2

Emerging Directions in Geophysical Inversion

Andrew P. Valentine and Malcolm Sambridge

Abstract: In this chapter, we survey some recent developments in the field of geophysical inversion. We aim to provide an accessible general introduction to the breadth of current research, rather than focusing in depth on particular topics. We hope to give the reader an appreciation for the similarities and connections between different approaches, and their relative strengths and weaknesses.

2.1 Introduction

Geophysics is built upon indirect information. We cannot travel deep into the Earth to directly measure rheological properties, nor journey back through geological time to record the planet's tectonic evolution. Instead, we must draw inferences from whatever observations we can make, constrained as we are to the Earth's surface and the present day. Inevitably, such datasets are sparse, incomplete, and contaminated with signals from many unknown events and processes. We therefore rely on a variety of mathematical, statistical, and computational techniques designed to help us learn from available data. Collectively, these are the tools of 'geophysical inversion', and they lie at the heart of all progress in geophysics.

To achieve this progress, geophysicists have long pioneered – and indeed driven – developments in the mathematical and statistical theory that underpins inference. The acclaimed French mathematician Pierre-Simon Laplace played a central role in our understanding of tidal forcing, developing the theory of spherical harmonics along the way. He is also credited (along with Gauss and Legendre) with the development of the least-squares algorithm and the underpinnings of modern Bayesian statistics – an approach which was subsequently extended and popularised within the physical sciences by Sir Harold Jeffreys (1931, 1939), who is, of course, also well known for his contributions to seismology and solid-earth geophysics (see, e.g., Cook, 1990). Technological developments have also been significant, with (for example) the challenges of handling and processing the huge volumes of data obtained from continuously operating terrestrial and satellite sensor systems stimulating innovation in computational science.

In this chapter, we discuss some current and emerging ideas that we believe to have significance for the broad field of geophysical inversion. In doing so, we aim to not just highlight novelty, but also demonstrate how such 'new' ideas can be connected into the canon of established techniques and methods. We hope that this can help provide insight into the potential strengths and weaknesses of different strategies, and support the interpretation and integration of results obtained using different approaches. Inevitably, constraints of time and space mean that our discussion here remains far from comprehensive; much interesting and important work must be omitted, and our account is undoubtedly biased by our own perspectives and interests. Nevertheless, we hope that the reader is able to gain some appreciation for the current state of progress in geophysical inversion.

In order to frame our discussion, and to enable us to clearly define notation and terminology, we begin with a brief account of the basic concepts of geophysical inversion. For a more in-depth account, readers are encouraged to refer to one of the many textbooks and monographs covering the subject, such as those by Menke (1989), Parker (1994), Tarantola (2005), or Aster et al. (2013).

2.2 Fundamentals

The starting point for any geophysical inversion must be a mathematical description of the earth system of interest. In practical terms, this amounts to specifying some relationship of the form

$$\mathcal{F}[m(\mathbf{x},t),u(\mathbf{x},t)]=0 \tag{2.1}$$

where $m(\mathbf{x},t)$ represents some property (or collection of properties) of the Earth with unknown value that may vary across space, $\mathbf{x}$, and/or time, t; and where $u(\mathbf{x},\ t)$ represents some quantity (or collection of quantities) that can – at least in principle – be measured or observed. Most commonly in geophysics, $\mathcal{F}$ has the form of an integrodifferential operator. Underpinning Eq. (2.1) will be some set of assumptions, $\mathcal{A}$, although these may not always be clearly or completely enunciated.

2.2.1 The Forward Problem

The fundamental physical theory embodied by Eq. (2.1) may then be used to develop predictions, often via a computational simulation. This invariably involves introducing additional

https://doi.org/10.1017/9781009180412.003 Published online by Cambridge University Press

assumptions, $\mathcal{B}$. In particular, it is common to place restrictions on the function m, so that it may be assumed to have properties amenable to efficient computation. For example, it is very common to assert that the function must lie within the span of a finite set of basis functions, $\psi_1, \ldots, \psi_M$, allowing it to be fully represented by a set of M expansion coefficients,

$$m(\mathbf{x}, t) = \sum_{i=1}^{M} m_i \psi_i(\mathbf{x}, t). \tag{2.2}$$

It is important to recognise that such restrictions are primarily motivated by computational considerations, but may impose certain characteristics – such as a minimum length-scale, or smoothness properties – upon the physical systems that can be represented. Nevertheless, by doing so, we enable Eq. (2.1) to be expressed, and implemented, as a 'forward model'

$$u(\mathbf{x}, t) = \mathcal{G}(\mathbf{x}, t, m), \tag{2.3}$$

which computes simulated observables for any 'input model' conforming to the requisite assumptions. Typically, the function $\mathcal{G}$ exists only in the form of a numerical computer code, and not as an analytical expression in any meaningful sense. As a result, we often have little concrete understanding of the function's global behaviour or properties, and the computational cost associated with each function evaluation may be high.

2.2.2 Observational Data and the Inverse Problem

We use $\mathbf{d}$ to represent a data vector, with each element d_i representing an observation made at a known location in space and time, $(\mathbf{x}_i, t_i)$. This is assumed to correspond to $u(\mathbf{x}_i, t_i)$, corrupted by 'noise' (essentially all processes not captured within our modelling assumptions, $\mathcal{A} \cup \mathcal{B}$), any limitations of the measurement system itself, and any preprocessing (e.g. filtering) that has been applied to the dataset. We address the latter two factors by applying transformations (e.g. equivalent preprocessing and filters designed to mimic instrument responses) to the output of our forward model; mathematically, this amounts to composing $\mathcal{G}$ with some transfer function $\mathcal{T}$. For notational convenience, we define a new function, $\mathbf{g}$, which synthesises the entire dataset $\mathbf{d}$: $[\mathbf{g}(m)]_i = \mathcal{T} \circ \mathcal{G}(\mathbf{x}_i, t_i, m)$. We also introduce the concept of a data covariance matrix, $\mathbf{C_d}$, which encapsulates our assumptions about the uncertainties and covariances within the dataset. The fundamental goal of inversion is then to find – or somehow characterise – m such that $\mathbf{g}(m)$ matches or explains $\mathbf{d}$.

Since $\mathbf{d}$ contains noise, we do not expect any model to be able to reproduce the data perfectly. Moreover, the forward problem may be fundamentally non-unique: it may generate identical predictions for two distinct models. As such, there will typically be a range of models that could be taken to 'agree with' observations. We must therefore make a fundamental decision regarding the approach we wish to take. We may:

1. Seek a single model, chosen to yield predictions that are 'as close as possible' to the data, usually with additional requirements that impose characteristics we deem desirable and ensure that a unique solution exists to be found, e.g. that the model be 'as smooth as possible';
2. Seek a collection or ensemble of models, chosen to represent the spectrum of possibilities that are compatible with observations – again, perhaps tempered by additional preferences;
3. Disavow the idea of recovering a complete model, and instead focus on identifying specific characteristics or properties that must be satisfied by any plausible model.

In the context of this chapter, we have deliberately framed these three categories to be quite general in scope. Nevertheless, readers may appreciate some specific examples: the first category includes methods based upon numerical optimisation of an objective function, including the familiar least-squares algorithm (e.g. Nocedal and Wright, 1999), while Markov chain Monte Carlo and other Bayesian methods fall within the second (e.g. Sambridge and Mosegaard, 2002); Backus–Gilbert theory (e.g. Backus and Gilbert, 1968) lies within the third. Each of these groups is quite distinct – at least in philosophy – from the others, and in the remainder of this chapter we address each in turn.

2.3 Single Models

Before we can set out to find the model that 'best' explains the data, we must introduce some measure of the agreement between observations and predictions. This 'misfit function' or 'objective function' is of fundamental importance in determining the properties of the recovered model and the efficiency of the solution algorithms that may be available to us. In general, misfit functions take the form

$$\phi(m) = \phi_d(\mathbf{d}, \mathbf{g}(m)) + \phi_m(m), \tag{2.4}$$

where ϕ_d is a metric defined in the 'data space', measuring how far a model's predictions are from observations, and ϕ_m is a 'regularisation' or 'penalty' term (see Fig. 2.1). This encapsulates any preferences we may have regarding the solution, and aims to ensure that the function ϕ has a unique minimum.

Once a misfit function has been defined, it is conceptually straightforward to search for the model that minimises $\phi(m)$. However, it is often challenging to achieve this in practice. The most complete characterisation of ϕ comes from a grid-search strategy, with systematic evaluation of the function throughout a discretised 'model space' (typically following Eq. 2.2). This is viable for small problems, and is commonly encountered in the geophysical literature (e.g. Sambridge and Kennett, 1986; Dinh and Van der Baan, 2019; Hejrani and Tkalčić, 2020), but the computational costs of evaluating the forward model,

https://doi.org/10.1017/9781009180412.003 Published online by Cambridge University Press

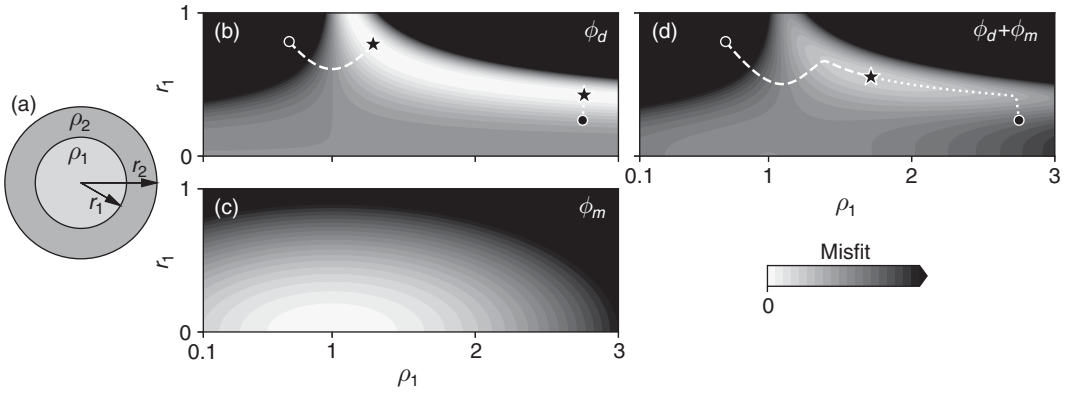


Figure 2.1 Misfit functions for a simple inverse problem (after Valentine and Sambridge, 2020a). (a) A planet is modelled as comprising two spherical layers: a core of radius r_1 and density ρ_1, and an outer unit of density ρ_2 extending to radius r_2. Defining units such that $r_2 = 1$ and $\rho_2 = 1$, we find the overall mass of the planet to be $M = 4.76 \pm 0.25$ units. What can be said about r_1 and ρ_1? (b) The data misfit, $\phi_d(d, g(r_1, \rho_1))$, as in Eq. (2.5), highlighting non-linear behaviours. Two gradient-based optimisation trajectories are shown for different starting points (circles), with convergence to distinct solutions (stars). The inverse problem is inherently non-unique. (c) Penalty term, $\phi_m(r_1, \rho_1)$, expressing a preference for a small core with density similar to that of the surface layer. (d) Combined (regularised) misfit, $\phi_d(d, g(r_1, \rho_1)) + \phi_m(r_1, \rho_1)$. Both optimisation trajectories now converge to the same point.

combined with the 'curse of dimensionality' (Curtis and Lomax, 2001; Fernández-Martínez and Fernández-Muñiz, 2020), rapidly become prohibitive. However, in many cases, it is possible to obtain Fréchet derivatives of the forward problem (Eq. 2.3) with respect to the model, $\delta\mathcal{G}/\delta m$, and this information can be used to guide a search towards the minimum of $\phi(m)$.

