PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

Transdimensional inference in the geosciences

M. Sambridge, T. Bodin, K. Gallagher and H. Tkalcic

Phil. Trans. R. Soc. A 2013 **371**, 20110547, published 31 December 2012

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A

rsta.royalsocietypublishing.org

Research

Cite this article: Sambridge M, Bodin T, Gallagher K, Tkalčić H. 2013 Transdimensional inference in the geosciences. Phil Trans R Soc A 371: 20110547.
http://dx.doi.org/10.1098/rsta.2011.0547

One contribution of 17 to a Discussion Meeting Issue 'Signal processing and inference for the physical sciences'.

Subject Areas:
geophysics

Keywords:
inversion, Bayesian inference, variable parametrization

Author for correspondence:
M. Sambridge
e-mail: malcolm.sambridge@anu.edu.au

Royal Society Publishing
Informing the science of the future

Transdimensional inference in the geosciences

M. Sambridge[1], T. Bodin[2], K. Gallagher[3] and H. Tkalčić[1]

[1]Research School of Earth Sciences, Australian National University, Canberra, Australian Capital Territory 0200, Australia
[2]Department of Earth and Planetary Science, University of California, Berkeley, CA 94720, USA
[3]Géosciences Rennes, Université de Rennes 1, Rennes 35042, France

Seismologists construct images of the Earth's interior structure using observations, derived from seismograms, collected at the surface. A common approach to such inverse problems is to build a single 'best' Earth model, in some sense. This is despite the fact that the observations by themselves often do not require, or even allow, a single best-fit Earth model to exist. Interpretation of optimal models can be fraught with difficulties, particularly when formal uncertainty estimates become heavily dependent on the regularization imposed. Similar issues occur across the physical sciences with model construction in ill-posed problems. An alternative approach is to embrace the non-uniqueness directly and employ an inference process based on parameter space sampling. Instead of seeking a best model within an optimization framework, one seeks an ensemble of solutions and derives properties of that ensemble for inspection. While this idea has itself been employed for more than 30 years, it is now receiving increasing attention in the geosciences. Recently, it has been shown that transdimensional and hierarchical sampling methods have some considerable benefits for problems involving multiple parameter types, uncertain data errors and/or uncertain model parametrizations, as are common in seismology. Rather than being forced to make decisions on parametrization, the level of data noise and the weights between data types in advance, as is often the case in an optimization framework, the choice can be informed by the data themselves. Despite the relatively high computational burden involved, the number of areas where sampling methods are now feasible is growing rapidly. The intention of this article is to introduce concepts of

transdimensional inference to a general readership and illustrate with particular seismological examples. A growing body of references provide necessary detail.

1. Introduction

Imaging Earth's interior through the use of seismic waves is a popular technique for constraining the internal structure and composition of our planet and has been actively developed by the seismological community for many decades. Usually, the Earth model is parametrized using basis functions covering a two- or three-dimensional volume, e.g. uniform local cells in two or three dimensions, and the observations are used to constrain some property within each cell, e.g. seismic wave speed. The details of the parametrization, e.g. cell size and shape, are almost always chosen in advance. Figure 1 shows some examples. The choice of cell size defines the volume of Earth material which is averaged over in the parametrization.

In seismology, and indeed geophysics more generally, data constraints are highly variable in space, owing to either logistical restrictions in making observations, e.g. seismological observations are largely restricted to being made within continental regions, or the uneven spatial distribution of natural sources, e.g. earthquakes are mostly restricted to the boundaries of tectonic plates. In global seismology, the path that energy travels from source to receiver samples Earth's interior highly unevenly. The usual way of addressing these limitations is to use a tomographic imaging technique and apply some spatial smoothing, norm damping or simply to coarsen the parametrization in ill-constrained volumes of the model [4]. In practice, this usually reduces to the need to solve a large linear system of equations with regularization [4]. A key point is that regularization is often applied uniformly across the entire model, which raises the possibility that, while the ill-constrained regions are being appropriately damped, the well-constrained regions are also being oversmoothed and hence information may be lost. This has prompted authors to consider alternate regularization mechanisms that allow sharp discontinuities and multi-scale features [5–7]. A second issue is that, in this situation, formalized estimates of uncertainty in the sought after model are often strongly influenced by the nature of the regularization imposed.

In this paper, we highlight some key developments in the application of transdimensional model inference in the geosciences, focusing to some extent on seismological problems. This constitutes an alternative approach to the estimation of single regularized models in inverse problems where the unknown is a function of space (in one, two or three dimensions) or time. We explain the main ideas and present a few recent examples of areas where transdimensional inversion techniques have been applied. Nothing in the methodology is particular to geophysical inverse problems. We argue that the same ideas could find fruitful application more broadly across the physical sciences.

2. Transdimensional inversion

Most inverse problems in the geosciences treat the number of unknowns (or model parameters), k, as a constant. Transdimensional inversion is the name given to the case where this assumption is relaxed, and k is treated as an unknown. In principle, one should distinguish between the cases where the number of unknowns is a measurable physical quantity, for example a distinct number of components in a mixture, and the cases where it merely represents the number of basis functions chosen for a model equation (2.1), e.g. layers in a subsurface model of seismic wave speeds. In what follows, we draw no such distinction and simply use the number of unknowns as a control on the number of degrees of freedom within the model. The motivation is that, by extending fixed dimensional inverse problems into the transdimensional regime, we can use the data themselves to constrain the maximum allowable complexity in the model rather than specifying this beforehand. Of course, within an optimization framework, one can nearly always

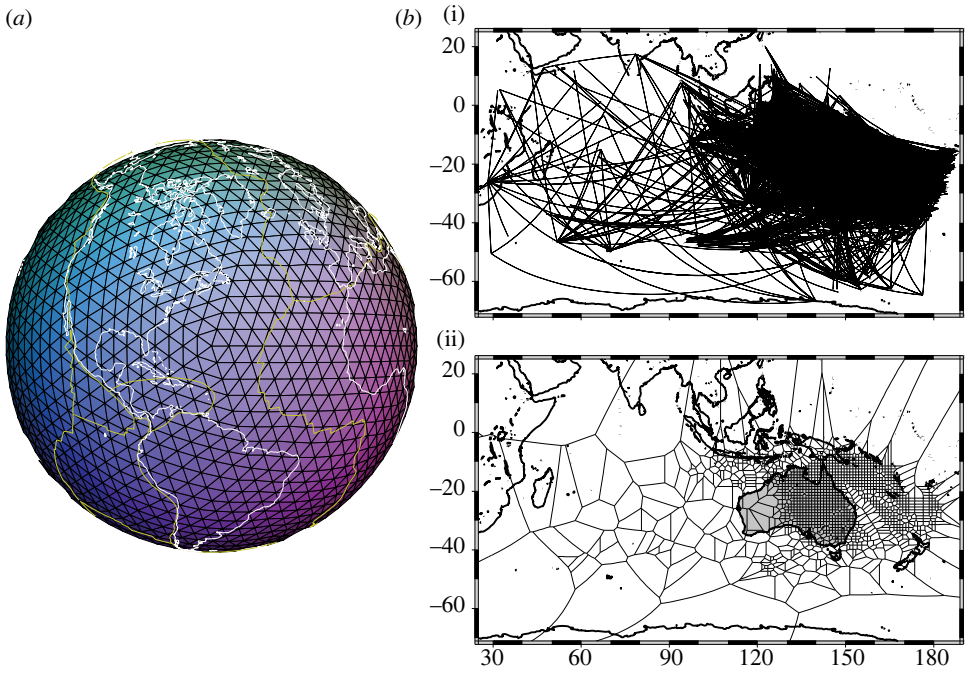

Figure 1. (*a*) A uniform triangular parametrization covering the Earth's surface used in seismic imaging studies [1]. (*b*) (i) Seismic surface wave paths in the Australasian region as used in the study of Debayle & Sambridge [2]. (ii) Voronoi cells [3] constructed on the Earth's surface that reflect the relative resolution of the model parameters constrained by the data. Larger cells lie in regions with poorer resolution. (Online version in colour.)

fit data better by introducing more unknowns, but as we shall see the situation is quite different within a Bayesian sampling framework which is naturally parsimonious (see [8] for a detailed discussion). This means that more complex models are not necessarily preferred over simpler ones and Ockham's razor prevails.

