issue has been considered at length in the GP literature, and various approximation strategies have been developed to allow reduction of costs, which may be applicable in the present case (e.g. Quiñonero Candela *et al.* 2007; Wilson & Nickisch 2015; Ambikasaran *et al.* 2016). A second computational issue arises from the need to compute weighting functions $w(x)$, and the various integrals in which they appear, in a manner consistent with the 'continuous' philosophy. Care must be taken here, as forward modelling codes often rely internally on some form of discretization, and it is important to ensure that this does not adversely affect results. Typically, this will require some form of adaptive scheme, allowing discretization errors to be kept below some user-determined bound. This can be achieved using numerical algorithms such as dense-output ODE solvers (e.g. Hairer *et al.* 1993), but many codes (especially older ones) use fixed-order methods.

Even in cases where it is not practical to use the GP approach directly, it may prove useful as part of a more general solution strategy. In particular, it may offer an efficient route towards defining prior distributions for sampling-based strategies. For example, the work of de Wit *et al.* (2013, 2014) required the generation of 1-D earth models distributed about PREM, but honouring physical constraints such as those on mass and moment of inertia. An *ad hoc* procedural strategy for achieving this was used, introducing some complexity into the prior. The results of this paper provide an alternative—and arguably simpler—route to defining, and generating samples from, such a prior.

It is fair to note that the GP inversion framework is built upon some quite strong assumptions about the form taken by the solution (or, perhaps, about the form of our knowledge surrounding that solution). Underpinning our entire theory is an assertion that this can be represented by a GP, with covariance properties conforming to a certain structure. The very nature of inverse problems makes it difficult—if not impossible—to rigorously assess whether these are appropriate assumptions in any given case. Often, physical knowledge will provide some insight into whether a given covariance structure is reasonable: we frequently have justifiable prior information about the interactions and length-scales that are thought to dominate a system. However, as Tarantola (2005, section 5.1.5) shows, knowledge of a correlation function alone is insufficient to characterize a random process. Imposing a Gaussian structure upon the function (in the sense we describe in Section 2) represents a considerable additional restriction. Doing so brings clear practical benefits: as we have seen, it enables calculations to be performed analytically. However, it may not provide a faithful representation of the statistical properties of the underlying physical system. It is incumbent on any potential user of the theory to recognize that such choices form part of the 'prior knowledge' being imposed, to satisfy themselves that doing so is reasonable, and to bear this in mind when interpreting results. We remark that this is not specific to the GP approach: every inversion framework is built on assumptions and choices, even if they are not always clearly stated.

Finally, we emphasize that this work is important on a broader theoretical level, as it connects several strands from the inverse theory literature, and links them to the vast body of research being undertaken in machine learning. In that field, Gaussian Processes are also seen as a central theoretical model, bridging between diverse techniques including neural networks and support vector machines (e.g. Wahba 1999; Rasmussen & Williams 2006; Lee *et al.* 2018). There is currently significant interest in applying machine learning tools to geoscience problems, and it is therefore valuable to have established a common thread connecting such techniques to the classic methods of geophysical inversion. We hope that this can help stimulate continued innovation across the field.

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

Abrahamsen, P., 1997, A review of Gaussian random fields and correlation functions, Tech. Rep. 917, Norweigian Computing Center, Oslo.

Ambikasaran, S., Foreman-Mackey, D., Greengard, L., Hogg, D. & O'Neil, M., 2016. Fast direct methods for Gaussian Processes, *IEEE Trans. Pattern Anal. Mach. Intell.,* **32,** 252–265.

Anderssen, R., Worthington, M. & Cleary, J., 1972. Density modelling by Monte Carlo inversion—I. Methodology, *Geophys. J. R. astr. Soc.,* **29,** 433–444.

Aster, R., Borchers, B. & Thurber, C., 2013. *Parameter Estimation and Inverse Problems,* Academic Press.

Backus, G. & Gilbert, F., 1967. Numerical applications of a formalism for geophysical inverse problems, *Geophys. J. R. astr. Soc.,* **13,** 247–276.

Backus, G. & Gilbert, F., 1968. The resolving power of gross Earth data, *Geophys. J. R. astr. Soc.,* **16,** 169–205.

Backus, G. & Gilbert, F., 1970. Uniqueness in the inversion of inaccurate gross Earth data, *Phil. Trans. of the R. Soc. Lond.,* **266,** 123–192.

Bergen, K., Johnson, P., de Hoop, M. & Beroza, G., 2019. Machine learning for data-drive discovery in solid Earth geoscience, *Science,* **363,** eaau0323.

Chambat, F. & Valette, B., 2001. Mean radius, mass, and inertia for reference Earth models, *Phys. Earth planet. Inter.,* **124,** 237–253.