2.3.1 Euclidean Data Metrics

Overwhelmingly, the conventional choice for ϕ_d is the squared L_2, or Euclidean, norm of the residuals weighted using the data covariance matrix, $\mathbf{C_d}$,

$$\begin{aligned}\phi_d(\mathbf{d}, \mathbf{g}(m)) &= \left|\left|\mathbf{C_d}^{-\frac{1}{2}}(\mathbf{d} - \mathbf{g}(m))\right|\right|_2^2 \\ &= (\mathbf{d} - \mathbf{g}(m))^{\mathbf{T}}\mathbf{C_d}^{-1}(\mathbf{d} - \mathbf{g}(m)). \end{aligned} \quad (2.5)$$

Relying on the Fréchet derivatives is essentially an assumption that $g(m)$ is (locally) linear. For the usual case, where the model has been discretised as in Eq. (2.2) and can be represented as a vector of coefficients, $\mathbf{m}$, we have $\mathbf{g}(\mathbf{m}) = \mathbf{g}(\mathbf{m_0}) + \mathbf{G}(\mathbf{m} - \mathbf{m_0})$, where $\mathbf{m_0}$ is the linearisation point and $[\mathbf{G}]_{ij} = \partial[\mathbf{g}(\mathbf{m})]_i/\partial m_j|_{\mathbf{m}=\mathbf{m_0}}$. We therefore find

$$\frac{\partial\phi}{\partial\mathbf{m}} = 2\mathbf{G}^{\mathbf{T}}\mathbf{C_d}^{-1}[\mathbf{G}(\mathbf{m} - \mathbf{m_0}) - (\mathbf{d} - \mathbf{g}(\mathbf{m_0}))] + \frac{\partial\phi_m}{\partial\mathbf{m}}. \quad (2.6)$$

This can be used to define an update to the model, following a range of different strategies. Setting $\mathbf{m} = \mathbf{m_0}$, we obtain the gradient of ϕ with respect to each coordinate direction at the point of linearisation: this information may then be used to take a step towards the optimum, using techniques such as conjugate-gradient methods (as in Bozdağ et al., 2016) or the L-BFGS algorithm of Liu and Nocedal (1989), as employed by Lei et al. (2020). Alternatively, we can exploit the fact that, at the optimum, the gradient should be zero: for a suitable choice of ϕ_m, it is possible to solve Eq. (2.6) directly for the $\mathbf{m}$ that should minimise the misfit within the linearised regime. This is 'the' least-squares algorithm, employed by many studies (e.g. Wiggins, 1972; Dziewonski et al., 1981; Woodhouse and Dziewonski, 1984). Few interesting problems are truly linear, and so it is usually necessary to adopt an iterative approach, computing a new linear approximation at each step.

2.3.1.1 Stochastic Algorithms

Since the fundamental task of optimising an objective function is also central to modern machine learning efforts, recent geophysical studies have also sought to exploit advances from that sphere. In particular, methods based on 'stochastic gradient descent' have attracted some attention (e.g. van Herwaarden et al., 2020; Bernal-Romero and Iturrarán-Viveros, 2021). These exploit the intuitive idea that the gradient obtained using all available data can be approximated by a gradient obtained using only a subset of the dataset – and that by using different randomly chosen subsets on successive iterations of gradient descent, one may reach a point close to the overall optimum. In appropriate problems, this can yield a substantial reduction in the overall computational effort expended on gradient calculations. It should be noted that the success of this approach relies on constructing approximate gradients that are, on average, unbiased; as discussed in Valentine and Trampert (2016), approximations that induce systematic errors into the gradient operator will lead to erroneous results.

https://doi.org/10.1017/9781009180412.003 Published online by Cambridge University Press

2.3.2 Sparsity

As has been discussed, we commonly assume that a model can be discretised in terms of some finite set of basis functions. Usually, these are chosen for computational convenience, and inevitably there will be features in the real earth system that cannot be represented within our chosen basis. This leads to the problem of 'spectral leakage' (Trampert and Snieder, 1996): features that are unrepresentable create artefacts within the recovered model.

In digital signal processing, the conditions for complete and accurate recovery of a signal are well known. According to Nyquist's theorem, the signal must be band limited and sampled at a rate at least twice that of the highest frequency component present (Nyquist, 1928). Failure to observe this leads to spurious features in the reconstructed signal, known as aliasing – essentially the same issue as spectral leakage. This has far-reaching consequences, heavily influencing instrument design, data collection, and subsequent processing and analysis.

However, recent work has led to the concept of 'compressed sensing' (Donoho, 2006; Candès and Wakin, 2008). Most real-world signals are, in some sense, sparse: when expanded in terms of an appropriately chosen basis (as per Eq. 2.2), only a few non-zero coefficients are required. If data is collected by random sampling, and in a manner designed to be incoherent with the signal basis, exploiting this sparsity allows the signal to be reconstructed from far fewer observations than Nyquist would suggest. The essential intuition here is that incoherence ensures that each observation is sensitive to many (ideally: all) coefficients within the basis function expansion; the principle of sparsity then allows us to assign the resulting information across the smallest number of coefficients possible.

In theory, imposing sparsity should require us to use a penalty term that counts the number of non-zero model coefficients: $\phi_m(\mathbf{m}) = \alpha^2\|\mathbf{m}\|_0$. However, this does not lead to a tractable computational problem. Instead, Donoho (2006) has shown that it is sufficient to penalise the L_1 norm of the model vector, $\phi_m(\mathbf{m}) = \alpha^2\|\mathbf{m}\|_1 = \alpha^2\sum_i |m_i|$. This can be implemented using a variety of algorithms, including quadratic programming techniques and the Lasso (Tibshirani, 1996). Costs are markedly higher than for L_2 -based penalty functions, but remain tolerable.

Sparsity-promoting algorithms have significant potential: they open up new paradigms for data collection, offering the opportunity to substantially reduce the burden of storing, transmitting, and handling datasets. The success of compressed sensing also suggests that the data misfit $\phi_d(\mathbf{d}, \mathbf{g}(m))$ may be accurately estimated using only a small number of randomly chosen samples: for certain classes of forward model, this may offer a route to substantially reduced computational costs. Again, work is ongoing to explore the variety of ways in which concepts of sparsity can be applied and exploited within the context of geophysical inversion (e.g. Herrmann et al., 2009; Wang et al., 2011; Simons et al., 2011; Bianco and Gerstoft, 2018; Muir and Zhang, 2021).

2.3.3 Non-Euclidean Data Metrics

A common challenge for gradient-based methods is convergence to a local – rather than global – minimum. This situation is difficult to identify or robustly avoid, since doing so would require knowledge of the global behaviour of the forward model. In this context, a particular downside to the use of a Euclidean data norm is that it treats each element of the data vector (i.e. each individual digitised data point) independently. For geophysical datasets, this is often undesirable: the spatial and temporal relationships connecting distinct data points are physically meaningful, and a model that misplaces a data feature (such as a seismic arrival) in time or space is often preferable to one that fails to predict it at all. This problem is particularly familiar in waveform-fitting tasks, where the Euclidean norm is unduly sensitive to any phase differences between data and synthetics. From an optimisation perspective, this can manifest as 'cycle-skipping', where waveforms end up misaligned by one or more complete periods.

As a result, there is interest – and perhaps significant value – in exploring alternative metrics for quantifying the agreement between real and observed data sets. A particular focus of current research is measures built upon the theory of Optimal Transport (e.g. Ambrosio, 2003; Santambrogio, 2015). This focuses on quantifying the 'work' (appropriately defined) required to transform one object into another, and the most efficient path between the two states. In particular, the p-Wasserstein distance between two densities, $f(x)$ and $g(x)$ may be defined

$$W_p(f,g) = \left[\inf_{T\in\mathcal{T}} \int c(x,T(x))^p f(x)\,\mathrm{d}x\right]^{1/p}, \tag{2.7}$$

where $\mathcal{T}$ is the set of all 'transport plans' $T(x)$ that satisfy

$$f(x) = g(T(x))|\nabla T(x)|, \tag{2.8}$$

and $c(x,y)$ is a measure of the distance between points x and y. The resulting metric provides a much more intuitive measure of the difference between two datasets, and perhaps offers a principled route to combining information from multiple distinct data types (sometimes known as 'joint inversion').

Pioneered in geophysics by Engquist and Froese (2014), this has subsequently been employed for numerous studies, including the work of Métivier et al. (2016a,b,c,d) and others (e.g. He et al., 2019; Hedjazian et al., 2019; Huang et al., 2019). However, numerous challenges remain to be fully overcome. Since Optimal Transport is conceived around density functions – which are inherently positive – signed datasets

https://doi.org/10.1017/9781009180412.003 Published online by Cambridge University Press

such as waveforms require special treatment. In addition, since computing the Wasserstein distance between two functions is itself an optimisation problem, there are practical challenges associated with employing it in large-scale inversion problems, and these are the focus of current work.

2.4 Ensemble-Based Methods

We now switch focus, and consider the second fundamental approach to geophysical inversion: instead of seeking a single model that explains the data, we now aim to characterise the collection, or ensemble, of models that are compatible with observations. Clearly, this has potential to be more informative, providing insight into uncertainties and trade-offs; however, it also brings new challenges. Computational costs may be high, and interpretation and decision-making may be complicated without the (illusion of) certainty promised by single-model strategies.

There are many different ways in which one might frame an ensemble-based inversion strategy: at the simplest, one might adapt the grid-search strategy of Section 2.3 so that the 'ensemble' is the set of all grid nodes for which $\phi(m)$ is below some threshold. This approach, with models generated randomly rather than on a grid, underpinned some of the earliest ensemble-based studies in geophysics (e.g. Press, 1970; Anderssen et al., 1972; Worthington et al., 1972). However, it is not particularly convenient from a computational perspective, since such an ensemble has little structure that can be exploited for efficiency or ease of analysis. Techniques exist that seek to address this (e.g. Sambridge, 1998), but the most common strategy is to adopt a probabilistic – and typically Bayesian – perspective. This involves a subtle, but important, change of philosophy: rather than seeking to determine the Earth structure directly, Bayesian inversion aims to quantify our state of knowledge (or 'degree of belief') about that structure (for more discussion, see, e.g., Scales and Snieder, 1997).

The hallmark of Bayesian methods is that the posterior distribution – $\mathcal{P}(m|\mathbf{d})$, the probability of a model m given the observations $\mathbf{d}$ – is obtained by taking the prior distribution ($\mathcal{P}(m)$, our state of knowledge before making any observations), and weighting it by the likelihood, $\mathcal{P}(\mathbf{d}|m)$, which encapsulates the extent to which the data support any given model (see Fig. 2.2a–c). When normalised to give a valid probability distribution, we obtain

$$\mathcal{P}(m|\mathbf{d}) = \frac{\mathcal{P}(\mathbf{d}|m)\mathcal{P}(m)}{\mathcal{P}(\mathbf{d})}, \tag{2.9}$$

which is well known as Bayes' Theorem (Bayes, 1763). We take this opportunity to remark that whereas a misfit function may be chosen in rather ad hoc fashion to exhibit whatever sensitivity is desired, a likelihood has inherent meaning as 'the probability that the observations arose from a given model', and ought to be defined by reference to the expected noise characteristics of the data. We also highlight the work of Allmaras et al. (2013), which provides a comprehensive but accessible account of the practical application of Bayes' Theorem to an experimental inference problem. However, it is usually challenging to employ Eq. (2.9) directly, since evaluating the 'evidence', $\mathcal{P}(\mathbf{d})$, requires an integral over the space of all allowable models, $\mathcal{M}$,

$$\mathcal{P}(\mathbf{d}) = \int_{\mathcal{M}} \mathcal{P}(\mathbf{d}|m)\mathcal{P}(m)\,\mathrm{d}m, \tag{2.10}$$

which is not computationally tractable for arbitrary large-scale problems. Instead, most Bayesian studies either make additional assumptions that enable analytic or semi-analytic evaluation of the evidence, or they exploit the fact that the ratio $\mathcal{P}(m_A|\mathbf{d})/\mathcal{P}(m_B|\mathbf{d})$ can be evaluated without knowledge of the evidence to obtain information about the *relative* probability of different models.

2.4.1 Bayesian Least Squares

The choice of prior is central to the success of any Bayesian approach – and also lies at the heart of many controversies and interpretational challenges, largely due to the impossibility of representing the state of no information (e.g. Backus, 1988). It is therefore apparent that within a Bayesian framework all inference is considered relative to a known prior. In principle, the prior should be chosen based on a careful consideration of what is known about the problem of interest; in practice, this is often tempered by computational pragmatism, and a distribution with useful analytic properties is adopted.

2.4.1.1 Gaussian Process Priors

A convenient choice when dealing with an unknown model function $m(\mathbf{x}, \mathbf{t})$ is a Gaussian Process prior,

$$m(\mathbf{x}, t) \sim \mathcal{GP}\Big(\mu(\mathbf{x}, t), k(\mathbf{x}, t, \mathbf{x}', t')\Big). \tag{2.11}$$

This is essentially the extension of the familiar normal distribution into function space, with our knowledge at any given point, $(\mathbf{x}, t)$, quantified by a mean $\mu(\mathbf{x}, t)$ and standard deviation $k(\mathbf{x}, t; \mathbf{x}, t)^{1/2}$; however, the covariance function k also quantifies our knowledge (or assumptions) about the expected covariances if m were to be measured at multiple distinct points. A comprehensive introduction to the theory of Gaussian Processes may be found in, for example, Rasmussen and Williams (2006).