Transdimensional inversion is a sampling-based approach in which the model is expanded in terms of a variable number of basis functions whose position and coefficients are unknowns

$$m(\mathbf{x}) = \sum_{i=1}^{k} m_i \phi_i(\mathbf{x}). \tag{2.1}$$

Here $m(\mathbf{x})$ represents the physical quantity of interest to be constrained by the data, which is a function of spatial position $\mathbf{x}$. For example, in seismic tomography, $m(\mathbf{x})$ represents either the speed at which the corresponding seismic wave travels through the Earth or its slowness (reciprocal of wave speed). $\phi_i(\mathbf{x})$ is the ith basis function, which parametrizes the mathematical model of the Earth. The parameter, m_i, where $i = 1, \ldots, k$, is the coefficient of the ith basis function and k is the total number of unknowns. Traditional approaches to imaging problems treat $\phi_i(\mathbf{x})$ and k as knowns, chosen in advance, with m_i found by some optimization process where predictions from the model $m(\mathbf{x})$ are made to fit the observations as well as regularizing criteria. In the transdimensional framework, m_i, k and $\phi_i(\mathbf{x})$ may all be treated as unknowns. The key idea is to use the observational data to constrain these parameters rather than simply fixing some in advance.

The approach we have taken is to adapt the Bayesian partition (or changepoint) modelling methodology described by Denison *et al*. [9]. Here we briefly summarize only the underlying theory. The reader is referred to the numerous papers cited for mathematical details. In a Bayesian formulation, information is represented by probability distributions. One begins with a prior

probability density function (PDF) on the unknowns $p(\mathbf{m})$, where the vector $\mathbf{m}$ represents all unknowns in the inverse problem. The choice of prior is the single most controversial component of the Bayesian approach with a long history of debate [10,11]. Whatever the choice, all results of a Bayesian inversion are dependent on the selected prior, which must be clearly stated and results interpreted accordingly. We will assume that some suitable subjective prior can be found. In our applications, it is often a simple uniform PDF between predetermined bounds. Next, one defines a likelihood function, $p(\mathbf{d}|\mathbf{m})$, which is literally interpreted as the probability of the observed data given the model. The bar notation indicates that the terms to the right are given or fixed (and the value of the likelihood is conditional on these values). To evaluate the likelihood function, one generally needs a statistical model of errors in the data as well as the ability to calculate predictions from a model ($\mathbf{d}_\mathrm{p} = g(\mathbf{m})$), often called a solution to the forward problem. In practice, the likelihood then measures the probability that the discrepancies between the observed data, $\mathbf{d}$, and predictions from a model, $g(\mathbf{m})$, are due to random error alone.

Bayes' rule [12] links these two with the posterior PDF on the model $p(\mathbf{m}|\mathbf{d})$, which is interpreted as the probability of the model given in the data. We have

$$p(\mathbf{m}|\mathbf{d}) = \lambda p(\mathbf{d}|\mathbf{m})p(\mathbf{m}). \tag{2.2}$$

The reciprocal of the constant of proportionality, λ, is a term called the evidence, which is not a function of the model and is often neglected in many studies. However, this term can play an important role in measuring the validity of the underlying assumptions upon which the formulation of the inverse problem is based, e.g. the statistics of noise process, or the physical theory behind the forward problem connecting model to data [13]. The solution to a Bayesian formulation is not a single optimal model but the entire posterior PDF in (2.2). In many geoscience inference problems, the length of $\mathbf{m}$ may vary between 10^0 and 10^7, rendering direct inspection of the posterior impractical. For problems with 10^1–10^3 unknowns, sampling methods may be used to generate models whose density follows the posterior PDF, and then properties of that ensemble may be determined, e.g. mean, covariance or marginal PDFs for inspection. Lower dimensional marginal PDFs are a projection of the full multi-dimensional PDF onto a subset of model parameters and can be a useful way of gleaning information on the model space. In all cases, posterior properties must be compared with their counterparts in the prior, as it is the difference between these two PDFs which represents the influence of the data. Over the past 30 years, many Bayesian computational methods have been developed to sample arbitrarily high-dimensional PDFs, the class of technique in most widespread use is Markov chain Monte Carlo (MCMC; see [14] for a review).

(a) Mobile basis functions

In a transdimensional setting, the dimension of the model vector $\mathbf{m}$ is itself of unknown length. In general, the basis functions in (2.1) may be defined by both location and scale parameters. For example, a Gaussian basis function would have a mean and standard deviation. In what follows we restrict attention to the case where the basis functions, $\phi_i(\mathbf{x})$, depend only on a single spatial reference vector $\mathbf{x}_i$, with the scale length implicitly defined by the relationship between position vectors. An example is given in figure 2, which shows a layered Earth model as a function of a single depth coordinate. This is a typical case in many geophysical problems, where seismic, electrical or heat flow observations are collected at the surface, and one wishes to recover physical properties, e.g. seismic wave speeds, conductivities or temperatures, within each layer. The number and thicknesses of the layers can be varied by choosing the ith basis function, which is unity in the ith layer and zero elsewhere

$$\phi_i(z) = \delta_{ij}, \quad z_{j-1} \le z < z_j. \tag{2.3}$$

In this case, there are n_L layers and $k = 2n_\mathrm{L} + 1$ unknowns in total, $m_i, z_i, n_\mathrm{L} (i = 1, \ldots, n_\mathrm{L})$, the lower half space being held fixed. Here the depths to each interface, z_i, are variable and are represented by black circles in figure 2. An alternative is to define the interfaces implicitly using

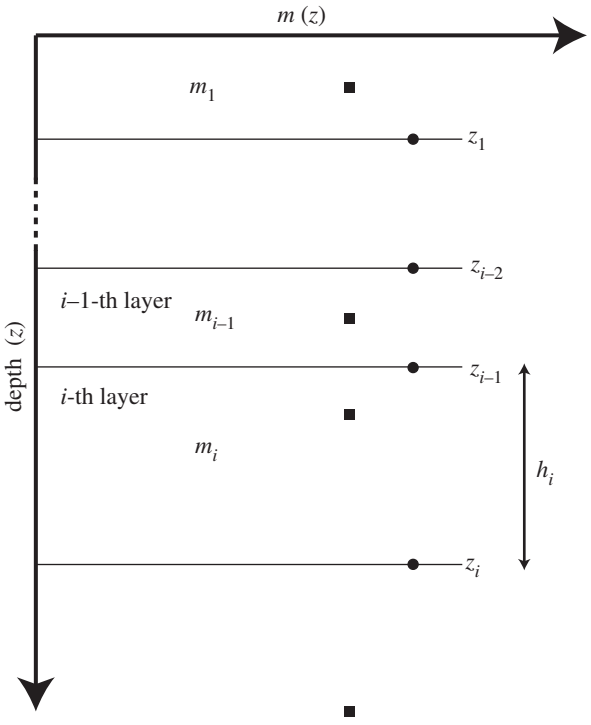

Figure 2. Parametrization of a one-dimensional transdimensional model. Each layer has a physical property represented by m_i, which is an unknown in the inverse problem. The number of layers, n_L, and their thicknesses are also variable. The basis functions, $\phi_i(z)$, in this class of model can be represented using either the depth of each interface (black circles) or the Voronoi nuclei (black squares). In the latter case, the interfaces are defined implicitly as the mid-point between adjacent nuclei.

depth nodes positioned such that the interfaces are equi-distant between nodes (see black squares in figure 2). While there is little benefit in one- and two-dimensional problems, this approach becomes a convenient way to locally partition the Euclidean plane or surface of a sphere. The nearest-neighbour regions built about a set of nuclei in two dimensions form the geometrical construct known as Voronoi cells [3]. An example is shown in figure 1*b*, where Voronoi cells are built around defining nuclei (not shown) on the surface of the Earth. This was used by Debayle & Sambridge [2] to parametrize the Earth in an inversion of seismic surface waves.