Christensen, N. & Mooney, W., 1995. Seismic velocity structure and composition of the continental crust: A global view, *Geophys. J. R. astr. Soc.,* **100,** 9761–9788.

de Wit, R., Käufl, P., Valentine, A. & Trampert, J., 2014. Bayesian inversion of free oscillations for Earth's radial (an)elastic structure, *Phys. Earth planet. Inter.,* **237,** 1–17.

de Wit, R., Valentine, A. & Trampert, J., 2013. Bayesian inference of Earth's radial seismic structure from body wave travel times using neural networks, *J. geophys. Int.,* **195**(1), 408–422

Dubrule, O., 2018. Kriging, splines, conditional simulation, Bayesian inversion and ensemble Kalman filtering, in *Handbook of Mathematical Geosciences,* Chap. 1, pp. 3–24, eds Sagar, B., Cheng, Q. & Agterberg, F., Springer.

Dziewonski, A. & Anderson, D., 1981. Preliminary reference Earth model, *Phys. Earth planet. Inter.,* **25,** 297–356.

Dziewonski, A., Chou, T.-A. & Woodhouse, J., 1981. Determination of earthquake source parameters from waveform data for studies of global and regional seismicity, *Geophys. J. R. astr. Soc.,* **86,** 2825–2852.

Hairer, E., Nørsett, S. & Wanner, G., 1993. *Solving Ordinary Differential Equations,* Springer.

Jackson, D., 1976. Most squares inversion, *Geophys. J. R. astr. Soc.,* **81,** 1027–1030.

Jeffreys, H., 1931. *Scientific Inference,* Cambridge Univ. Press.

Jeffreys, H., 1939. *The Theory of Probability,* Oxford Univ. Press.

Krige, D., 1951, A statistical approach to some mine valuations and allied problems at the Witwatersrand. *Master's thesis*, University of Witwatersrand.

Kullback, S. & Leibler, R., 1951. On information and sufficiency, *Ann. Math. Stat.,* **22,** 79–86.

Käufl, P., Valentine, A., de Wit, R. & Trampert, J., 2016. Solving probabilistic inverse problems rapidly with prior samples, *J. geophys. Int.,* **205,** 1710–1728.

Lee, J., Bahri, Y., Novak, R., Schoenholz, S., Pennington, J. & Sohl-Dickstein, J., 2018. Deep neural networks as Gaussian Processes, in *International Conference on Learning Representations*.

Meju, M. & Hutton, V., 1992. Iterative most-squares inversion: application to magnetotelluric data, *J. geophys. Int.,* **108,** 758–766.

Menke, W., 1989. *Geophysical Data Analysis: Discrete Inverse Theory,* Academic Press.

Montagner, J.-P. & Nataf, H.-C., 1988. Vectorial tomography — I. Theory, *Geophys. J.,* **94,** 295–307.

Montagner, J.-P. & Tanimoto, T., 1990. Global anisotropy in the upper mantle inferred from the regionalization of phase velocities, *Geophys. J. R. astr. Soc.,* **95,** 4797–4819.

Montagner, J.-P. & Tanimoto, T., 1991. Global upper mantle tomography of seismic velocities and anisotropies, *Geophys. J. R. astr. Soc.,* **96,** 20337–20351.

Murphy, K., 2012. *Machine Learning: A Probabilistic Perspective,* MIT Press.

Petersen, K. & Pedersen, M., 2015, The matrix cookbook, Tech. rep., Technical University of Denmark.

Pijpers, F. & Thompson, M., 1992. Faster formulations of the optimally localized averages method for helioseismic inversions, *Astron. Astrophys.,* **262,** L33–L36.

Press, F., 1968. Earth models obtained by Monte Carlo inversion, *Geophys. J. R. astr. Soc.,* **73,** 5223–5234.

Quiepo, N., Haftka, R., Shyy, W., Goel, T., Vaidyanathan, R. & Tucker, P., 2005. Surrogate-based analysis and optimization, *Prog. Aerospace Sci.,* **41,** 1–28.

Quiñonero Candela, J., Rasmussen, C. & Williams, C., 2007. Approximation methods for Gaussian Process regression, in *Large-Scale Kernel Machines,* pp. 203–224 , MIT Press.

Rasmussen, C. & Williams, C., 2006. *Gaussian Processes for Machine Learning,* MIT Press.

Sambridge, M., 1999a. Geophysical inversion with a neighbourhood algorithm–I. Searching a parameter space, *J. geophys. Int.,* **138,** 479–494.

Sambridge, M., 1999b. Geophysical inversion with a neighbourhood algorithm–II. Appraising the ensemble, *J. geophys. Int.,* **138,** 727–746.

Sambridge, M., Bodin, T., Gallagher, K. & Tkalcic, H., 2012. Transdimensional inference in the geosciences, *Phil. Trans. R. Soc.,* **371,** doi:10.1098/rsta.2011.0547.

Sambridge, M. & Mosegaard, K., 2002. Monte Carlo methods in geophysical inverse problems, *Rev. Geophys.,* **40**(3), 3–1-3-29.

Shahriari, B., Swersky, K., Wang, Z., Adams, R. & de Freitas, N., 2016. Taking the human out of the loop: a review of Bayesian optimization, *Proc. IEEE,* **104,** 148–175.