In some geophysical problems, the data–model relationship is – or can usefully be approximated as – linear (see also Section 2.3.1), and so can be expressed in the form

https://doi.org/10.1017/9781009180412.003 Published online by Cambridge University Press

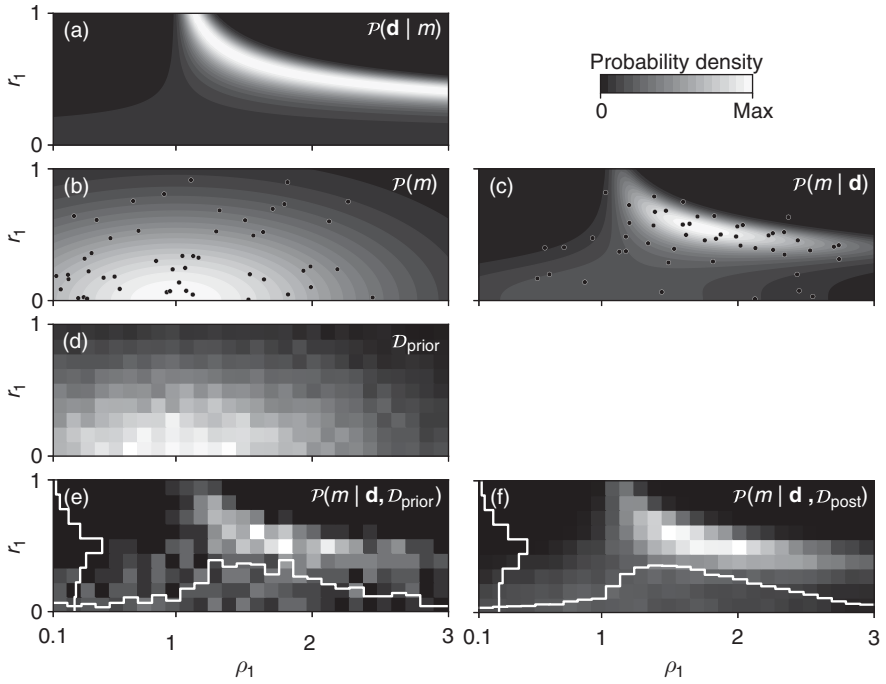


Figure 2.2 Bayesian analysis for the simple inverse problem introduced in Fig. 2.1. (a) The likelihood, $\mathcal{P}(\mathbf{d}|m)$, quantifies the extent to which any given choice of model can explain the data. (b) The prior distribution, $\mathcal{P}(m)$, encapsulates our beliefs before observing any data, and can be 'sampled' to generate a collection of candidate models ($\mathcal{D}_{\text{prior}}$; dots; 50 shown). (c) The posterior distribution, $\mathcal{P}(m|\mathbf{d})$ combines prior and likelihood (Eq. 2.9) to encapsulate our state of knowledge *after* taking account of the data. In realistic problems, visualising the posterior is intractable, but we can generate samples from it ($\mathcal{D}_{\text{post}}$; 50 shown). (d) We can evaluate the forward model $g(m)$ for each example within an ensemble of prior samples, and additionally simulate the effects of noise processes. This can be completed without reference to any data. The information can be stored in many forms, including as a machine learning model. (e) Once data becomes available, this information can be queried to identify regions of parameter space that may explain observations; see Section 2.4.2. This provides an approximation to the posterior; we additionally show 1-D marginals for each model parameter. (f) A similarly sized set of posterior samples provides a much better approximation to the true posterior, as it is targeted towards explaining one specific set of observations: see Section 2.4.3. However, computational costs may be prohibitive for some applications.

$$d_i = \int_0^T \int_{\mathcal{X}} q_i(\mathbf{x}, t)\, m(\mathbf{x}, t)\, \mathrm{d}\mathbf{x}\, \mathrm{d}t, \tag{2.12}$$

where $q_i(\mathbf{x}, t)$ is some 'data kernel', and where $\mathcal{X}$ represents the domain upon which the model is defined. Moreover, we assume that the noise process represented by $\mathbf{C_d}$ is explicitly Gaussian. These assumptions permit analytic evaluation of the evidence, and the posterior distribution can be written in the form (Valentine and Sambridge, 2020a)

$$\widetilde{m}(\mathbf{x}, t) \sim \mathcal{GP}\left(\widetilde{\mu}(\mathbf{x}, t), \widetilde{k}(\mathbf{x}, t; \mathbf{x}', t')\right), \tag{2.13}$$

where we use a tilde to denote a posterior quantity, and where

$$\widetilde{\mu}(\mathbf{x}, t) = \mu(\mathbf{x}, t) + \sum_{ij} w_i(\mathbf{x}, t) \left[(\mathbf{W} + \mathbf{C_d})^{-1}\right]_{ij} (d_j - \omega_j), \tag{2.14}$$

$$\widetilde{k}(\mathbf{x}, t; \mathbf{x}', t') = k(\mathbf{x}, t; \mathbf{x}', t') - \sum_{ij} w_i(\mathbf{x}, t) \left[(\mathbf{W} + \mathbf{C_d})^{-1}\right]_{ij} w_j(\mathbf{x}', t') \tag{2.15}$$

with

$$w_i(\mathbf{x}, t) = \int_0^T \int_{\mathcal{X}} k(\mathbf{x}, t; \mathbf{x}', t') q_i(\mathbf{x}', t') \mathrm{d}\mathbf{x}'\, \mathrm{d}t', \tag{2.16}$$

$$W_{ij} = \int_0^T \int_0^T \iint_{\mathcal{X}^2} q_i(\mathbf{x}, t)\, k\,(\mathbf{x}, t; \mathbf{x}', t') q_j(\mathbf{x}', t')\ \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{x}'\ \mathrm{d}t\, \mathrm{d}t', \tag{2.17}$$

$$\omega_i = \int_0^T \int_{\mathcal{X}} \mu(\mathbf{x}, t) q_i(\mathbf{x}, t) \mathrm{d}\mathbf{x}\ \mathrm{d}t. \tag{2.18}$$

https://doi.org/10.1017/9781009180412.003 Published online by Cambridge University Press

This approach has formed the basis for a variety of geophysical studies (e.g. Tarantola and Nercessian, 1984; Montagner and Tanimoto, 1990, 1991; Valentine and Davies, 2020) and has the attractive property that the inference problem is posed directly in a function space, avoiding some of the difficulties associated with discretisation (such as spectral leakage).

2.4.1.2 Discretised Form

Nevertheless, if one chooses to introduce a finite set of basis functions, as in Eq. (2.2), it is possible to express Eqs. (2.11–2.18) in discretised form (for full discussion, see Valentine and Sambridge, 2020b). The prior distribution on the expansion coefficients becomes

$$\mathbf{m} \sim \mathcal{N}(\mathbf{m_p}, \mathbf{C_m}) \tag{2.19}$$

and the linear data–model relationship is expressed in the form $\mathbf{g}(\mathbf{m}) = \mathbf{Gm}$. The posterior distribution may be written in a variety of forms, including

$$\mathbf{m} \sim \mathcal{N}(\widetilde{\mathbf{m}}, \widetilde{\mathbf{C}}_\mathbf{m}), \tag{2.20}$$

where

$$\widetilde{\mathbf{m}} = \mathbf{m_p} + (\mathbf{G^T C_d^{-1} G} + \mathbf{C_m^{-1}})^{-1} \mathbf{G^T C_d^{-1}} (\mathbf{d} - \mathbf{G m_p}) \tag{2.21}$$

$$\widetilde{\mathbf{C}}_\mathbf{m} = (\mathbf{G^T C_d^{-1} G} + \mathbf{C_m^{-1}})^{-1}. \tag{2.22}$$

This well-known result, found in Tarantola and Valette (1982), has formed the basis of much work in geophysics. The expression for $\widetilde{\mathbf{m}}$ is also often applied in non-Bayesian guise – compare with the discussion in Section 2.3.1 – with the prior covariance matrix $\mathbf{C_m}$ regarded as a generic 'regularisation matrix' without probabilistic interpretation.

2.4.2 Prior Sampling

The results of Section 2.4.1 are built upon assumptions that our prior knowledge is Gaussian and the forward model is linear. This is computationally convenient, but will rarely be an accurate representation of the true state of affairs. Unfortunately, more general assumptions tend not to support analytic expressions for the posterior, and hence it becomes necessary to adopt 'sampling-based methods'. These rely on evaluating the forward problem for a large number of models, in order to accumulate information about the relationship between model and data. Various strategies exist, which can be characterised by the manner in which sampling is performed.

The first group of strategies are those where candidate models are generated according to the prior distribution, and predicted data (potentially including simulated 'noise') is computed for each. This provides a set of samples

$$\mathcal{D}_{\text{prior}} = \{(\mathbf{m}_i, \mathbf{g}(\mathbf{m}_i)), \qquad i = 1, \ldots, N\}, \tag{2.23}$$

which may then be interpolated as necessary to address inversion questions (see Fig. 2.2d–e). This family of approaches is known as 'prior sampling' (Käufl et al., 2016a), with different examples characterised by differing approaches to interpolation. Many of the recent studies that exploit machine learning to perform inversion may be seen within the prior sampling framework, although not all are explicitly Bayesian in design.

2.4.2.1 Mixture Density Networks

If we *do* take a Bayesian approach, then we may note that the density of samples within $\mathcal{D}_{\text{prior}}$ approximates – by construction – the joint probability density, $\mathcal{P}(\mathbf{m}, \mathbf{d})$. If we can fit an appropriate parametric density function to the samples, it is then straightforward to interpolate to obtain the conditional density $\mathcal{P}(\mathbf{m}|\mathbf{d})$ corresponding to observations (which we recognise to be the posterior distribution). One currently popular way to achieve this is to employ Mixture Density Networks (MDNs; Bishop, 1995), which involve an assumption that the conditional distribution can be written as a Gaussian Mixture Model (GMM),

$$\mathcal{P}(m_j|\mathbf{d}) \approx \sum_{k=1}^{K} \frac{w_k(\mathbf{d})}{\sqrt{2\pi\sigma_k^2(\mathbf{d})}} \exp\left(-\frac{\left(m_j - \mu_k(\mathbf{d})\right)^2}{2\sigma_k^2(\mathbf{d})} \right), \tag{2.24}$$

where the weights w_k (which are subject to an additional constraint, $\Sigma_k w_k = 1$), means μ_k, and standard deviations σ_k that define the GMM are assumed to be functions of the data. These relationships may in turn be represented by a neural network. The set of prior samples, $\mathcal{D}_{\text{prior}}$, is then used to optimise the neural network parameters, such that the expected value

$$\mathbb{E}_{\mathcal{D}_{\text{prior}}}\{\mathcal{P}(m_j|\mathbf{d})\} = \frac{1}{N} \sum_{i=1}^{N} \mathcal{P}\Big([\mathbf{m_i}]_j | \mathbf{g}(\mathbf{m_i}) \Big) \tag{2.25}$$

is maximised. This approach has been applied to a variety of geophysical problems, including structural studies at global (e.g. Meier et al., 2007; de Wit et al., 2014) and local (e.g. Earp et al., 2020; Mosher et al., 2021) scales, seismic source characterisation (Käufl et al., 2014), and mineral physics (e.g. Rijal et al., 2021).

2.4.2.2 Challenges and Opportunities

The principal downside to prior sampling – discussed in detail by Käufl et al. (2016a) in the context of MDNs, but applicable more broadly – is the fact that only a few of the samples within $\mathcal{D}_{\text{prior}}$ will provide useful information about any given set of observations. In realistic problems, the range of models encompassed by the prior is large in comparison to the range encompassed by the posterior, and much computational effort is expended on generating predictions that turn out to have little similarity to

https://doi.org/10.1017/9781009180412.003 Published online by Cambridge University Press

observations. This is exacerbated by issues associated with the 'curse of dimensionality', motivating the common choice (implicit in our notation for Eq. 2.24) to use prior sampling to infer low or uni-dimensional marginal distributions rather than the full posterior. Overall, the consequence is that prior sampling tends to yield rather broad posteriors, representing 'our state of knowledge in the light of the simulations we have performed', rather than 'the most we can hope to learn from the available data'. We also emphasise that results are wholly dependent on the choice of prior, and will be meaningless if this does not encompass the real earth system. This is perhaps obvious in an explicitly Bayesian context, but may be lost when studies are framed primarily from the perspective of machine learning.