(b) Transdimensional sampling

The extension of (2.2) to include transdimensional model vectors, $\mathbf{m}$, is straightforward. If the total number of unknowns is k, then we can rewrite the fixed dimensional Bayes' theorem in (2.2) as

$$p(\mathbf{m}|\mathbf{d},k) \propto p(\mathbf{d}|\mathbf{m},k)p(\mathbf{m}|k), \tag{2.4}$$

where a k to the right of the conditional bar in each term denotes that the number of unknowns is fixed. It can be shown [13] that the appropriate extension to the transdimensional setting is

$$p(\mathbf{m},k|\mathbf{d}) \propto p(\mathbf{d}|\mathbf{m},k)p(\mathbf{m}|k)p(k). \tag{2.5}$$

Here, k is now a variable on the left-hand side with its own prior PDF, $p(k)$. The data, $\mathbf{d}$, are now used to jointly constrain $\mathbf{m}$ and k. The relationship between the fixed and transdimensional posteriors can be shown to be

$$p(\mathbf{m},k|\mathbf{d}) = p(\mathbf{m}|\mathbf{d},k)p(k|\mathbf{d}), \tag{2.6}$$

which is eqn (12) in [13]. This shows that the posterior for the variable dimension case is equal to the fixed dimension posterior multiplied by the posterior for the dimension, $p(k|\mathbf{d})$, as expected in a hierarchical probability model. The latter term is the information provided by the data on the dimension alone and is found by integrating the variable dimensional posterior over the model parameters

$$p(k|\mathbf{d}) = \int p(\mathbf{m}, k|\mathbf{d})\, \mathrm{d}\mathbf{m}. \tag{2.7}$$

In standard Bayesian inference, it is usual to fix k and the nature of the basis function in advance, and then use MCMC methods to draw samples from the fixed k posterior in (2.4). In transdimensional problems, samples must be drawn from the variable k posterior in (2.5). An extension of the well-known MCMC method to sample arbitrary dimension PDFs is the reversible jump MCMC technique given by Green [15]. A slightly less general but more commonly used special case of the reversible jump algorithm is the birth–death MCMC algorithm of Geyer & Møller [16], which has found many applications in the geosciences.

The first use of these techniques in geophysics was by Malinverno [17] in the inversion of surface sounding data for electrical resistivity depth profiles, and later for inversion of seismic travel time observations for depth profiles of seismic wave speed [18]. Subsequent applications have appeared in a variety of geophysical and geochemical inference settings, including low-temperature thermochronology [19,20], regression problems [13], estimation of basin stratigraphy and borehole lithology [21,22], borehole temperature inversion for palaeo-climate histories [23,24], geochemical mixing problems [25], exploration geophysics [18,26], seismic tomography [27–29], inversion of seismic surface waves and receiver functions for crustal structure [30,31], geoacoustics [32], airborne imaging [33] and microtremor arrays [34]. Useful reviews of theory and applications to statistical and geochemical inference problems can be found in the earlier studies [9,35,36].

We do not go into the details of the transdimensional sampling algorithms here, as they are described in many of the above texts. Although experience is growing with their use in geophysical problems, they are still the subject of ongoing research, especially in the areas of efficiency and choice of the class of basis function suited to particular inference problems [37,38]. In cases where the maximum model dimension is high, typically k greater than a few 100, most transdimensional sampling techniques are likely to become inefficient owing to the well-known curse of dimensionality [39]. In the next sections, we give some examples of their application and conclude with some comments on extension to much higher-dimensional linear inverse problems.

3. Examples

(a) Regression

A straightforward multi-component regression problem illustrates the main points of a transdimensional sampling algorithm. This example follows Bodin *et al.* [29]. In figure 3*a*, the dots are the observations of a piecewise constant function with nine partitions (thick piecewise constant line) contaminated with Gaussian random noise with standard deviation of 10 units in the y-direction, $\mathbf{d}^{\mathrm{T}} = [y_1^o, \ldots, y_N^o]$. The major discontinuous steps, or change points, in the true function lie at x positions of 1.0, 2.3, 6.5 and 8. Here the problem is to reconstruct the thick piecewise constant line from the dots. The model parametrization is built from zeroth-order polynomials, i.e. simple constants within an unknown number of partitions just as in figure 2 and equation (2.3). The thin piecewise constant line in figure 3*a* shows an example model with $n_{\mathrm{p}} = 6$ partitions and y-values randomly selected between the range of the data. The partition boundaries are defined implicitly using the positions, x_i, of the Voronoi nuclei (squares). Hence there are $k = 2n_{\mathrm{p}} + 1$ unknowns in total that must be constrained by the data, i.e. $\mathbf{m}^{\mathrm{T}} = [y_i, x_i, n_{\mathrm{p}}](i = 1, \ldots, n_{\mathrm{p}})$. Flat prior PDFs are used for all model parameters. For $y_i(i = 1, \ldots n_{\mathrm{p}})$ the prior ranges between

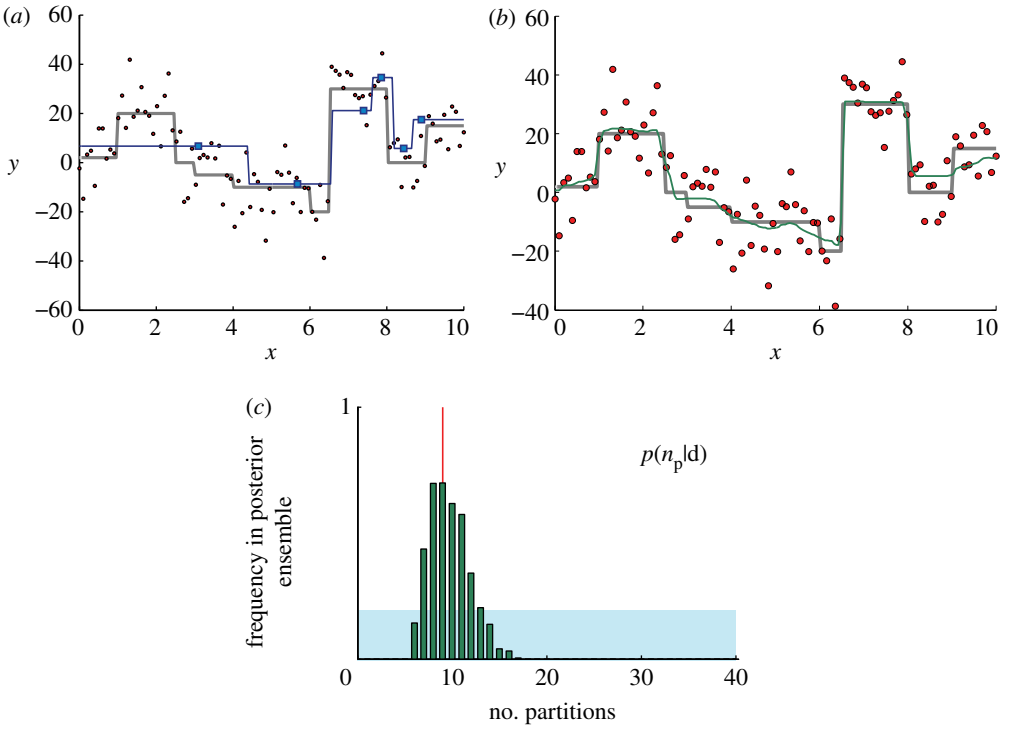

Figure 3. (*a*) Example regression problem. The 100 dots are the data generated by adding Gaussian random noise with s.d., $\sigma = 10$ units, to the grey piecewise 'true' model made from nine partitions along the x-axis. The thin piecewise constant line is a typical model in the variable parametrization containing six partitions each defined by a single nucleus (square). (*b*) Same as (*a*) but showing the mean of the ensemble of models (thin line) generated by transdimensional sampling. Extraneous features in individual models of the ensemble are cancelled out in the mean model, while common features can be reinforced. The mean model represents the true model well, including the major change points (y-jumps). (*c*) Histogram of the number of partitions in the posterior ensemble. The prior PDF is flat between 1 and 40 (shaded), while the posterior distribution is approximately centred around the true value of 9. The complexity of the model is effectively controlled by the data themselves. (Online version in colour.)