Simons, F. & Olhede, S., 2013. Maximum-likelihood estimation of lithospheric flexural rigidity, initial-loading fraction and load correlation, under isotropy, *J. geophys. Int.,* **193,** 1300–1342.

Sóbester, A. & Forrester, A., 2007, On the use of surrogate models in global optimization—a practical approach, Tech. rep., University of Southampton.

Tarantola, A., 2005. *Inverse Problem Theory and Methods for Model Parameter Estimation,* SIAM.

Tarantola, A. & Nercessian, A., 1984. Three-dimensional inversion without blocks, *Geophys. J. R. astr. Soc.,* **76,** 299–306.

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. & Kalnins, L., 2016. An introduction to learning algorithms and potential applications in geomorphometry and earth surface dynamics, *Earth Surf. Dyn.,* **4,** 445–460.

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

Valentine, A.P. & Sambridge, M., 2019. Gaussian process models—II. Lessons for discrete inversion, *Geophys. J. Int.,* in press, doi: 10.1093/gji/ggz521.

Wahba, G., 1999. Support vector machines, reproducing kernel Hilbert spaces and the randomized GACV, *Adv. Kernel Methods-Support Vector Learn.,* **6,** 69–87.

Wang, Z., Hutter, F., Zoghi, M., Matheson, D. & de Freitas, N., 2016. Bayesian optimization in a billion dimensions via random embeddings, *J. Artif. Intellig. Res.,* **55,** 361–387.

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

Wilson, A.G. & Adams, R.P., 2013. Gaussian process kernels for pattern discovery and extrapolation, *Int. Conf. Mach. Learn*., **28**(3), 1067–1075.

Wilson, A.G. & Nickisch, H., 2015. Kernel interpolation for scalable structured Gaussian Processes, in *Proceedings of the 32nd International Conference on Machine Learning,* pp. 1775–1784.

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

Worthington, M., Cleary, J. & Anderssen, R., 1972. Density modelling by Monte Carlo inversion—II. Comparison of recent Earth models, *Geophys. J. R. astr. Soc.,* **29,** 445–457.

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

Zarolli, C., 2019. Seismic tomography using parameter-free Backus-Gilbert-inversion, *J. geophys. Int.,* **218,** 619–630.

APPENDIX: THE POSTERIOR COVARIANCE FUNCTION

To derive eq. (21), we start from

$$\iint_{\mathcal{X}^2} [\delta(x-u) - R(x,u)]\, k(u,v) \left[\delta(x'-v) - R(x',v)\right] \mathrm{d}u\, \mathrm{d}v = k(x,x') - 2\int_{\mathcal{X}} R(x,u)k(u,x')\, \mathrm{d}u + \iint_{\mathcal{X}^2} R(x,u)k(u,v)R(x',v)\, \mathrm{d}u\, \mathrm{d}v, \tag{A1}$$

where we have relied upon the fact that k is symmetric in its arguments. Using eqs (13), (15b) and (19), we see that

$$\int_{\mathcal{X}} R(x,u)k(u,x')\, \mathrm{d}u = k(x,x') - \tilde{k}(x,x') \tag{A2}$$

and

$$\begin{aligned}\iint_{\mathcal{X}^2} R(x,u)k(u,v)R(x',v)\, \mathrm{d}u\, \mathrm{d}v &= \hat{\mathbf{w}}^{\mathbf{T}}(x)\left(\hat{\mathbf{W}} + \mathbf{C_d}\right)^{-1} \hat{\mathbf{W}} \left(\hat{\mathbf{W}} + \mathbf{C_d}\right)^{-1} \hat{\mathbf{w}}(x') \\ &= \hat{\mathbf{w}}^{\mathbf{T}}(x)\left(\hat{\mathbf{W}} + \mathbf{C_d}\right)^{-1} \hat{\mathbf{w}}(x') - \hat{\mathbf{w}}^{\mathbf{T}}(x)\left(\hat{\mathbf{W}} + \mathbf{C_d}\right)^{-1} \mathbf{C_d} \left(\hat{\mathbf{W}} + \mathbf{C_d}\right)^{-1} \hat{\mathbf{w}}(x') \\ &= \int_{\mathcal{X}} R(x,u)k(u,x')\, \mathrm{d}u - \mathbf{a}^{\mathbf{T}}(x)\mathbf{C_d}\mathbf{a}(x'),\end{aligned} \tag{A3}$$

where the middle line follows from the result that $(\mathbf{A}+\mathbf{B})^{-1}\mathbf{A} = \mathbf{I} - (\mathbf{A}+\mathbf{B})^{-1}\mathbf{B}$, for general matrices $\mathbf{A}$ and $\mathbf{B}$. Thus, eq. (A1) becomes

$$\iint_{\mathcal{X}^2} [\delta(x-u) - R(x,u)]\, k(u,v) \left[\delta(x'-v) - R(x',v)\right] \mathrm{d}u\, \mathrm{d}v = \tilde{k}(x,x') - \mathbf{a}^{\mathbf{T}}(x)\mathbf{C_d}\mathbf{a}(x') \tag{A4}$$

from which eq. (21) follows directly.