The great benefit of prior sampling is that nearly all of the computational costs are incurred *before* any knowledge of observed data is required. As a result, it may be effective in situations where it is desirable to obtain results as rapidly or cheaply as possible following data collection – for example, to enable expensive numerical wave propagation simulations to be employed for earthquake early warning (Käufl et al., 2016b). We note parallels here to the use of scenario-matching approaches in the field of tsunami early warning (e.g. Steinmetz et al., 2010). It is also well-suited to applications where the same fundamental inverse problem must be solved many times for distinct datasets, perhaps representing observations repeated over time, or at many localities throughout a region.

Prior sampling may also be effective in settings requiring what we term 'indirect' inference, where the primary goal is to understand some quantity derived from the model, rather than the model itself. For example, in an earthquake early warning setting, one might seek to determine seismic source information with a view to then using this to predict tsunami run-up, or the peak ground acceleration at critical infrastructure sites (Käufl, 2015). In a prior sampling setting, one may augment $\mathcal{D}_{\text{prior}}$ to incorporate a diverse suite of predictions, $\mathcal{D}_{\text{prior}} = \{(\mathbf{m}_i, \mathbf{g_1}(\mathbf{m}_i), \mathbf{g_2}(\mathbf{m}_i), \ldots),\ i = 1, \ldots, N\}$, and then employ some interpolation framework to use observations of the process associated with (say) $\mathbf{g_1}$ to make inferences about $\mathbf{g_2}$. From a Bayesian perspective, this can be seen as a process of marginalisation over the model parameters themselves.

2.4.3 Posterior Sampling

As an alternative to prior sampling, one may set out to generate a suite of samples, $\mathcal{D}_{\text{post}}$, distributed according to the posterior (see Fig. 2.2f). Again, there are a variety of ways this can be achieved – for example, a simple (but inefficient) approach might involve rejection sampling. More commonly, Markov chain Monte Carlo (McMC) methods are employed, with the posterior forming the equilibrium distribution of a random walk. Encompassed within the term McMC lie a broad swathe of algorithms, of which the Metropolis–Hastings is probably most familiar, and the field is continually the subject of much development. We do not attempt to survey these advances, but instead direct the reader to one of the many recent reviews or tutorials on the topic (e.g. Brooks et al., 2011; Hogg and Foreman-Mackey, 2018; Luengo et al., 2020).

As set out in Käufl et al. (2016a), prior and posterior sampling procedures generate identical results in the theoretical limit. However, in practical settings they are suited to different classes of problems. Posterior sampling approaches are directed towards explaining a specific dataset: this allows computational resources to be targeted towards learning the specifics of the problem at hand, but prevents expensive simulations from being 'recycled' in conjunction with other datasets. It should also be noted that the 'solution' obtained via posterior sampling takes the form of an ensemble of discrete samples. This can be challenging to store, represent, and interrogate in a meaningful way: many studies resort to reducing the ensemble to a single maximum-likelihood or mean model, and perhaps some statistics about the (co)variances associated with different parameters, and thereby neglect much of the power of McMC methods. Effective solutions to this issue may be somewhat problem dependent, but remain the focus of much work.

2.4.3.1 Improving Acceptance Ratios

Generation of ensembles of posterior samples is inherently wasteful: by definition, one does not know in advance where samples should be placed, and hence for every 'useful' sample, a large numbers of candidate models must be tested (i.e. we must evaluate the forward problem) and rejected. This is exacerbated by requirements for 'burn-in' (so that the chain is independent of the arbitrary starting point) and 'chain thinning' (to reduce correlations between consecutive samples), which also cause substantial numbers of samples to be discarded. Much effort is therefore expended on developing a variety of strategies to improve 'acceptance ratios' (i.e. the proportion of all tested models that end up retained within the final ensemble).

One route forward involves improving the 'proposal distribution', that is, the manner in which samples are generated for testing. Ideally, we wish to make the proposal distribution as close as possible to the posterior, so that nearly all samples may be retained. Of course, the difficulty in doing so is that the posterior is not known in advance. An avenue currently attracting considerable interest is Hamiltonian McMC (HMC) methods, which exploit analogies with Hamiltonian dynamics to guide the random walk process towards 'acceptable' samples (see, e.g., Neal, 2011; Betancourt, 2017). In order to do so, HMC methods require, and exploit, knowledge of the gradient of the likelihood with respect to the model parameters at each sampling point. This provides additional information about the underlying physical problem, enabling extrapolation away

https://doi.org/10.1017/9781009180412.003 Published online by Cambridge University Press

from the sample point, and the identification of 'useful' directions for exploration. To apply this idea, we must be able to compute the required gradients efficiently; early applications in geophysics have included seismic exploration and full-waveform inversion (e.g. Sen and Biswas, 2017; Fichtner et al., 2019; Aleardi and Salusti, 2020).

In many cases, the fundamental physical problem of Eq. (2.1) is amenable to implementation (Eq. 2.3) in a variety of ways, depending on the assumptions made ($\mathcal{B}$). Usually, simplified assumptions lead to implementations with lower computational costs, at the expense of introducing systematic biases into predictions. Recently, Koshkholgh et al. (2021) has exploited this to accelerate McMC sampling, by using a low-cost physical approximation to help define a proposal distribution. Likelihood evaluations continue to rely on a more complex physical model, so that accuracy is preserved within the solution to the inverse problem – but the physically motivated proposal distribution improves the acceptance rate and reduces overall computational costs. This is an attractive strategy, and seems likely to underpin future theoretical developments.

2.4.3.2 Transdimensional Inference

In practice, McMC studies typically assume a discretised model, expressed relative to some set of basis functions in as in Eq. (2.2), and the choice of basis functions is influential in determining the characteristics of the solution. In particular, the number of terms in the basis function expansion typically governs the flexibility of the solution, and the scale-lengths that can be represented. However, it also governs the dimension of the search space: as the number of free parameters in the model grows, so does the complexity (and hence computational cost) of the Monte Carlo procedure. Transdimensional approaches arise as an attempt to strike a balance between these two competing considerations: both basis set and expansion coefficients are allowed to evolve during the random walk process (Green, 1995; Sambridge et al., 2006; Bodin and Sambridge, 2009; Sambridge et al., 2012).

The transdimensional idea has been applied to a wide variety of geoscience problems, including source (e.g. Dettmer et al., 2014) and structural (Burdick and Lekić, 2017; Galetti et al., 2017; Guo et al., 2020) studies using seismic data, in geomagnetism (Livermore et al., 2018), and in hydrology (Enemark et al., 2019). It can be particularly effective in settings where basis functions form a natural hierarchy of scale lengths, such as with wavelets and spherical harmonics, although keeping track of information creates computational challenges (Hawkins and Sambridge, 2015). We note that model complexity is not confined only to length-scales: one can also employ a transdimensional approach to the physical theory, perhaps to assess whether mechanisms such as anisotropy are truly mandated by available data. The approach can also be employed to identify change-points or discontinuities within a function (e.g. Gallagher et al., 2011), and used in combination with other techniques such as Gaussian Processes (Ray and Myer, 2019; Ray, 2021).

2.4.4 Variational Methods

One of the drawbacks of posterior sampling is the fact that the sampling procedure must achieve two purposes: it not only 'discovers' the form of the posterior distribution, but also acts as our mechanism for representing the solution (which takes the form of a collection of appropriately distributed samples). Large numbers of samples are often required to ensure stable statistics and 'convincing' figures, even if the underlying problem itself is rather simple. To address this, one may introduce a parametric representation of the posterior distribution, and frame the inference task as determination of the optimal values for the free parameters – much as with the Mixture Density Network (Section 2.4.2.1). This approach, often known as Variational Inference (e.g. Blei et al., 2017), transforms inference for ensembles into an optimisation problem, and offers potentially large efficiency gains.

We sketch the basic concept here, noting that a galaxy of subtly different strategies can be found in recent literature (see, e.g. Zhang et al., 2019, for a review). As usual, our goal is to determine the posterior distribution, $\mathcal{P}(m \mid \mathbf{d})$. To approximate this, we introduce a distribution function $\mathcal{Q}(m \mid \boldsymbol{\theta})$, that has known form, parameterised by some set of variables $\boldsymbol{\theta}$ – for example, we might decide that Q should be a Gaussian mixture model, in which case $\boldsymbol{\theta}$ would encapsulate the weights, means, and variances for each mixture component. Our basic goal is then to optimise the parameters $\boldsymbol{\theta}$ such that $\mathcal{Q}(m \mid \boldsymbol{\theta}) \approx \mathcal{P}(m \mid \mathbf{d})$.

To make this meaningful, we must – much as in Section 2.3 – first define some measure of the difference between the two distributions. In Variational Inference, the usual choice is the Kullback–Leibler divergence (Kullback and Leibler, 1951),

$$D_{\mathrm{KL}}(\mathcal{Q}\|\mathcal{P}) = \int \mathcal{Q}(m \mid \boldsymbol{\theta}) \log \frac{\mathcal{Q}(m \mid \boldsymbol{\theta})}{\mathcal{P}(m \mid \mathbf{d})} \mathrm{d}m$$
$$= \mathbb{E}_{\mathcal{Q}(m \mid \boldsymbol{\theta})} \left\{ \log \frac{\mathcal{Q}(m \mid \boldsymbol{\theta})}{\mathcal{P}(m \mid \mathbf{d})} \right\}, \tag{2.26}$$

where the notation $\mathbb{E}_{\mathcal{Q}(m)}\{f(m)\}$ signifies 'the expected value of $f(m)$ when m m is distributed according to Q'. Exploiting the properties of logarithms, and applying Bayes' Theorem, we can rewrite this in the form

$$D_{\mathrm{KL}}(Q\|\mathcal{P}) = \log \mathcal{P}(\mathbf{d}) + \mathbb{E}_{\mathcal{Q}(m \mid \boldsymbol{\theta})}\{\log Q(m|\boldsymbol{\theta}) - \log \mathcal{P}(\mathbf{d}|m) - \log \mathcal{P}(m)\}, \tag{2.27}$$

where $\mathcal{P}(\mathbf{d})$ has been moved outside the expectation since it is independent of m. While this quantity is unknown, it is

https://doi.org/10.1017/9781009180412.003 Published online by Cambridge University Press

also constant – and so can be neglected from the perspective of determining the value of $\boldsymbol{\theta}$ at which D_{KL} is minimised. The quantity $\mathcal{P}(\mathbf{d}) - D_{\mathrm{KL}}(Q\|\mathcal{P})$ is known as the 'evidence lower bound' (ELBO), and maximisation of this is equivalent to minimising the Kullback–Leibler divergence. Because the variational family $Q(m\,|\,\boldsymbol{\theta})$ has a known form, the ELBO can be evaluated, as can the derivatives $\partial D_{\mathrm{KL}}/\partial\theta_i$. Thus, it is conceptually straightforward to apply any gradient-based optimisation scheme to determine the parameters such that Q best approximates the posterior distribution.

2.4.4.1 A Gaussian Approximation

To illustrate this procedure, and to highlight connections to other approaches, we consider an inverse problem where: (i) the model is discretised, as in Eq. (2.2), so that we seek an M-component model vector $\mathbf{m}$; (ii) the prior distribution on those model coefficients is Gaussian with mean $\mathbf{m_p}$ and covariance $\mathbf{C_m}$; and (iii) the likelihood takes the form $\mathcal{P}(\mathbf{m}\,|\mathbf{d}) = k\,\exp(-\frac{1}{2}\phi(\mathbf{m}))$ for some appropriate function ϕ. We choose to assert that the solution can be approximated by a Gaussian of mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, and seek the optimal values of these quantities. Thus, we choose

$$Q(\mathbf{m}\,|\,\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{M/2}(\det\boldsymbol{\Sigma})^{1/2}}\exp\{-(\mathbf{m}-\boldsymbol{\mu})^{\mathbf{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}(\mathbf{m}-\boldsymbol{\mu})\}. \tag{2.28}$$

To proceed, we need to determine the expectation of various functions of $\mathbf{m}$ under this distribution – and their gradients with respect to $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$.