the limits of the observations, for $x_i (i = 1, \ldots n_p)$ the prior ranges between 0 and 10, and for n_p between 1 and 50. Since the noise statistics are i.i.d. Gaussian random variables then the likelihood is given by

$$p(\mathbf{d}|\mathbf{m}, k) = \frac{1}{(2\pi\sigma)^{N/2}} \exp\left\{-\frac{1}{2}\sum_{j=1}^{N}\frac{(y_j^o - y_i(x_i))^2}{\sigma^2}\right\}. \tag{3.1}$$

Here y_j^o is the jth datum value (dots in figure 3*b*), and y_i is the model parameter in the ith partition defined by the nucleus at x_i, where the ith partition contains the jth datum point. An implementation of the birth–death MCMC algorithm was run on this problem. Details can be found in Bodin *et al.* [29]. This generates an ensemble of piecewise constant curves $\mathbf{m}_l (l = 1, \ldots, M)$ with a variable number of partitions distributed according to the posterior density equation (2.7). Taking a mean curve by averaging the ensemble at each point along the x-axis gives the thin curve in figure 3*b*, which is clearly a reasonable recovery of the true (thick piecewise constant curve). In particular, the major change points in the true curve have all been well recovered, even though averaging is inherently a smoothing process. The key point in this example is that the scale lengths of the mean curve in both the x- and y-directions are entirely constrained by the data through the mobile basis function parametrization and transdimensional sampling.

(i) Parsimony

Although the focus is on generating an ensemble of solutions, it is tempting to take the mean curve as an estimator of the true solution, a process often called frequentist or pragmatic Bayes. From this viewpoint, the transdimensional inversion can be thought of as a sophisticated regularizer, estimating the true model by self-adapting to the data in a local manner, without introducing unnecessary detail. Figure 3*c* shows a histogram of the number of partitions in the curves of the posterior ensemble, i.e. an estimate of $p(n_p|\mathbf{d})$. Comparing the posterior with the prior, which corresponds to a constant density between 1 and 50 partitions (shaded region), one clearly sees that the effect of the data is to concentrate the posterior about the true value of 9 (thin vertical bar). Notably, there is no preference for curves of large numbers of partitions, even though these can more easily fit the data and reduce the likelihood. This is an example of the parsimony of transdimensional sampling described above. A more detailed discussion of parsimony and its dependence on the prior PDF is given by Mackay [8] and Gallagher *et al*. [35].

(ii) Hierarchical Bayes

In the previous example, it was assumed that the variance of the data noise ($\sigma^2 = 100$) was known. This is a crucial assumption, because the complexity of the final ensemble and the mean curve are determined by the noise levels in the data. As is apparent in figure 3*c*, transdimensional sampling produces models whose complexity is consistent with the level of data noise. If the data variance, i.e. σ^2 in the likelihood function (3.1), is assumed smaller than the actual value then more complexity is required to fit to the apparently higher quality data. Hence more structure will be introduced into the mean model. Figure 4*a* demonstrates this point. Here the sampling is repeated using a data noise of $\sigma^2 = 16$ with the result that considerably more structure is introduced into the mean model. This is evident from the plot of the marginal on the number of partitions $p(n_p|\mathbf{d})$, which on average is much larger than the true value of 9, as well as the complexity of the mean model (continuous thin line curve in figure 4*a*). Figure 4*b* shows results from the converse case where data noise variance is assumed too high, $\sigma^2 = 900$. In this case, the mean model is much too smooth and the marginal on n_p clearly shows that the number of partitions is on average much smaller than the true value.

It is clear then that model complexity is directly linked to the level of data noise. The transdimensional sampling algorithm introduces only enough complexity to fit the data to within the assumed noise level. This is somewhat similar to the tuning of regularization terms in linear inversion with the discrepancy principle [40]. If the nature of the data noise is uncertain then ideally it should be parametrized and included as an unknown in the inverse problem. This is done in the following example using the so-called 'hierarchical Bayes' approach [9,18,29]. Again we omit the details that can be found in the references. For the regression problem, the sampling algorithm is extended by addition of a single parameter, σ. The hierarchical name comes from the fact that parameters of different type may grouped into separate classes. In this case, it refers to just two classes, i.e. the single data noise parameter, σ, and non-noise parameters ($[y_i, x_i, n_p](i = 1, \ldots, n_p)$). The hierarchical division between parameters simply acknowledges the fact that for fixed values of the x and y parameters it is straightforward to evaluate the likelihood function (3.1) for a range of σ values. As σ is increased the exponent in the likelihood function becomes smaller in absolute size, thereby increasing the overall likelihood. However, σ also appears in the denominator of the normalizing factor in (3.1) and so an increase would tend to decrease the likelihood. Hence the constraints on σ from the data will reflect a balance between these competing factors.

Figure 4*c*,*d* shows results of the hierarchical scheme with the data noise parameter, σ, included as a variable. A flat prior, $p(\sigma)$, is used ranging between 1 and 40 units. By including data noise as an unknown, the mean model (continuous thin line) is again a reasonable recovery of the true curve (piecewise constant line), similar to the case where the value of σ was fixed at the correct value (figure 3). The complexity of the model has also adapted to produce a mean solution with major change points recovered, while also recovering the level of data noise. Figure 4*d* shows

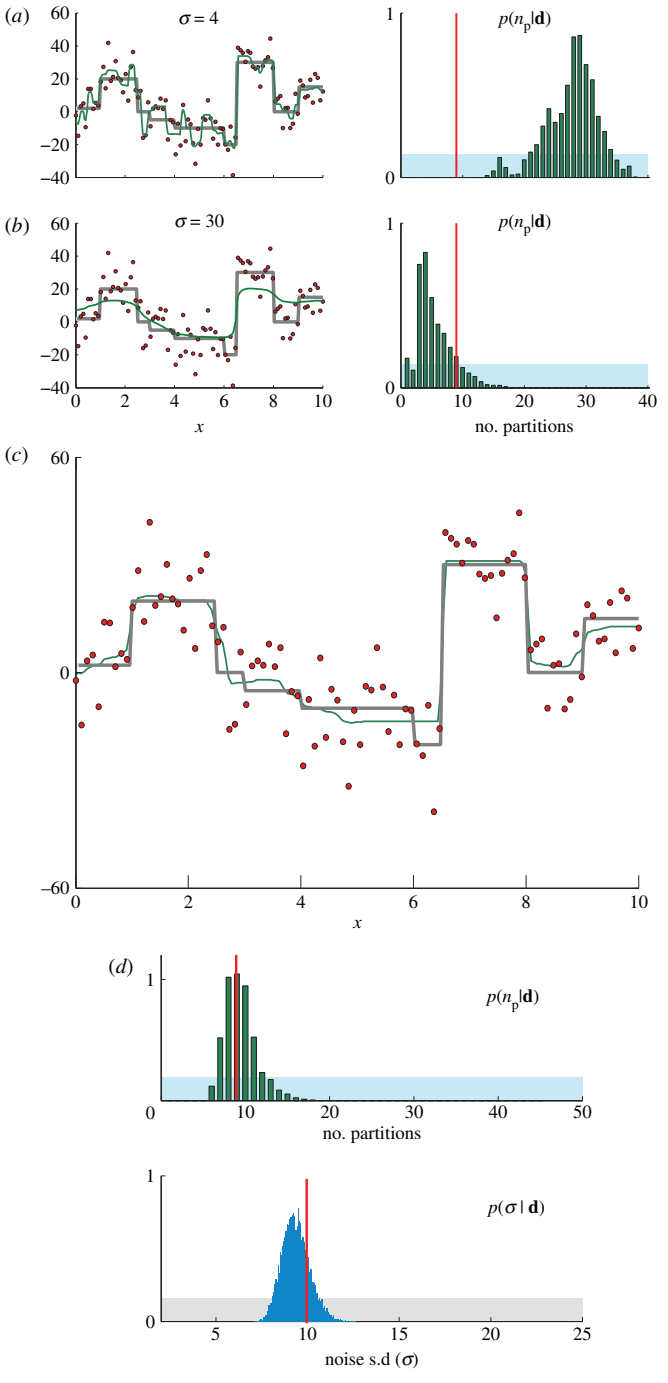

Figure 4. (*a*) Similar to figure 3*b*,*c*, showing the mean model and posterior on the number of partitions $p(k|\mathrm{d})$ for the case when the data noise variance is underestimated. The transdimensional sampling algorithm introduces additional complexity in order to fit the data to a higher level than is necessary. (*b*) Same as (*a*) for the case where the data noise variance is overestimated. Now fewer partitions are introduced and the mean model is too smooth. (*c*) Results of the hierarchical scheme where the data noise, σ, is also treated as an unknown with flat prior PDF between 1 and 40 units. Results are similar to figure 3*b* when data noise was set at the correct value. The mean model again recovers the true model (piecewise constant line) reasonably well, including changepoints. (*d*) Histograms of the number of partitions/layers and noise parameter σ in the posterior ensemble which is approximately centred about the true value of 9. Priors are shaded in all cases. (Online version in colour.)

