Geophys. J. Int. (2009) **178,** 1411–1436 doi: 10.1111/j.1365-246X.2009.04226.x

Seismic tomography with the reversible jump algorithm

Thomas Bodin and Malcolm Sambridge

Research School of Earth Sciences, Australian National University, Canberra ACT 0200, *Australia. E-mail: thomas.bodin@anu.edu.au*

Accepted 2009 April 24. Received 2009 April 21; in original form 2009 February 6

SUMMARY

The reversible jump algorithm is a statistical method for Bayesian inference with a variable number of unknowns. Here, we apply this method to the seismic tomography problem. The approach lets us consider the issue of model parametrization (i.e. the way of discretizing the velocity field) as part of the inversion process. The model is parametrized using Voronoi cells with mobile geometry and number. The size, position and shape of the cells defining the velocity model are directly determined by the data. The inverse problem is tackled within a Bayesian framework and explicit regularization of model parameters is not required. The mobile position and number of cells means that global damping procedures, controlled by an optimal regularization parameter, are avoided. Many velocity models with variable numbers of cells are generated via a transdimensional Markov chain and information is extracted from the ensemble as a whole. As an aid to interpretation we visualize the expected earth model that is obtained via Monte Carlo integration in a straightforward manner. The procedure is particularly adept at imaging rapid changes or discontinuities in wave speed. While each velocity model in the final ensemble consists of many discontinuities at cell boundaries, these are smoothed out in the averaged ensemble solution while those required by the data are reinforced. The ensemble of models can also be used to produce uncertainty estimates and experiments with synthetic data suggest that they represent actual uncertainty surprisingly well. We use the fast marching method in order to iteratively update the ray geometry and account for the non-linearity of the problem. The method is tested here with synthetic data in a 2-D application and compared with a subspace method that is a more standard matrix-based inversion scheme. Preliminary results illustrate the advantages of the reversible jump algorithm. A real data example is also shown where a tomographic image of Rayleigh wave group velocity for the Australian continent is constructed together with uncertainty estimates.

Key words: Inverse theory; Probability distribution; Tomography; Seismic tomography; Computational seismology; Australia.

1 INTRODUCTION

Tomography is a well-established tool for investigating the internal structure and composition of our planet and has been actively developed by the seismology community for 30 yr. In most cases, Earth models are parametrized with basis functions of uniform local cells in 2-D or 3-D whose size and shapes are fixed in advance. As is well known, the choice of cell size is a compromise between model resolution and model uncertainty. If the cells are large, independent information can be integrated to give a mean velocity value that is not biased by the noise in the data. The uncertainty on the estimated velocity will be small at the expense of the resolution, which in turn is directly linked to the size of the cells. As the cells become smaller, the noise in the data maps into large uncertainty in the model parameters and, quickly, the solution will become non-unique. Regularization is often imposed to obtain a single model that biases linearized uncertainty estimates.

The information obtained in seismic tomography strongly depends on the location of the sources (mostly at plate boundaries) and the positions of the receivers that are not evenly distributed on the Earth surface. This leads to having an irregular spatial distribution of the information, that is, some regions are traversed by a lot of seismic rays whereas other regions are left with poor ray coverage. It is self evident that uneven ray path sampling leads to limited resolution in regions of poor data coverage. In tomography, the usual ways of dealing with ill-constrained parts of a model are to apply some spatial smoothing, norm damping or simply to coarsen the parametrization, for example, increase cell sizes. Traditionally these forms of 'regularization' have been applied uniformly across the entire model, which raises the possibility that, while the ill-constrained regions are being damped, the well-constrained regions are being oversmoothed and hence information may be lost.