A number of useful analytical results and expressions can be found in Petersen and Pedersen (2012). It is straightforward to determine that

$$\frac{\partial}{\partial\boldsymbol{\mu}}D_{\mathrm{KL}}(Q\|\mathcal{P}) = \mathbf{C_m^{-1}}(\boldsymbol{\mu}-\mathbf{m_p}) + \frac{1}{2}\frac{\partial}{\partial\boldsymbol{\mu}}\mathbb{E}_Q\{\phi(\mathbf{m})\}, \tag{2.29}$$

$$\frac{\partial}{\partial\boldsymbol{\Sigma}}D_{\mathrm{KL}}(Q\|\mathcal{P}) = -\frac{1}{2}(\boldsymbol{\Sigma^{-1}} - \mathbf{C_m^{-1}}) + \frac{1}{2}\frac{\partial}{\partial\boldsymbol{\Sigma}}\mathbb{E}_Q\{\phi(\mathbf{m})\}. \tag{2.30}$$

These expressions can be used to drive an iterative optimisation procedure to determine the optimal variational parameters. In implementing this, the result

$$\frac{\partial}{\partial\theta_i}\mathbb{E}_{Q(m\,|\,\boldsymbol{\theta})}\{f[m]\} = \mathbb{E}_{Q(m\,|\,\boldsymbol{\theta})}\left\{f(m)\frac{\partial}{\partial\theta_i}\log Q(m\,|\,\boldsymbol{\theta})\right\} \tag{2.31}$$

may be useful.

In the case where $\mathbf{g}(\mathbf{m})$ is (or is assumed to be) linear, and where the function ϕ is defined as the L_2 norm of the residuals, the expected value can be evaluated analytically. The misfit is quadratic in form,

$$\phi(\mathbf{m}) = \mathbf{d^T C_d^{-1} d} - 2\mathbf{d^T C_d^{-1} G m} + \mathbf{m^T G^T C_d^{-1} G m}, \tag{2.32}$$

and the Gaussian Q, can be determined, along with its derivatives, as in Section 2.4.1.1. Hence the expected value, given m is distributed according to the Gaussian Q, can be determined, along with its derivatives

$$\mathbb{E}_Q\{\phi(\mathbf{m})\} = -2\mathbf{d^T C_d^{-1} G}\boldsymbol{\mu} + \mathrm{Tr}(\mathbf{G^T C_d^{-1} G}\boldsymbol{\Sigma}) + \boldsymbol{\mu}^{\mathbf{T}}\mathbf{G^T C_d^{-1} G}\boldsymbol{\mu} \tag{2.33}$$

$$\frac{\partial}{\partial\boldsymbol{\mu}}\mathbb{E}_Q\{\phi(\mathbf{m})\} = -2\mathbf{G^T C_d^{-1} d} + 2\mathbf{G^T C_d^{-1} G}\boldsymbol{\mu}, \tag{2.34}$$

$$\frac{\partial}{\partial\boldsymbol{\Sigma}}\mathbb{E}_Q\{\phi(\mathbf{m})\} = \mathbf{G^T C_d^{-1} G}. \tag{2.35}$$

Substituting these expressions into Eqs. (2.29–2.30), and solving for the $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ such that the gradients of D_{KL} are zero (as is required at a minimum), we find that the optimal distribution Q is identical to the posterior distribution obtained in Eq. (2.20). This is unsurprising, since our underlying assumptions are also identical, but demonstrates the self-consistency of, and connections between, the different approaches. Again, we also highlight the similarity with the expressions obtained in Section 2.3.1, although the underlying philosophy differs.

2.4.4.2 Geophysical Applications

Variational methods offer a promising route to flexible but tractable inference. As the preceding example illustrates, they provide opportunities to balance the (assumed) complexity and expressivity of the solution against computational costs. A number of recent studies have therefore explicitly sought to explore their potential in particular applications, including for earthquake hypocentre determination (Smith et al., 2022), seismic tomography (Zhang and Curtis, 2020; Siahkoohi and Herrman, 2021; Zhao et al., 2022), and hydrogeology (Ramgraber et al., 2021). However, given the fairly broad ambit of variational inference, many past studies could also be seen as falling under this umbrella.

2.4.5 Generative Models

Many of the methods discussed so far rely on strong assumptions about the form of prior and/or posterior distributions: we suppose that these belong to some relatively simple family, with properties that we can then exploit for efficient calculations. However, such assertions are typically justified by their convenience – perhaps aided by an appeal to the principle known as Occam's Razor – and not through any fundamental physical reasoning (see, e.g., Constable

https://doi.org/10.1017/9781009180412.003 Published online by Cambridge University Press

et al., 1987). This is unsatisfactory, and may contribute substantial unquantifiable errors into solutions and their associated uncertainty estimates.

Recently, a number of techniques have emerged that allow representation of, and computation with, relatively general probability distributions. In broad terms, these are built upon the idea that arbitrarily complex probability distributions can be constructed via transformations of simpler distributions. This is familiar territory: whenever we need to generate normally distributed random numbers, a technique such as the Box–Muller transform (Box and Muller, 1958) is applied to the uniformly distributed output of a pseudorandom number generator. However, the versatility of such approaches is vastly increased in conjunction with the tools and techniques of modern machine learning.

This is an area that is currently the focus of rapid development; recent reviews include those of Bond-Taylor et al. (2022) and Ruthotto and Haber (2021). Clearly, the concept is closely connected to the idea of variational inference, as discussed in Section 2.4.4. Several major techniques have emerged, including 'generative adversarial networks' (GANs) (e.g. Goodfellow et al., 2014; Creswell et al., 2018), 'variational autoencoders' (Kingma and Welling, 2014), and 'normalising flows' (Rezende and Mohamed, 2015; Kobyzev et al., 2021). A variety of recent studies have explored diverse applications of these concepts within the context of geophysical inversion: examples include Mosser et al. (2020), Lopez-Alvis et al. (2021), Scheiter et al. (2022), and Zhao et al. (2022). We have no doubt that this area will lead to influential developments, although the precise scope of these is not yet clear.

2.5 Model Properties

The third fundamental approach builds on the work of Backus and Gilbert (1968) and Backus (1970a,b,c), and we sketch it briefly for completeness. For certain classes of problem, as in Eq. (2.12), each of the observables d_i can be regarded as representing an average of the model function weighted by some data kernel $q_i(\mathbf{x}, t)$. It is then straightforward to write down a weighted sum of the observations,

$$D_{\boldsymbol{\alpha}} = \sum_i \alpha_i d_i = \int_0^T \int_{\mathcal{X}} Q(\boldsymbol{\alpha}, \mathbf{x}, t) m(\mathbf{x}, t) \mathrm{d}\mathbf{x}\, \mathrm{d}t, \tag{2.36}$$

where $Q(\boldsymbol{\alpha}, \mathbf{x}, t) = \sum_i \boldsymbol{\alpha}_i q_i(\mathbf{x}, t)$, and $\boldsymbol{\alpha} = (\alpha_1, \alpha_2 \ldots)$ represents some set of tunable weights. By adjusting these, one may vary the form of the averaging kernel Q, and frame a functional optimisation problem to determine the $\boldsymbol{\alpha}$ that brings Q as close as possible to some desired form. In this way, the value of the average that is sought can be estimated as a linear combination of the observed data.

Backus–Gilbert theory has an inherent honesty: it is data led, with a focus on understanding what the available data can – or cannot – constrain within the system. On the other hand, this can be seen as a downside: it is not usually possible to use the results of a Backus–Gilbert style analysis as the foundation for further simulations. Moreover, interpretation can be challenging in large-scale applications, as the 'meaning' of each result must be considered in the light of the particular averaging kernel found. Perhaps for this reason – and because it is designed for strictly linear problems (although we note the work of Snieder, 1991) – the method is well known but has found comparatively little use. Notable early examples include Green (1975), Chou and Booker (1979), and Tanimoto (1985, 1986). More recently, it has been adopted by the helioseismology community (Pijpers and Thompson, 1992), and applied to global tomography (Zaroli, 2016) and to constrain mantle discontinuities (Lau and Romanowicz, 2021). Concepts from Backus–Gilbert theory are also sometimes used to support interpretation of models produced using other approaches: for example, Ritsema et al. (1999) presents Backus–Gilbert kernels to illustrate the resolution of a model obtained by least-squares inversion.

2.6 Miscellanea

The preceding sections have focused on the range of different philosophies, and associated techniques, by which geophysical inversion can be framed. We now turn to consider some additional concepts and developments that are not themselves designed to solve inverse problems, but which can potentially be employed in conjunction with one or other of the approaches described in this chapter.

2.6.1 Approximate Forward Models

One of the major limiting factors in any geophysical inversion is computational cost. High-fidelity numerical models tend to be computationally expensive, and costs may reach hundreds or even thousands of CPU-hours per simulation. In such cases, resource availability may severely constrain the number of simulations that may be performed, rendering certain approaches infeasible. There is therefore considerable potential value in any technique that may lower the burden of simulation.

2.6.1.1 Surrogate Modelling

One possible solution to this lies in 'surrogate modelling': using techniques of machine learning to mimic the behaviour of an expensive forward model, but at much lower computational cost. This is an idea that has its origins in engineering design (see, e.g., Quiepo et al., 2005; Forrester

https://doi.org/10.1017/9781009180412.003 Published online by Cambridge University Press

and Keane, 2009), and typically involves tuning the free parameters of a neural network or other approximator to match a database of examples obtained via expensive computations (or, indeed, physical experiments). The approximate function can then be interrogated to provide insights, or to serve as a drop-in replacement for the numerical code.

Although the term 'surrogate modelling' only appears relatively recently in the geophysics literature, the underlying idea has a long history. For example, seismologists have long recognised that travel times of seismic arrivals from known sources can be interpolated, and the resulting travel-time curves used to assist in the location of new events (e.g. Jeffreys and Bullen, 1940; Kennett and Engdahl, 1991; Nicholson et al., 2004). One may also regard the Neighbourhood Algorithm (Sambridge, 1999a,b) within this framework: it uses computational geometry to assemble a surrogate approximation to evaluation of (typically) the likelihood for any given model. By employing and refining this within a Markov chain, it is possible to substantially reduce the computational costs of McMC-based inference. In doing so, we exploit the fact that the mapping from models to likelihood (a scalar quantity) is, typically, much simpler than the mapping from models to data. Closely related is the field of 'Bayesian optimisation', which relies on a surrogate (often a Gaussian Process) to encapsulate incomplete knowledge of an objective function, and takes this uncertainty into account within the optimisation procedure (e.g. Shahriari et al., 2016; Wang et al., 2016).

Latterly, surrogate models (also known as emulators) have been explicitly adopted for geophysical studies. Similar to the Neighbourhood Algorithm, Chandra et al. (2020) employed a neural network-based surrogate to replace likelihood calculations within a landscape evolution model; on the other hand, Das et al. (2018) and Spurio Mancini et al. (2021) both developed a surrogate that directly replaces a forward model and outputs synthetic seismograms. Other geophysical examples include modelling of climate and weather (e.g. Field et al., 2011; Castruccio et al., 2014), and applications in hydrology (Hussain et al., 2015) and planetary geophysics (Agarwal et al., 2020).

2.6.1.2 Physics-Informed Neural Networks

A number of recent studies have also explored the concept and applications of 'physics informed neural networks' (PINNs; see, e.g., Raissi et al., 2019; Karniadakis et al., 2021). As with surrogate models, these exploit machine learning techniques to provide a version of the forward model that has significantly lower computational cost than 'conventional' implementations. However, whereas a surrogate is constructed using a suite of examples obtained by running the conventional model (at substantial expense), a PINN is directly trained to satisfy the physical constraints. Typically, this amounts to defining a neural network to represent the observable function, $u(\mathbf{x}, t)$, and then employing a training procedure to minimise the deviation from Eq. (2.1). This is potentially a more efficient approach, and provides the researcher with greater oversight of the behaviour and limitations of the learned model.

A number of recent examples may be found, particularly in the seismological literature. Moseley et al. (2020), Song et al. (2021), and Smith et al. (2020) all use PINNs to solve problems related to the wave equation, with the latter underpinning the variational inference approach of Smith et al. (2022). A range of potential applications in climate science and meteorology are discussed in Kashinath et al. (2021), while He and Tartakovsky (2021) consider hydrological problems. Again, it is clear that PINNs present a promising opportunity that is likely to bring substantial benefits for geophysics, but it is not yet clear how the field will evolve.