the estimated marginals for the number of partitions, $p(n_p|\mathbf{d})$, and the noise parameter, $p(\sigma|\mathbf{d})$, calculated from histograms of the posterior ensemble. In both cases, the vertical bar represents the true values and the range of the figure equals that of the flat priors (shaded). Both the complexity and the noise level have been simultaneously recovered within a reasonable range of uncertainty in this case, indicating that they can be jointly constrained by the data.

For one-dimensional problems, such as regression, ensemble properties other than the mean can easily be calculated and inspected, including 95 per cent; credible intervals at each x-point, one-dimensional marginal PDFs at each x-point and model covariances (some examples of these appear below).

(b) Joint inversion of multiple data types

A second example is taken from seismology and addresses the case of constraining subsurface seismic wave speeds jointly from two different classes of observations. As in the regression example, a layered or partitioned structure is used to represent the shear wave speed profile with depth. As mentioned earlier, this consists of a variable number of interfaces, n_L, at depths $z_i (i = 1, \ldots, n_L)$, defining each layer as in figure 2 with shear wave speeds $v_i (i = 1, \ldots, n_L + 1)$. The lowermost layer is unbounded from below and so the total number of unknowns $k = 2n_L + 2$. This model is to be constrained by measurements of receiver functions and surface wave dispersion measurements collected at a single receiver on the surface. Seismic receiver functions are time-dependent signals derived from deconvolving the vertical from the horizontal displacement waveforms of distant earthquakes recorded at a broadband seismometer. Figure 5*a* shows an example. Their calculation and use in constraining near-receiver structure has been the subject of much study over the past 20 years [30,41]. It is well known that receiver functions are sensitive to the location of changepoints, or discontinuities in the seismic wavespeed as a function of depth in the crust.

The second class of observations, seismic surface wave dispersion measurements, are obtained from mapping the frequency dependence of seismic surface wave energy from an earthquake detected at a receiver. An example is seen in figure 5*b*. This type of observation is less sensitive to discontinuities in Earth properties but provides useful constraints on absolute values of shear wavespeed in the crust. Ideally, a joint inversion of both classes of data for crustal shear wave profiles would be preferable, as we obtain shear wave profiles with depth reflecting the information in the combined data. Commonly, this is done in an optimization framework, and the nonlinear dependence of the observables on the model results in a nonlinear optimization problem. Furthermore, a recurring difficulty for joint inversion is to determine how best to weight one class of observation relative to the other. Practitioners tend to make 'informed guesses' as to the relative weighting. In principle, the answer is to build a likelihood function which scales each class of data by its respective observational errors. For example, assuming that the errors are Gaussian and independent between data types, we have

$$p(\mathbf{d}|\mathbf{m}, k) = \frac{1}{[(2\pi)^{n_r+n_s}|C_r||C_s|]^{1/2}} \exp\left\{-\frac{1}{2}\Phi(\mathbf{m}, k)\right\}, \tag{3.2}$$

where n_r and n_s are the numbers of measurements of receiver function and surface wave data used, C_r and C_s are the covariance matrices of the receiver function and surface wave errors, respectively, and $\Phi(\mathbf{m}, k)$ is the misfit function measuring discrepancies between observations $(\mathbf{d}_r, \mathbf{d}_s)$ and predictions $(g_r(\mathbf{m}), g_s(\mathbf{m}))$ from the shear wave profile $\mathbf{m}$. The misfit, expressed in vector form, can be written

$$\Phi(\mathbf{m}, k) = (\mathbf{d}_r - g_r(\mathbf{m}))^T C_r^{-1} (\mathbf{d}_r - g_r(\mathbf{m})) + (\mathbf{d}_s - g_s(\mathbf{m}))^T C_s^{-1} (\mathbf{d}_s - g_s(\mathbf{m})), \tag{3.3}$$

where the weighting between the two classes of data is determined by the relative size of entries in the covariance matrices, C_r and C_s. Typically, surface wave dispersion errors are assumed to be independent and hence C_s is diagonal. For receiver functions, each value in $\mathbf{d}_r$ represents the

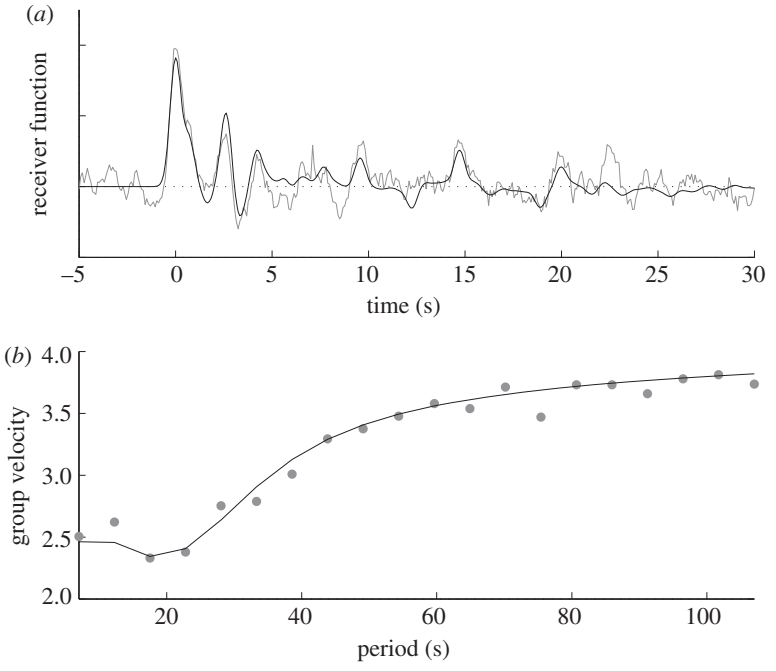

Figure 5. (*a*) Example synthetic receiver function (black, no noise) calculated using techniques described by Bodin *et al*. [31], with correlated Gaussian random noise added (grey, noise). The grey curve is discretized and used as the input data for the joint inversion example; (*b*) synthetic dispersion curve (black, no noise) corresponding to the same shear wave velocity model as (*a*). Dispersion data used in the joint inversion appear as dots (noise), which are samples of the black curve with i.i.d. Gaussian random noise added. In both cases, noise variance is comparable to that of real observations.

amplitude of a time-varying signal, and so errors are correlated between samples and C_r has a banded structure (see [31] for details).