2.6.1.3 Conventional Approximations

Surrogate models and PINNs both rely on machine learning, and their 'approximate' nature arises from this: they are constructed to give good average performance for a particular task, but there are few hard constraints on their accuracy in any specific case. In many geophysical problems, an alternate route exists, and has long been exploited: rather than seeking an approximate solution to a complex physical problem, we can use conventional methods to obtain an accurate solution for a simplified physical system (i.e. adopting a more restrictive set of assumptions, $\mathcal{A} \cup \mathcal{B}$). Thus, for example, seismic waves might be modelled under the assumption that propagation is only affected by structure in the great-circle plane between source and receiver (Woodhouse and Dziewonski, 1984) at far lower cost than (almost) physically complete simulation (e.g. Komatitsch et al., 2002). Depending on circumstances, it may be beneficial to exploit a known approximation of this kind, where impacts can be understood and interpretations adjusted accordingly. We also highlight that it may be desirable to vary the level of approximation used for forward simulations within an inversion framework, using a fast approximate technique for initial characterisation, and increasing accuracy as solutions are approached. In the ideal case, one might envisage a forward model where the level of approximation is itself a tuneable parameter (e.g. via the coupling band-width in a normal-mode–based solver, Woodhouse, 1980), enabling a smooth transition from simplified to complete modelling as a solution is approached.

2.6.2 Computational Advances

Modern geophysics is computationally intensive, and – as we have seen – the feasibility of various inversion strategies is directly linked to the available resources. As such,

https://doi.org/10.1017/9781009180412.003 Published online by Cambridge University Press

computational developments are often important in driving the development and adoption of novel inference approaches. In particular, current progress leverages a number of technological advances that have been stimulated by the rapid growth of 'machine learning' applications across society. This includes general-purpose computational libraries such as Tensorflow (Abadi et al., 2016) and Pytorch (Paszke et al., 2019), along with more specialist tools such as Edward (Tran et al., 2016). A key feature of these libraries is native support for auto-differentiation, making it easy to exploit gradient-based optimisation strategies. This is an area that has previously been highlighted as ripe for exploitation in geophysics (Sambridge et al., 2007), although its use is not yet widespread. Another interesting development is the rise of packages such as FEniCS (Logg et al., 2012), which aim to automatically generate forward models from a statement of the relevant physical equations (e.g. Reuber and Simons, 2020). This has the potential to greatly expand the range of problems that it is feasible to address.

2.6.3 Novel Data – Novel Strategies

An ongoing theme of geophysics is the growth in data quantity. This is often driven by concerted efforts to collect high-resolution datasets: examples include high-quality satellite gravity measurements (e.g. Kornfeld et al., 2019), and systematic continental scale surveys such as USArray (Meltzer et al., 1999) or AusAEM (Ley-Cooper et al., 2020). Handling and processing such massive datasets has necessitated new tools and standards designed to enable easy exploitation of high-performance computing (e.g. Krischer et al., 2016; Hassan et al., 2020). On the other hand, we have also seen exciting recent developments in planetary seismology, with the recent breakthrough analysis of Martian seismic data from the InSight mission (Khan et al., 2021; Knapmeyer-Endrun et al., 2021; Stähler et al., 2021). In this context, the available dataset is very limited: we must work with a single instrument, limited capacity for data transmission, and with data characteristics quite different from those of Earth. Undoubtedly techniques will need to develop accordingly.

Another driver for innovation in geophysical inversion is innovation in data collection. Recent advances in sensor technology include the growth of distributed acoustic sensing (e.g. Daley et al., 2013; Parker et al., 2014), which uses fibre-optic cables to measure strain rates, and nodal seismic acquisition systems (Dean et al., 2018), which enable dense deployments of semi-autonomous instruments. Fully exploiting these technologies within an inversion context will doubtless motivate a new generation of analysis techniques (e.g. Lythgoe et al., 2021; Muir and Zhang, 2021), and ongoing innovation in the field of geophysical inversion.

2.7 Concluding Remarks

Athanasius Kircher published his *Mundus Subterraneus* in 1665, with his now-famous images of fiery chambers criss-crossing the Earth's interior to feed its volcanoes. What was his evidence for this structure? He acknowledges: '*sive ea jam hoc modo, sive alio*' – 'either like this, or something else'. As Waddell (2006) writes, this

> makes very clear that Kircher was not interested in whether his images had managed to capture exactly the subterranean structure of the Earth. Such large and detailed copper engravings must have been extremely expensive to commission and print, suggesting that Kircher did believe them to be important. But their value lay in their ability to encourage speculation and consideration.

Some 350 years later, geophysical images are produced with more emphasis on rigour – but otherwise, perhaps little has changed.

In this chapter, we have sought to survey and summarise the state of the art of geophysical inversion, and to highlight some of the theoretical and conceptual connections between different approaches. As we hope is clear, the field continues to develop at pace: driven by the need to better address geoscience questions; drawn on towards exciting horizons across mathematics, statistics, and computation. In particular, the growth of machine learning has focused much attention on techniques of regression, model building, and statistical inference, and the fruits of this have been evident throughout our discussion. We have no doubt that geophysical inversion will continue to produce images and models that can inspire and stimulate geoscientists for many years to come.

Acknowledgements. We are grateful to the many students, colleagues, and collaborators who have contributed to our understanding of the topics discussed in this chapter. We also thank several colleagues, and an anonymous reviewer, for helpful comments and suggestions on a draft of this work. We acknowledge financial support from the CSIRO Future Science Platform in Deep Earth Imaging, and from the Australian Research Council under grant numbers DP180100040 and DP200100053.

References

Abadi, M., Barham, P., Chen, J. et al. 2016. Tensorflow: A system for large-scale machine learning. *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, pp. 265–83, USENIX Association, Savannah, GA.

Agarwal, S., Tosi, N., Breuer, D. et al. 2020. A machine-learning-based surrogate model of Mars' thermal evolution. *Geophysical Journal International*, 222, 1656–70.

Aleardi, M., and Salusti, A. 2020. Hamiltonian Monte Carlo algorithms for target and interval-oriented amplitude versus angle inversions. *Geophysics*, 85, R177–R194.

Allmaras, M., Bangerth, W., Linhart, J. et al. 2013. Estimating parameters in physical models through Bayesian inversion: A complete example. *SIAM Review*, 55, 149–67.

https://doi.org/10.1017/9781009180412.003 Published online by Cambridge University Press

Ambrosio, L. 2003. Lecture notes on optimal transport problems. In L. Ambriosio, K. Deckelnick, G. Dziuk, M. Mimura, V. Solonnikov, and H. Soner, eds., *Mathematical Aspects of Evolving Interfaces*. Heidelberg: Springer, pp. 1–52.

Anderssen, R., Worthington, M., and Cleary, J. 1972. Density modelling by Monte Carlo inversion – I. Methodology. *Geophysical Journal of the Royal Astronomical Society*, 29, 433–44.

Aster, R., Borchers, B., and Thurber, C. 2013. *Parameter estimation and inverse problems*. Amsterdam: Academic Press.

Backus, G. 1970a. Inference from inadequate and inaccurate data, i. *Proceedings of the National Academy of Sciences*, 65, 1–7.

Backus, G. 1970b. Inference from inadequate and inaccurate data, ii. *Proceedings of the National Academy of Sciences*, 65, 281–7.

Backus, G. 1970c. Inference from inadequate and inaccurate data, iii. *Proceedings of the National Academy of Sciences*, 67, 282–9.

Backus, G. 1988. Bayesian inference in geomagnetism. *Geophysical Journal*, 92, 125–42.

Backus, G., and Gilbert, F. 1968. The resolving power of gross Earth data. *Geophysical Journal of the Royal Astronomical Society*, 16, 169–205.

Bayes, T. 1763. An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions*, 53, 370–418.

Bernal-Romero, M., and Iturrarán-Viveros, U. 2021. Accelerating full-waveform inversion through adaptive gradient optimization methods and dynamic simultaneous sources. *Geophysical Journal International*, 225, 97–126.

Betancourt, M. 2017. A conceptual introduction to Hamiltonian Monte Carlo. arXiv:1701.02434v1.

Bianco, M., and Gerstoft, P. 2018. Travel time tomography with adaptive dictionaries. *IEEE Transactions on Computational Imaging*, 4, 499–511.

Bishop, C. 1995. *Neural Networks for Pattern Recognition*. Oxford: Oxford University Press.

Blei, D., Kucukelbir, A., and McAuliffe, J. 2017. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112, 859–77.

Bodin, T., and Sambridge, M. 2009. Seismic tomography with the reversible jump algorithm. *Geophysical Journal International*, 178, 1411–36.

Bond-Taylor, S., Leach, A., Long, Y., and Willcocks, C. 2022. Deep generative modelling: A comparative review of VAEs, GANs, normalizing flows, energy-based and autoregressive models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11), 7327–47. https://doi.org/10.1109/TPAMI.2021.3116668.

Box, G., and Muller, M. 1958. A note on the generation of random normal deviates. *Annals of Mathematical Statistics*, 29, 610–611.

Bozdağ, E., Peter, D., Lefebvre, M. et al. 2016. Global adjoint tomography: First-generation model. *Geophysical Journal International*, 207, 1739–66.

Brooks, S., Gelman, A., Jones, G., and Meng, X.-L. 2011. *Handbook of Markov Chain Monte Carlo*. Boca Raton, FL: CRC Press.

Burdick, S., and Lekić, V. 2017. Velocity variations and uncertainty from trans dimensional *P*-wave tomography of North America. *Geophysical Journal International*, 209, 1337–51.

Candès, E., and Wakin, B. 2008. An introduction to compressive sampling. *IEEE Signal Processing Magazine*, 25, 21–30.

Castruccio, S., McInerney, D., Stein, M. et al. 2014. Statistical emulation of climate model projections based on precomputed GCM runs. *Journal of Climate*, 27, 1829–44.

Chandra, R., Azam, D., Kapoor, A., and Mu¨ller, R. 2020. Surrogate-assisted Bayesian inversion for landscape and basin evolution models. *Geoscientific Model Development*, 13, 2959–79.

Chou, C., and Booker, J. 1979. A Backus–Gilbert approach to inversion of travel-time data for three-dimensional velocity structure. *Geophysical Journal of the Royal Astronomical Society*, 59, 325–44.

Constable, S., Parker, R., and Constable, C. 1987. Occam's inversion: A practical algorithm for generating smooth models from electromagnetic sounding data. *Geophysics*, 52, 289–300.

Cook, A. 1990. Sir Harold Jeffreys. *Biographical Memoirs of Fellows of the Royal Society*, 36, 303–33.

Creswell, A., White, T., Dumoulin, V. et al. 2018. Generative adversarial networks: An overview. *IEEE Signal Processing Magazine*, 35, 53–65.

Curtis, A., and Lomax, A. 2001. Prior information, sampling distributions, and the curse of dimensionality. *Geophysics*, 66, 372–78.

Daley, T., Freifeld, B., Ajo-Franklin, J. et al. 2013. Field testing of fiberoptic distributed acoustic sensing (DAS) for subsurface seismic monitoring. *The Leading Edge*, 32, 699–706.

Das, S., Chen, X., Hobson, M. et al. 2018. Surrogate regression modelling for fast seismogram generation and detection of microseismic events in heterogeneous velocity models. *Geophysical Journal International*, 215, 1257–90.

de Wit, R., Käufl, P., Valentine, A., and Trampert, J. 2014. Bayesian inversion of free oscillations for Earth's radial (an) elastic structure. *Physics of the Earth and Planetary Interiors*, 237, 1–17.

Dean, T., Tulett, J., and Barwell, R. 2018. Nodal land seismic acquisition: The next generation. *First Break*, 36, 47–52.

Dettmer, J., Benavente, R., Cummins, P., and Sambridge, M. 2014. Trans-dimensional finite-fault inversion. *Geophysical Journal International*, 199, 735–51.

Dinh, H., and Van der Baan, M. 2019. A grid-search approach for 4d pressure-saturation discrimination. *Geophysics*, 84, IM47–IM62.

Donoho, D. 2006. Compressed sensing. *IEEE Transactions on Information Theory*, 52, 1289–306.

Dziewonski, A., Chou, T.-A., and Woodhouse, J. 1981. Determination of earthquake source parameters from waveform data for studies of global and regional seismicity. *Journal of Geophysical Research*, 86, 2825–52.

Earp, S., Curtis, A., Zhang, X., and Hansteen, F. 2020. Probabilistic neural network tomography across Grane field (North Sea) from surface wave dispersion data. *Geophysical Journal International*, 223, 1741–57.

Enemark, T., Peeters, L., Mallants, D. et al. 2019. Hydrogeological Bayesian hypothesis testing through trans-dimensional sampling of a stochastic water balance model. *Water*, 11(7), 1463. https://doi.org/10.3390/w11071463.

https://doi.org/10.1017/9781009180412.003 Published online by Cambridge University Press

Engquist, B., and Froese, B. 2014. Application of the Wasserstein metric to seismic signals. *Communications in Mathematical Sciences*, 12, 979–88.

Fernández-Martínez, J. L., and Fernández-Muñiz, Z. 2020. The curse of dimensionality in inverse problems. *Journal of Computational and Applied Mathematics*, 369, 112571. https://doi.org/10.1016/j.cam.2019.112571.

Fichtner, A., Zunino, A., and Gebraad, L. 2019. Hamiltonian Monte Carlo solution of tomographic inverse problems. *Geophysical Journal International*, 216, 1344–63.

Field, R., Constantine, P., and Boslough, M. 2011. Statistical surrogate models for prediction of high-consequence climate change, Tech. Rep. SAND2011-6496, Sandia National Laboratories.

Forrester, A., and Keane, A. 2009. Recent advances in surrogate-based optimization. *Progress in Aerospace Sciences*, 45, 50–79.

Galetti, E., Curtis, A., Baptie, B., Jenkins, D., and Nicolson, H. 2017. Transdimensional Love-wave tomography of the British Isles and shear-velocity structure of the East Irish Sea Basin from ambient-noise interferometry. *Geophysical Journal International*, 208, 35–58.

Gallagher, K., Bodin, T., Sambridge, M. et al. 2011. Inference of abrupt changes in noisy geochemical records using transdimensional changepoint models. *Earth and Planetary Science Letters*, 311, 182–94.

Goodfellow, I., Pouget-Abadie, J., Mirza, M. et al. 2014. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, eds., *Advances in Neural Information Processing Systems*, vol. 27. Red Hook, NY: Curran Associates. https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf.

Green, P. 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82, 711–32.

Green, W. 1975. Inversion of gravity profiles by a Backus–Gilbert approach. *Geophysics*, 40, 763–72.

Guo, P., Visser, G., and Saygin, E. 2020. Bayesian trans-dimensional full waveform inversion: Synthetic and field data application. *Geophysical Journal International*, 222, 610–27.

Hassan, R., Hejrani, B., Medlin, A., Gorbatov, A., and Zhang, F. 2020. High-performance seismological tools (HiPerSeis). In K. Czarnota, I. Roach, S. Abbott, M. Haynes, N. Kositcin, A. Ray, and E. Slatter, eds., *Exploring for the Future: Extended Abstracts*. Canberra: Geoscience Australia, pp. 1–4.

Hawkins, R., and Sambridge, M. 2015. Geophysical imaging using trans-dimensional trees. *Geophysical Journal International*, 203, 972–1000.

He, Q., and Tartakovsky, A. 2021. Physics-informed neural network method for forward and backward advection-dispersion equations. *Water Resources Research*, 57, e2020WR029479.

He, W., Brossier, R., Métivier, L., and Plessix, R.-E. 2019. Land seismic multiparameter full waveform inversion in elastic VTI media by simultaneously interpreting body waves and surface waves with an optimal transport based objective function. *Geophysical Journal International*, 219, 1970–88.

Hedjazian, N., Bodin, T., and Métivier, L. 2019. An optimal transport approach to linearized inversion of receiver functions. *Geophysical Journal International*, 216, 130–47.

Hejrani, B., and Tkalčić, H. 2020. Resolvability of the centroid-moment-tensors for shallow seismic sources and improvements from modeling high-frequency waveforms. *Journal of Geophysical Research*, 125, e2020JB019643.

Herrmann, F., Erlangga, Y., and Lin, T. 2009. Compressive simultaneous full-waveform simulation. *Geophysics*, 74, A35–A40.

Hogg, D., and Foreman-Mackey, D. 2018. Data analysis recipes: Using Markov chain Monte Carlo. *The Astrophysical Journal Supplement Series*, 236:11 (18 pp.). https://doi.org/10.3847/1538-4365/aab76e.

Huang, G., Zhang, X., and Qian, J. 2019. Kantorovich–Rubinstein misfit for inverting gravity-gradient data by the level-set method. *Geophysics*, 84, 1–115.

Hussain, M., Javadi, A., Ahangar-Asr, A., and Farmani, R. 2015. A surrogate model for simulation-optimization of aquifer systems subjected to seawater intrusion. *Journal of Hydrology*, 523, 542–554.

Jeffreys, H. 1931. *Scientific Inference*. Cambridge: Cambridge University Press.

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

Jeffreys, H., and Bullen, K. 1940. *Seismological Tables*. London: British Association for the Advancement of Science.

Karniadakis, G., Kevrekidis, I., Lu, L. et al. 2021. Physics-informed machine learning. *Nature Reviews Physics*, 3, 422–40.

Kashinath, K., Mustafa, M., Albert, A. et al. 2021. Physics-informed machine learning: Case studies for weather and climate modelling. *Philosophical Transactions*, 379, 20200093.

Käufl, P. 2015. Rapid probabilistic source inversion using pattern recognition. Ph.D. thesis, University of Utrecht.

Käufl, P., Valentine, A., O'Toole, T., and Trampert, J. 2014. A framework for fast probabilistic centroid – moment-tensor determination – inversion of regional static displacement measurements. *Geophysical Journal International*, 196, 1676–93.

Käufl, P., Valentine, A., de Wit, R., and Trampert, J. 2016a. Solving probabilistic inverse problems rapidly with prior samples. *Geophysical Journal International*, 205, 1710–28.

Käufl, P., Valentine, A., and Trampert, J. 2016b. Probabilistic point source inversion of strong-motion data in 3-D media using pattern recognition: A case study for the 2008 *Mw* 5.4 Chino Hills earthquake. *Geophysical Research Letters*, 43, 8492–8.

Kennett, B., and Engdahl, E. 1991. Traveltimes for global earthquake location and phase identification. *Geophysical Journal International*, 105, 429–65.

Khan, A., Ceylan, S., van Driel, M. et al. 2021. Upper mantle structure of Mars from InSight seismic data. *Science*, 373, 434–8.

Kingma, D., and Welling, M. 2014. Auto-encoding variational Bayes, in *2nd International Conference on Learning Representations*, ICLR 2014, Banff, AB, Canada, 14–16 April 2014, Conference Track Proceedings. https://arxiv.org/abs/1312.6114.

https://doi.org/10.1017/9781009180412.003 Published online by Cambridge University Press

Kircher, A. 1665. *Mundus subterraneus*. Amsterdam: Joannem Janssonium & ElziumWegerstraten.

Knapmeyer-Endrun, B., Panning, M. P., Bissig, F. et al. 2021. Thickness and structure of the martian crust from In Sight seismic data. *Science*, 373, 438–43.

Kobyzev, I., Prince, S., and Brubaker, M. 2021. Normalizing flows: An introduction and review of current methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43, 3964–79.

Komatitsch, D., Ritsema, J., and Tromp, J. 2002. The spectral-element method, Beowulf computing, and global seismology. *Science*, 298, 1737–42.

Kornfeld, R., Arnold, B., Gross, M. et al. 2019. GRACE-FO: The gravity recovery and climate experiment follow-on mission. *Journal of Spacecraft and Rockets*, 56, 931–51.

Koshkholgh, S., Zunino, A., and Mosegaard, K. 2021. Informed proposal Monte Carlo. *Geophysical Journal International*, 226, 1239–48.

Krischer, L., Smith, J., Lei, W. et al. 2016. An adaptable seismic data format. *Geophysical Journal International*, 207, 1003–11.

Kullback, S., and Leibler, R. 1951. On information and sufficiency. *The Annals of Mathematical Statistics*, 22, 79–86.

Lau, H., and Romanowicz, B. 2021. Constraining jumps in density and elastic properties at the 660 km discontinuity using normal mode data via the Backus-Gilbert method. *Geophysical Research Letters*, 48, e2020GL092217.

Lei, W., Ruan, Y., Bozdağ, E. et al. 2020. Global adjoint tomography: Model GLAD-M25. *Geophysical Journal International*, 223, 1–21.

Ley-Cooper, A., Brodie, R., and Richardson, M. 2020. AusAEM: Australia's airborne electromagnetic continental-scale acquisition program. *Exploration Geophysics*, 51, 193–202.

Liu, D., and Nocedal, J. 1989. On the limited memory BFGS method for large-scale optimization. *Mathematical Programming*, 45, 503–28.

Livermore, P., Fournier, A., Gallet, Y., and Bodin, T. 2018. Transdimensional inference of archeomagnetic intensity change. *Geophysical Journal International*, 215, 2008–34.

Logg, A., Mardal, K.-A., and Wells, G. N. 2012. *Automated Solution of Differential Equations by the Finite Element Method*. Heidelberg: Springer.

Lopez-Alvis, J., Laloy, E., Nguyen, F., and Hermans, T. 2021. Deep generative models in inversion: The impact of the generator's nonlinearity and development of a new approach based on a variational autoencoder. *Computers & Geosciences*, 152, 104762.

Luengo, D., Martino, L., Bugallo, M., Elvira, V., and Särkkä, S. 2020. A survey of Monte Carlo methods for parameter estimation. *EURASIP Journal on Advances in Signal Processing*, 2020, 25. https://doi.org/10.1186/s13634-020-00675-6.

Lythgoe, K., Loasby, A., Hidayat, D., and Wei, S. 2021. Seismic event detection in urban Singapore using a nodal array and frequency domain array detector: Earthquakes, blasts and thunderquakes. *Geophysical Journal International*, 226, 1542–57.

Meier, U., Curtis, A., and Trampert, J. 2007. Global crustal thickness from neural network inversion of surface wave data. *Geophysical Journal International*, 169, 706–722.

Meltzer, A., Rudnick, R., Zeitler, P. et al. 1999. USArray initiative. *GSA Today*, 11, 8–10.

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

Métivier, L., Brossier, R., Mérigot, Q., Oudet, E., and Virieux, J. 2016a. Increasing the robustness and applicability of full-waveform inversion: An optimal transport distance strategy. *The Leading Edge*, 35, 1060–7.

Métivier, L., Brossier, R., Mérigot, Q., Oudet, E., and Virieux, J. 2016b. Measuring the misfit between seismograms using an optimal transport distance: Application to full waveform inversion. *Geophysical Journal International*, 205, 345–77.

Métivier, L., Brossier, R., Mérigot, Q., Oudet, e., and Virieux, J. 2016c. An optimal transport approach for seismic tomography: Application to 3D full waveform inversion. *Inverse Problems*, 32, 115008.

Métivier, L., Brossier, R., Oudet, E., Mérigot, Q., and Virieux, J. 2016d. An optimal transport distance for full-waveform inversion: Application to the 2014 chevron benchmark data set. In C. Sicking and J. Ferguson *SEG Technical Program Expanded Abstracts*. Tulsa, OK: Society of Exploration Geophysicists, pp. 1278–83.

Montagner, J.-P., and Tanimoto, T. 1990. Global anisotropy in the upper mantle inferred from the regionalization of phase velocities. *Journal of Geophysical Research*, 95, 4797–819.

Montagner, J.-P., and Tanimoto, T. 1991. Global upper mantle tomography of seismic velocities and anisotropies. *Journal of Geophysical Research*, 96, 20337–51.

Moseley, B., Nissen-Meyer, T., and Markham, A. 2020. Deep learning for fast simulation of seismic waves in complex media. *Solid Earth*, 11, 1527–49.

Mosher, S., Eilon, Z., Janiszewski, H., and Audet, P. 2021. Probabilistic inversion of seafloor compliance for oceanic crustal shear velocity structure using mixture density networks. *Geophysical Journal International*, 227, 1879–92.

Mosser, L., Dubrule, O., and Blunt, M. 2020. Stochastic seismic waveform inversion using generative adversarial networks as a geological prior. *Mathematical Geosciences*, 52, 53–79.