As has been seen in the regression example, even with transdimensional sampling of the model $\mathbf{m}$, knowledge of the data noise is important to recover models with the correct complexity. However, in many situations, one has inadequate knowledge of the error statistics to calculate both covariance matrices fully. In such cases, the hierarchical approach developed in the regression problem can be extended to this case. The key idea is to parametrize each data covariance matrix by introducing additional unknowns, in a way that represents the degree of information available on the respective errors. We use a synthetic example from Bodin *et al*. [31] to illustrate. It is reasonable to assume that errors in dispersion measurements are independent and identically distributed, in which case the covariance matrix becomes proportional to the identity

$$C_s = \sigma_s^2 I, \tag{3.4}$$

and the standard deviation parameter, σ_s^2, becomes a new unknown. For the time-correlated receiver functions, we assume an exponential covariance function described by two more unknowns, a variance, σ_r^2, and a correlation parameter, r. In this case, C_r becomes a symmetric Toeplitz matrix

$$(C_r)_{i,j} = \sigma_r^2 r^{|i-j|}. \tag{3.5}$$

Although other parametrizations are possible, this one leads to simple expressions for the inverse and determinant of each covariance matrix, and makes it straightforward to evaluate the likelihood function (3.2)–(3.3) for a given combination of shear wave model $\mathbf{m}$ and noise parameters (σ_s, σ_r, r). For this set-up, the hierarchical transdimensional algorithm now consists of sampling over the shear wave velocity parameters, v_i, the interface depth parameters, z_i, the

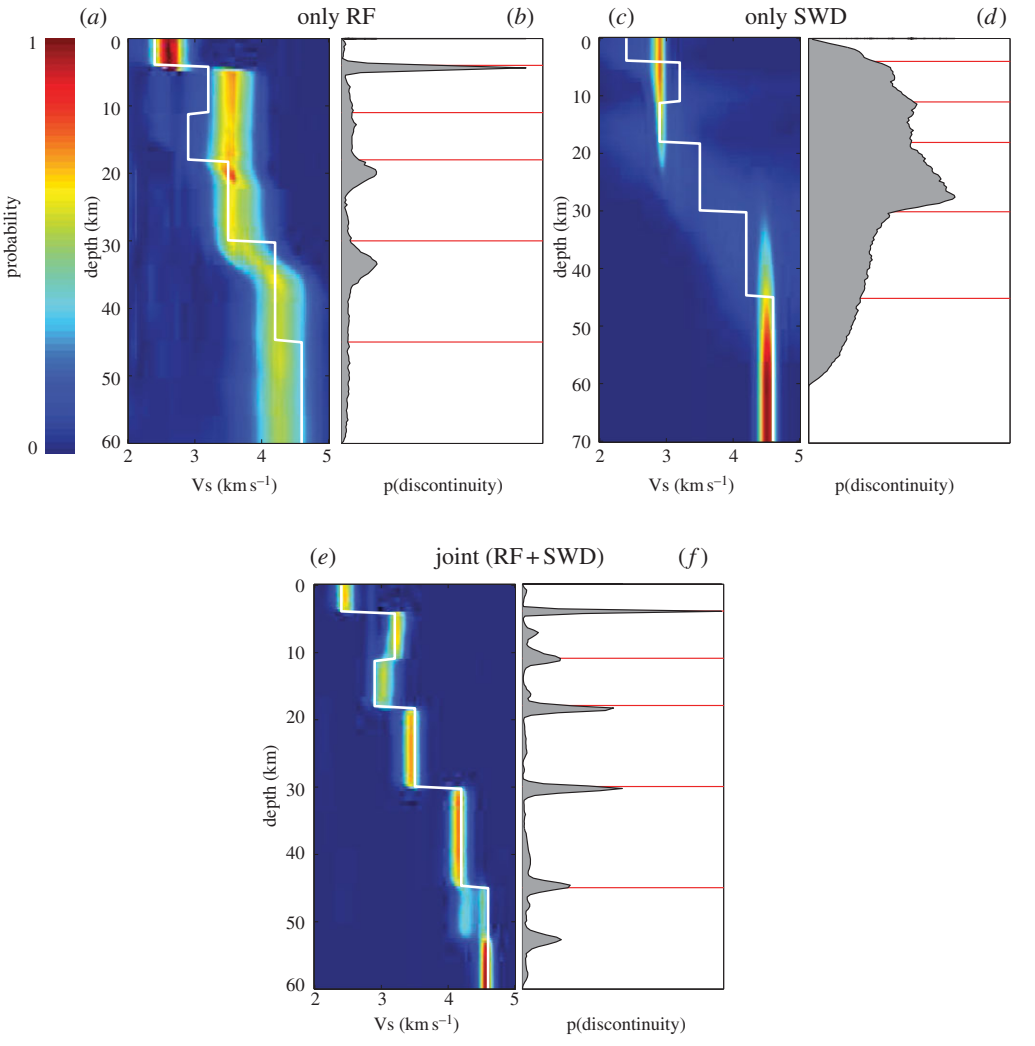

Figure 6. Density of shear wave profile as a function of depth for all models in the posterior ensemble produced by transdimensional sampling. Light colours indicate higher probability, dark colours lower. The white curve is the true model. Plots (*a*,*b*) show results with only receiver function data, (*c*,*d*) for only surface wave dispersion data and (*e*,*f*) when both are included. In all cases, data noise parameters are allowed to vary. (*b*,*d*,*f*) The densities of the interface positions from the models in the ensemble. The combination of the two classes of data provides much improved constraints on both velocities and interface positions when they are jointly inverted with the hierarchical scheme. (Online version in colour.)

number of layers, n_L, and the three noise parameters (σ_s, σ_r, r). To illustrate the algorithm, flat priors were chosen on all parameters. Details of all sampling algorithms can be found in the study of Bodin *et al*. [31].

Figure 6 shows the results for the velocity depth profile for three separate cases. The first is obtained by inversion of receiver function data only, i.e. treating noise parameters σ_r and r in equation (3.5) as unknowns. The second is obtained with surface wave dispersion data only, and hence a single noise parameter σ_s (3.4) is unknown. The third is obtained using the datasets combined, treating all three noise parameters as variables. All three cases are hierarchical in this sense. Figure 6*a* shows the densities of the models in the posterior ensemble when only receiver functions are used. Light colours represent high posterior density and dark colours lower density. Figure 6*b* shows the density of all interface values superimposed. The thin line is the true model in

all cases. While there is some sensitivity to velocity and interface position when receiver function data are used, the position of interfaces is poorly defined, as are velocities within layers. For surface wave dispersion data, the interface position is almost completely unresolved, as seen in figure 6*d*, and so is the velocity jump across interfaces. In contrast, the average velocities appear well constrained (figure 6*c*). It is only when the two are combined (figure 6*e*,*f*) that absolute velocity, velocity jumps across interfaces as well as the positions of interfaces are all well constrained. This is because the noise parameters control the weighting between data types in the joint inversion, and so it is only by allowing them to vary simultaneously that information can usefully be extracted.

The joint inversion of different classes of data is a common problem in the geosciences and arises in many other fields. Although this example specifically concerns a seismological application, the same mathematical set-up occurs in almost all cases where two or more independent data types are used to constrain a common set of parameters. The transdimensional and hierarchical sampling framework may well have applications more broadly.

4. Embedding transdimensional sampling

A common criticism of MCMC sampling methods is that large numbers of likelihood evaluations are required for convergence to the posterior PDF. Each likelihood evaluation requires solution of the forward problem, i.e. predictions to be made from the model $\mathbf{m}$, represented, for example, as $g_{\mathrm{r}}(\mathbf{m})$ and $g_{\mathrm{s}}(\mathbf{m})$ in equation (3.3). Even though efficient sampling schemes have been the subject of much research, this is a potential weakness. In particular cases, the computational burden can be reduced by embedding the MCMC algorithm within a linearized framework. Figure 7 illustrates the general idea. The steps in this iterative sequence are similar to that in a gradient-based optimization algorithm. In the main loop of figure 7, the forward problem is linearized about the reference model, $m_i(\mathbf{x})$. Details will vary with application but in general the nonlinear dependence of data on the model is represented by

$$\mathbf{d} = g(m(\mathbf{x})), \tag{4.1}$$

which can be linearized in the form

$$\mathbf{d} \approx \mathbf{d}_i + \delta\mathbf{d}(\mathbf{m}_k), \tag{4.2}$$

where $\mathbf{m}_k$ are perturbations to basis function coefficients of the reference model $m_i(\mathbf{x})$, $\mathbf{d}_i = g(m_i(\mathbf{x}))$ are the data predictions from the reference model in the ith iteration, and subscript k is the model dimension. For examples of inverse problems linearized in this way, see Aster *et al.* [40]. In the linearized scheme, the kernels of the forward problem are calculated once in the outer loop. In seismic travel time tomography [28], this corresponds to solving geometric ray tracing equations between source and receiver through the model $m_i(\mathbf{x})$ and storing the rays [42]. Within the MCMC step (dashed border in figure 7) perturbations to the data predictions, $\delta\mathbf{d}(\mathbf{m}_k)$, for each new model $\mathbf{m}_k$ can often be efficiently determined by using the kernels determined in the reference model $m_i(\mathbf{x})$. In the tomography case, this corresponds to integrating the model perturbation $\mathbf{m}_k$ along known seismic rays.

Once the MCMC algorithm has converged and a posterior ensemble obtained, a single model is constructed from the ensemble 'in some way' to be the next reference model. An example is to take a pointwise spatial average of each model in the ensemble (as was done in the one-dimensional example in figure 6). The key point is that the actual forward problem is only solved once per iteration in the outer loop, but avoided in the MCMC step. This approach was used by Bodin & Sambridge [28] to significantly speed up transdimensional sampling in two-dimensional seismic travel time tomography. In that case, 1156 travel time data measurements were inverted by completing three iterations of the outer loop each with 1.2×10^6 steps of the MCMC algorithm.

In principle, embedded MCMC could be applied to any linearizable inverse problem and thereby avoid computationally expensive forward solutions of every new model generated along the Markov chain. However, success is dependent on the convergence of the linearized

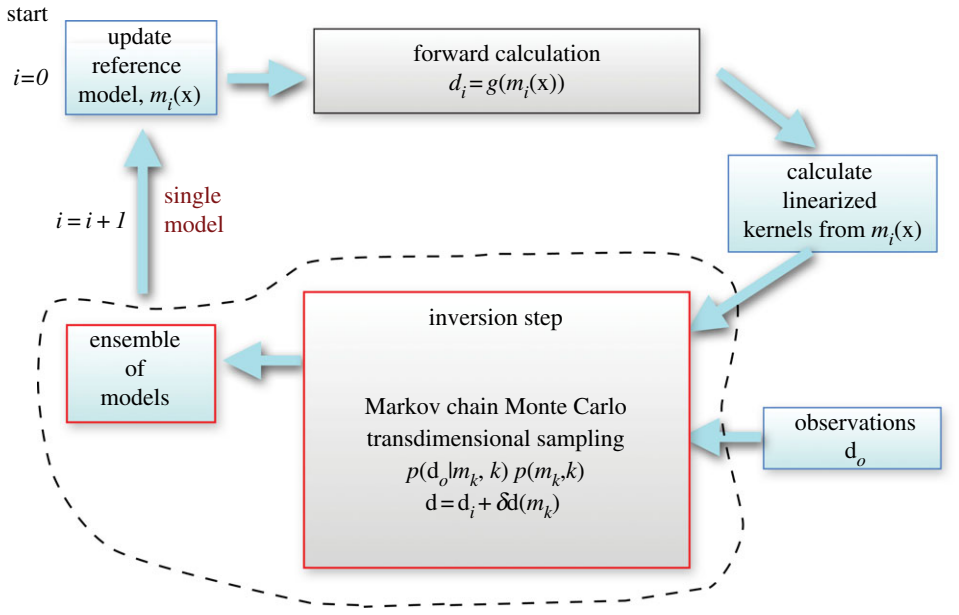

Figure 7. Schematic showing how the transdimensional sampling algorithm may be embedded within a linearized iterative inversion scheme to reduce computational overheads. Sampling is performed within the dashed region where an ensemble of variable parametrization models are generated. Computational time is saved by avoiding solving the forward problem for each model in the ensemble, and instead adopting a linearized approximate solution. A single model, $m_i(x)$, is produced from the ensemble and passed to the outer loop where the forward problem is again solved and a new linearization performed. (Online version in colour.)

scheme. It may be of interest to contrast the embedded MCMC algorithm with that of a standard optimization scheme which iteratively updates a single 'best-so-far' model. The transdimensional sampling generates a complete ensemble and provides a single 'average' model to the next iteration, whereas a gradient-based algorithm would update a single model using predetermined smoothing and other regularization operators [4]. The appeal of embedded MCMC is that it can take advantage of the efficiencies of linearization while using the data to constrain all classes of model parameter, including the parametrization and data noise parameters. In this context, the transdimensional sampling can be viewed as a sophisticated model update scheme (including regularization) with the ability to limit features in the model to those required by the data.

5. A transdimensional software library

An open source software library which implements a range of transdimensional sampling algorithms has been developed and is available for distribution. The suite is divided into four parts all built into a single C source code library. The first implements one-dimensional regression problems of the type described in the first example above. The number and location of partitions as well as the order of the polynomial within each partition may be variable. Hierarchical schemes are implemented for determining noise parameters. This will be suitable for a range of regression problems encountered in the sciences. The second part deals with the same style of one-dimensional spatial model only with a user-supplied forward problem as described in the seismic example in figure 6. The third and fourth parts deal with similar two-dimensional spatial models with and without a forward problem, i.e. regression of scattered data over two parameters and inversion for a two-dimensional field where a user-supplied forward problem is needed to make predictions of observables. An example of the later case is Bodin *et al*. [31]. Interfaces to the C-library using the Python, R and Fortran languages are under construction. The hope is that these tools will find uses across a range of inverse problems in the physical sciences.

6. Concluding remarks

We have outlined the concepts behind transdimensional MCMC sampling and presented some illustrative examples of its use to invert for one-dimensional models (e.g. functions of depth or time). Extensions to two- and three-dimensional spatial models are also possible. The key motivation for transdimensional inversion techniques is a desire to avoid making restrictive assumptions concerning details of parametrization and data noise, and instead use the data themselves to constrain these properties. In contrast to optimization approaches, the sampling-based algorithm generates an ensemble of candidate solutions, none of which is necessarily any more meaningful than another (although clearly can fit the data differently). Properties of the ensemble as a whole are used to infer information about the unknown spatial model. In this paper, we have focused on examining density plots and deriving average models for inspection, although many other properties can be calculated from ensembles, such as credible intervals, marginal PDFs and trade-off information between parameters [36]. Examples of each of these can be found in the reference cited. These techniques are increasingly finding applications in the geosciences but could equally well be applied to similar problems elsewhere.

We are grateful to Andrew Curtis and an anonymous reviewer for numerous constructive criticisms of an earlier draft of this manuscript. Aspects of research reported here were supported under the Australian Research Council Discovery grant scheme (project no. DP110102098). Calculations were performed on the Terrawulf II cluster, a computational facility supported through the AuScope Australian Geophysical Observing System (AGOS). Auscope Ltd is funded under the National Collaborative Research Infrastructure Strategy (NCRIS) and the Education Investment Fund (EIF3), both Australian Commonwealth Government programmes. Software construction was possible with support from the AuScope inversion laboratory. Rhys Hawkins of the Australian National University Supercomputing Facility has contributed to the *ilab* software library, which implements algorithms described in this paper and is available from M.S.

References

1. Sambridge M, Faletič R. 2003 Adaptive whole earth tomography. *Geochem. Geophys. Geosyst.* **4**, 1022. (doi:10.1029/2001GC000213)
2. Debayle E, Sambridge M. 2004 Inversion of massive surface wave data sets: model construction and resolution assessment. *J. Geophys. Res.* **109**, B02316. (doi:10.1029/2003JB002652)
3. Okabe A, Boots B, Sugihara K. 1992 *Spatial tessellations: concepts and applications of Voronoi diagrams*. Chichester, UK: John Wiley & Sons.
4. Rawlinson N, Sambridge M. 2003 Seismic traveltime tomography of the crust and lithosphere. *Adv. Geophys.* **46**, 81–198. (doi:10.1016/S0065-2687(03)46002-0)
5. Hu W, Abubakar A, Habashy T. 2009 Simultaneous multifrequency inversion of full-waveform seismic data. *Geophysics* **74**, R1–R14. (doi:10.1190/1.3073002)
6. Loris I, Nolet G, Daubechies I, Dahlen F. 2007 Tomographic inversion using 1-norm regularization of wavelet coefficients. *Geophys. J. Int.* **170**, 359–370. (doi:10.1111/j.1365-246X.2007.03409.x)
7. Loris I, Douma H, Nolet G, Daubechies I, Regone C. 2010 Nonlinear regularization techniques for seismic tomography. *J. Comp. Phys.* **229**, 890–905. (doi:10.1016/j.jcp.2009.10.020)
8. Mackay DJC. 2003 *Information theory, inference, and learning algorithms*. Cambridge, UK: Cambridge University Press.
9. Denison DGT, Holmes C, Mallick B, Smith AFM. 2002 *Bayesian methods for nonlinear classification and regression*. Hoboken, NJ: John Wiley & Sons.
10. Scales JA, Snieder R. 1997 To Bayes or not to Bayes. *Geophysics* **62**, 1045–1046.
11. McGrayne SB. 1987 *The theory that would not die: how Bayes' rule cracked the enigma code, hunted down Russian submarines, and emerged triumphant from two centuries of controversy*. New Haven, CT: Yale University Press.
12. Bayes T. 1763 An essay towards solving a problem in the doctrine of chances. *Phil. Trans. R. Soc. Lond.* **53**, 370–418. (doi:10.1098/rstl.1763.0053) (Reprinted, with biographical note by G. A. Barnard, 1958, *Biometrika* **45**, 293–315).