Muir, J., and Zhang, Z. 2021. Seismic wavefield reconstruction using a pre-conditioned wavelet-curvelet compressive sensing approach. *Geophysical Journal International*, 227, 303–15.

Neal, R. 2011. MCMC using Hamiltonian dynamics. In S. Brooks, A. Gelman, G. Jones, and X.-L. Meng, eds., *Handbook of Markov Chain Monte Carlo*. Boca Raton, FL: CRC Press.

Nicholson, T., Sambridge, M., and Gudmundsson, O. 2004. Three-dimensional empirical traveltimes: Construction and applications. *Geophysical Journal International*, 156, 307–28.

Nocedal, J., and Wright, S. 1999. *Numerical Optimization*. New York: Springer.

Nyquist, H. 1928. Certain topics in telegraph transmission theory. *Transactions of the American Institute of Electrical Engineers*, 47, 617–44.

Parker, R. 1994. *Geophysical Inverse Theory*. Princeton, NJ: Princeton University Press.

Parker, T., Shatalin, S., and Farhadiroushan, M. 2014. Distributed acoustic sensing – a new tool for seismic applications. *First Break*, 32, 61–69.

https://doi.org/10.1017/9781009180412.003 Published online by Cambridge University Press

Paszke, A., Gross, S., Massa, F. et al. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. A. Fox, and R. Garnett, eds. *Advances in Neural Information Processing Systems*, vol. 32. Red Hook, NY: Curran Associates. https://proceedings.neurips.cc/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf.

Petersen, K., and Pedersen, M. 2012. The matrix cookbook, Tech. rep. Technical University of Denmark, Kongens Lyngby. 72 p.

Pijpers, F., and Thompson, M. 1992. Faster formulations of the optimally localized averages method for helioseismic inversions. *Astronomy and Astrophysics*, 262, L33–L36.

Press, F. 1970. Earth models consistent with geophysical data. *Physics of the Earth and Planetary Interiors*, 3, 3–22.

Quiepo, N., Haftka, R., Shyy, W. et al. 2005. Surrogate-based analysis and optimization. *Progress in Aerospace Sciences*, 41, 1–28.

Raissi, M., Perdikaris, P., and Karniadakis, G. 2019. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378, 686–707.

Ramgraber, M., Weatherl, R., Blumensaat, F., and Schirmer, M. 2021. Non-Gaussian parameter inference for hydrogeological models using Stein variational gradient descent. *Water Resources Research*, 57, e2020WR029339.

Rasmussen, C., and Williams, C. 2006. *Gaussian Processes for Machine Learning*. Cambridge, MA: MIT Press.

Ray, A. 2021. Bayesian inversion using nested trans-dimensional Gaussian processes. *Geophysical Journal International*, 226, 302–26.

Ray, A., and Myer, D. 2019. Bayesian geophysical inversion with trans-dimensional Gaussian process machine learning. *Geophysical Journal International*, 217, 1706–26.

Reuber, G., and Simons, F. 2020. Multi-physics adjoint modeling of Earth structure: combining gravimetric, seismic and geodynamic inversions. *International Journal on Geomathematics*, 11(30), 1–38.

Rezende, D., and Mohamed, S. 2015. Variational inference with normalizing flows. In F. Bach and D. Blei, eds., *Proceedings of the 32nd International Conference on Machine Learning*, vol. 37; International Conference on Machine Learning, 7–9 July 2015, Lille, France, pp.1530–8. http://proceedings.mlr.press/v37/rezende15.pdf.

Rijal, A., Cobden, L., Trampert, J., Jackson, J., and Valentine, A. 2021. Inferring equations of state of the lower mantle minerals using mixture density networks. *Physics of the Earth and Planetary Interiors*, 319, 106784.

Ritsema, J., van Heijst, H., and Woodhouse, J. 1999. Complex shear wave velocity structure imaged beneath Africa and Iceland., *Science*, 286, 1925–31.

Ruthotto, L., and Haber, E. 2021. An introduction to deep generative modelling. *GAMM-Mitteilungen*, 44, e202100008.

Sambridge, M. 1998. Exploring multidimensional landscapes without a map. *InverseProblems*, 14, 427–40.

Sambridge, M. 1999a. Geophysical inversion with a neighbourhood algorithm – I. Searching a parameter space. *Geophysical Journal International*, 138, 479–94.

Sambridge, M. 1999b. Geophysical inversion with a neighbourhood algorithm – II. Appraising the ensemble *Geophysical Journal International*, 138, 727–46.

Sambridge, M., and Kennett, B. 1986. A novel method of hypocentre location. *Geophysical Journal International*, 87, 679–97.

Sambridge, M., and Mosegaard, K. 2002. Monte Carlo methods in geophysical inverse problems. *Reviews of Geophysics*, 40, 3-1–3-29. https://doi.org/10.1029/2000RG000089.

Sambridge, M., Gallagher, K., Jackson, A., and Rickwood, P. 2006. Trans-dimensional inverse problems, model comparison and the evidence. *Geophysical Journal International*, 167, 528–42.

Sambridge, M., Rickwood, P., Rawlinson, N., and Sommacal, S. 2007. Automatic differentiation in geophysical inverse problems. *Geophysical Journal International*, 170, 1–8.

Sambridge, M., Bodin, T., Gallagher, K., and Tkalcic, H. 2012. Transdimensional inference in the geosciences. *Philosophical Transactions of the Royal Society*, 371.

Santambrogio, F. 2015. *Optimal Transport for Applied Mathematicians*. Basel: Birkhäuser.

Scales, J., and Snieder, R. 1997. To Bayes or not to Bayes. *Geophysics*, 62, 1045–46.

Scheiter, M., Valentine, A., and Sambridge, M. 2022. Upscaling and downscaling Monte Carlo ensembles with generative models. *Geophysical Journal International*, 230(2), 916–31. https://doi.org/10.1093/gji/ggac100.

Sen, M., and Biswas, R. 2017. Transdimensional seismic inversion using the reversible jump Hamiltonian Monte Carlo algorithm. *Geophysics*, 82, R119–R134.

Shahriari, B., Swersky, K., Wang, Z., Adams, R., and de Freitas, N. 2016. Taking the human out of the loop: A review of Bayesian optimization. *Proceedings of the IEEE*, 104, 148–75.

Siahkoohi, A., and Herrman, F. 2021. Learning by example: Fast reliability-aware seismic imaging with normalizing flows. arXiv:2104.06255v1. https://arxiv.org/abs/2104.06255

Simons, F., Loris, I., Nolet, G. et al. 2011. Solving or resolving global tomographic models with spherical wavelets, and the scale and sparsity of seismic heterogeneity. *Geophysical Journal International*, 187, 969–88.

Smith, J., Azizzadenesheli, K., and Ross, Z. 2020. EikoNet: Solving the eikonal equation with deep neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, 59 (12), 10685–96. https://doi.org/10.1109/TGRS.2020.3039165.

Smith, J., Ross, Z., Azizzadenesheli, K., and Muir, J. 2022. HypoSVI: Hypocenter inversion with Stein variational inference and physics informed neural networks. *Geophysical Journal International*, 228, 698–710.

Snieder, R. 1991. An extension of Backus–Gilbert theory to nonlinear inverse problems. *Inverse Problems*, 7, 409–433.

Song, C., Alkhalifah, T., and Bin Waheed, U. 2021. Solving the frequency-domain acoustic VTI wave equation using physics-informed neural networks. *Geophysical Journal International*, 225, 846–59.

Spurio Mancini, A., Piras, D., Ferreira, A., Hobson, M., and Joachimi, B. 2021. Accelerating Bayesian microseismic event location with deep learning. *Solid Earth*, 12, 1683–705.

Stähler, S., Khan, A., Banerdt, W. et al. 2021. Seismic detection of the Martian core. *Science*, 373, 443–48.

https://doi.org/10.1017/9781009180412.003 Published online by Cambridge University Press

Steinmetz, T., Raape, U., Teßmann, S. et al. 2010. Tsunami early warning and decision support. *Natural Hazards and Earth System Sciences*, 10, 1839–50.

Tanimoto, T. 1985. The Backus–Gilbert approach to the three-dimensional structure in the upper mantle – I. Lateral variation of surface wave phase velocity with its error and resolution. *Geophysical Journal International*, 82, 105–23.

Tanimoto, T. 1986. The Backus–Gilbert approach to the three-dimensional structure in the upper mantle – II. *SH* and *SV* velocity. *Geophysical Journal International*, 84, 49–69.

Tarantola, A. 2005. *Inverse Problem Theory and Methods for Model Parameter Estimation*. Philadelphia, PA: Society of Industrial and Applied Mathematics.

Tarantola, A., and Nercessian, A. 1984. Three-dimensional inversion without blocks. *Geophysical Journal of the Royal Astronomical Society*, 76, 299–306.

Tarantola, A., and Valette, B. 1982. Generalized nonlinear inverse problems solved using the least squares criterion. *Reviews of Geophysics and Space Physics*, 20, 219–32.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B*, 58, 267–88.

Trampert, J., and Snieder, R. 1996. Model estimations biased by truncated expansions: Possible artifacts in seismic tomography. *Science*, 271, 1257–60.

Tran, D., Kucukelbir, A., Dieng, A. B. et al. 2016. Edward: A library for probabilistic modeling, inference, and criticism. arXiv:1610.09787.

Valentine, A., and Davies, D. 2020. Global models from sparse data: A robust estimate of earth's residual topography spectrum. *Geochemistry, Geophysics, Geosystems*, e2020GC009240.

Valentine, A., and Sambridge, M. 2020a. Gaussian process models – I. A framework for probabilistic continuous inverse theory. *Geophysical Journal International*, 220, 1632–47.

Valentine, A., and Sambridge, M. 2020b. Gaussian process models – II. Lessons for discrete inversion. *Geophysical Journal International*, 220, 1648–56.

Valentine, A., and Trampert, J. 2016. The impact of approximations and arbitrary choices on geophysical images. *Geophysical Journal International*, 204, 59–73.

van Herwaarden, D. P., Boehm, C., Afansiev, M. et al. 2020. Accelerated full-waveform inversion using dynamic mini-batches. *Geophysical Journal International*, 221, 1427–38.

Waddell, M. 2006. The world, as it might be: Iconography and probabalism in the Mundus subterraneus of Athanasius Kircher. *Centaurius*, 48, 3–22. https://doi.org/10.1111/j.1600-0498.2006.00038.x.

Wang, Y., Cao, J., and Yang, C. 2011. Recovery of seismic wavefields based on compressive sensing by an l_1-norm constrained trust region method and the piecewise random subsampling. *Geophysical Journal International*, 187, 199–213.

Wang, Z., Hutter, F., Zoghi, M., Matheson, D., and de Freitas, N. 2016. Bayesian optimization in a billion dimensions via random embeddings. *Journal of Artificial Intelligence Research*, 55, 361–87.

Wiggins, R. 1972. The general linear inverse problem: Implication of surface waves and free oscillations for Earth structure. *Reviews of Geophysics and Space Physics*, 10, 251–85.

Woodhouse, J. 1980. The coupling and attenuation of nearly resonant multiplets in the Earth's free oscillation spectrum. *Geophysical Journal of the Royal Astronomical Society*, 61, 261–83.

Woodhouse, J., and Dziewonski, A. 1984. Mapping the upper mantle: Three-dimensional modelling of Earth structure by inversion of seismic waveforms. *Journal of Geophysical Research*, 89, 5953–86.

Worthington, M., Cleary, J., and Anderssen, R. 1972. Density modelling by Monte Carlo inversion – II. Comparison of recent Earth models. *Geophysical Journal of the Royal Astronomical Society*, 29, 445–57.

Zaroli, C. 2016. Global seismic tomography using Backus–Gilbert inversion. *Geophysical Journal International*. 207, 876–88.

Zhang, C., Bütepage, J., Kjellström, H., and Mandt, S. 2019. Advances in variational inference. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41, 2008–26.

Zhang, X., and Curtis, A. 2020. Seismic tomography using variational inference methods. *Journal of Geophysical Research*, 125, e2019JB018589.

Zhao, X., Curtis, A., and Zhang, X. 2022. Bayesian seismic tomography using normalizing flows. *Geophysical Journal International*, 228, 213–39.

https://doi.org/10.1017/9781009180412.003 Published online by Cambridge University Press