13. Sambridge M, Gallagher K, Jackson A, Rickwood P. 2006 Trans-dimensional inverse problems, model comparison and the evidence. *Geophys. J. Int.* **167**, 528–542. (doi:10.1111/j.1365-246X.2006.03155.x)
14. Gilks W, Richardson S, Spiegelhalter D. 1996 *Markov chain Monte Carlo in practice*. London, UK: Chapman & Hall.
15. Green PJ. 1995 Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711–732. (doi:10.1093/biomet/82.4.711)
16. Geyer CJ, Møller J. 1994 Simulation procedures and likelihood inference for spatial point processes. *Scand. J. Stat.* **21**, 359–373.
17. Malinverno A. 2002 Parsimonious Bayesian Markov chain Monte Carlo inversion in a nonlinear geophysical problem. *Geophys. J. Int.* **151** 675–688. (doi:10.1046/j.1365-246X.2002.01847.x)
18. Malinverno A, Briggs VA. 2004 Expanded uncertainty quantification in inverse problems: hierarchical Bayes and empirical Bayes. *Geophysics* **69**, 1005–1016. (doi:10.1190/1.1778243)
19. Stephenson J, Gallagher K, Holmes CC. 2006 Low temperature thermochronology and strategies for multiple samples 2: partition modelling for 2D/3D distributions with discontinuites. *Earth Planet. Sci. Lett.* **241**, 557–570. (doi:10.1016/j.epsl.2005.11.027)
20. Gallagher K. 2012 Transdimensional inverse thermal history modelling for quantitative thermochronology. *J. Geophys. Res.* **117**, B02408. (doi:10.1029/2011JB008825)
21. Charvin K, Gallagher K, Hampson G, Labourdette R. 2009 A Bayesian approach to inverse modelling of stratigraphy, part 1: method. *Basin Res.* **21**, 5–25. (doi:10.1111/j.1365-2117.2008.00369.x)
22. Reading AM, Bodin T, Sambridge M, Howe S, Roach M. 2010 Down the borehole but outside the box: innovative approaches to wireline log data interpretation. In *Conf. Handbook: 21st Int. Geophysical Conf. & Exhibition, Future Discoveries are in our Hands, Sydney, Australia, 22–24 August 2010*, vol. 147, pp. 1–4. Perth, WA: Australian Society of Exploration Geophysics.
23. Hopcroft P, Gallagher K, Pain C. 2007 Inference of past climate from borehole temperature data using Bayesian reversible jump Markov chain Monte Carlo. *Geophys. J. Int.* **171**, 1430–1439. (doi:10.1111/j.1365-246X.2007.03596.x)
24. Hopcroft P, Gallagher K, Pain C. 2009 A Bayesian partition modelling approach to resolve spatial variability in climate records from borehole temperature inversion. *Geophys. J. Int.* **178**, 651–666. (doi:10.1111/j.1365-246X.2009.04192.x)
25. Jasra A, Stephens D, Gallagher K, Holmes C. 2006 Bayesian mixture modelling in geochemistry via Markov chain Monte Carlo. *Math. Geol.* **38**, 269–300. (doi:10.1007/s11004-005-9109-3)
26. Malinverno A, Leaney WS. 2005 Monte Carlo Bayesian look-ahead inversion of walkaway vertical seismic profiles. *Geophys. Prosp.* **53**, 689–703. (doi:10.1111/j.1365-2478.2005.00496.x)
27. Bodin T, Sambridge M, Gallagher K. 2009 A self-parameterising partition model approach to tomographic inverse problems. *Inverse Probl.* **25**, 055009. (doi:10.1088/0266-5611/25/5/055009)
28. Bodin T, Sambridge M. 2009 Seismic tomography with the reversible jump algorithm. *Geophys. J. Int.* **178**, 1411–1436. (doi:10.1111/j.1365-246X.2009.04226.x)
29. Bodin T, Sambridge M, Rawlinson N, Arroucau P. 2012 Transdimensional tomography with unknown data noise. *Geophys. J. Int.* **189**, 1536–1556. (doi:10.1111/j.1365-246X.2012.05414.x)
30. Piana-Agostinetti N, Malinverno A. 2010 Receiver function inversion by transdimensional Monte Carlo sampling. *Geophys. J. Int.* **181**, 858–872. (doi:10.1111/j.1365-246X.2010.04530.x)
31. Bodin T, Sambridge M, Tkalčić H, Arroucau P, Gallagher K, Rawlinson N. 2012 Transdimensional inversion of receiver functions and surface wave dispersion. *J. Geophys. Res.* **178**, B02301. (doi:10.1029/2011JB008560)
32. Dettmer J, Dosso SE, Holland C. 2010 Trans-dimensional geoacoustic inversion. *J. Acoust. Soc. Am.* **128**, 3393–4005. (doi:10.1121/1.3500674)
33. Minsley B. 2011 A trans-dimensional Bayesian Markov chain Monte Carlo algorithm for model assessment using frequency-domain electromagnetic data. *Geophys. J. Int.* **187**, 252–272. (doi:10.1111/j.1365-246X.2011.05165.x)
34. Dettmer J, Molnar S, Steininger G, Dosso SE, Cassidy JF. 2012 Trans-dimensional inversion of microtremor array dispersion data with hierarchical autoregressive error models. *Geophys. J. Int.* **188**, 719–734. (doi:10.1111/j.1365-246X.2011.05302.x)
35. Gallagher K, Charvin K, Nielsen S, Sambridge M, Stephenson J. 2009 Markov chain Monte Carlo (McMC) sampling methods to determine optimal models, model resolution and

model choice for earth science problems. *J. Mar. Petrol. Geol.* **26**, 525–535. (doi:10.1016/j.marpetgeo.2009.01.003)
36. Gallagher K, Bodin T, Sambridge M, Weiss D, Kylander M, Large D. 2011 Inference of abrupt changes in noisy geochemical records using transdimensional changepoint models. *Earth Planet. Sci. Lett.* **311**, 182–194. (doi:10.1016/j.epsl.2011.09.015)
37. Brooks SP, Giudici P, Roberts GO. 2003 Efficient construction of reversible jump MCMC proposal distributions (with discussion). *J. Roy. Stat. Soc. B* **65**, 3–55. (doi:10.1111/1467-9868.03711)
38. Al-Awadhi F, Hurn M, Jennison C. 2004 Improving the acceptance rate of reversible jump MCMC proposals. *Stat. Probab. Lett.* **69**, 189–198. (doi:10.1016/j.spl.2004.06.025)
39. Curtis A, Lomax A. 2001 Prior information, sampling distributions and the curse of dimensionality. *Geophysics* **66**, 372–378. (doi:10.1190/1.1444928)
40. Aster R, Borchers B, Thurber CH. 2005 *Parameter estimation and inverse problems*, vol. 90. International Geophysics Series. Amsterdam, The Netherlands: Elsevier.
41. Ammon CJ, Randall GE, Zandt G. 1990 On the nonuniqueness of receiver function inversions. *J. Geophys. Res.* **95**, 15 303–15 318. (doi:10.1029/JB095iB10p15303)
42. Sambridge MS. 1990 Non-linear arrival time inversion: constraining velocity anomalies by seeking smooth models in 3-D. *Geophys. J. Int.* **102**, 653–677. (doi:10.1111/j.1365-246X.1990.tb04588.x)