In order to deal with the irregular distribution of information, some seismologists have used irregular meshes (see Sambridge & Rawlinson 2005, for a review). In a study to image the *P*-wave velocity structure beneath Papua New Guinea, Abers & Roecker (1991) joined fine-scale regular 3-D blocks together to form larger non-uniform cells. Non-uniform-sized rectangular 3-D blocks were also used by Fukao *et al.* (1992) and Spakman & Bijwaard (1998) to account for uneven ray path sampling in regional models. In a cross-borehole tomographic problem, Curtis & Snieder (1997) used a genetic algorithm to arrange the vertices of a triangular parametrization so as to minimisze the condition number of the resulting tomographic system of equations. Sambridge *et al.* (1995a), Sambridge & Gudmundsson (1998) and Zhang & Thurber (2005) proposed the use of structures from computational geometry, Delaunay tetrahedra and Voronoi polyhedra for tomographic problems. These were used as the basis of completely unstructured meshes, for example, not based on a cubic or other regular grid (Voronoi 1908; Okabe *et al.* 1992), and adapted to the resolving power of the data. The same geometric structures were used by Sambridge & Faletic (2003) to adapt the spatial mesh to the data distribution during the inversion itself. Nolet & Montelli (2005) proposed an algorithm for spacing mesh vertices in the Earth so that their density mirrored that of the local resolving length of the data.

Most of these studies have a fixed number of unknowns decided before hand, for example, the number of layers or cells. A range of statistical techniques has been developed for judging whether the choice of the model dimension is warranted by the data, for example, *F*-tests (Aster *et al.* 2005). But the idea of determining the model dimension during the inversion, that is treating the number of parameters as a parameter itself, has received relatively little attention. In the paper presented here, we propose to invert for the cell geometries and the seismic velocities at the same time. That is, the parametrization will be directly determined by the data and treated as an actual unknown to be inverted for by the tomographic algorithm. At first glance this may sound like an unrealistic prospect. Experience shows that it is always easier to fit data with more unknowns, so within an optimization framework (where we seek best-fit models) there would seem to be little to prevent an algorithm introducing ever more detail into a model. As will be shown in this paper, however, this is not the case within a Bayesian sampling framework. It turns out that high-dimensional (many cell) models are naturally discouraged. Statisticians have considered the problem of a variable number of unknowns, for more than 10 yr. As a consequence Markov chain Monte Carlo (MCMC) methods that admit transitions between states of differing dimension have been actively developed. This new family of sampling algorithms have recently been termed transdimensional Markov chains [for a review, see Sisson (2005)]. At present, the reversible jump algorithm (rj-MCMC) of Green (1995) (see also Green & Mira 2001; Sambridge *et al.* 2006) is the most common MCMC tool for exploring variable dimension statistical models. To date, the majority of areas in which transdimensional Markov chain have been applied tend to be computationally or biologically related. Overall, one in every five citations of Green (1995) can be broadly classified as genetics-based research. The reversible jump algorithm was first applied in the geophysical literature by Malinverno & Leaney (2000) to the inversion of zero-offset vertical seismic profiles. [For a more complete treatment see Malinverno & Leaney (2005).] Subsequent paper was by Malinverno (2002) who applied it to electrical resistivity sounding.

In this paper, we develop an entirely new approach to the tomography problem based on the reversible jump algorithm. The scheme we propose here is closely related to a process known as Bayesian partition modelling (e.g. Denison *et al.* 2002a,b), which is a statistical analysis tool for non-linear classification and regression. Partition modelling has been applied successfully in disease mapping (e.g. Denison & Holmes 2001) and more recently, introduced into the Earth Sciences for applications in geostatistics (Stephenson *et al.* 2004), thermochronology (e.g. Stephenson *et al.* 2006) and palaeoclimate inference (e.g. Hopcroft *et al.* 2007).

Partition modelling is a way of using variable dimension irregular meshes. The velocity (or slowness) field is partitioned by a variable number of non-overlapping regions defined by a Voronoi tessellation as in Fig. 1 (Okabe *et al.* 1992). The parametrization is defined by a discrete set of points (or Voronoi nuclei) and each region (Voronoi cell) encloses all the points of the space that are closer to its nucleus than to any other Voronoi nucleus. That is, the number and the position of the cells defining the geometry of the velocity field, as well as the velocity field itself are unknowns in the inversion. The inversion is carried out with a fully non-linear parameter search method based on a transdimensional Markov chain.

The method presented here was first developed by Bodin *et al.* (2009) with a fixed number of partitions (i.e. fixed number of unknowns) in the general context of straight ray (i.e. linear) tomographic problems. In this paper, we extend the method to transdimensionality and apply

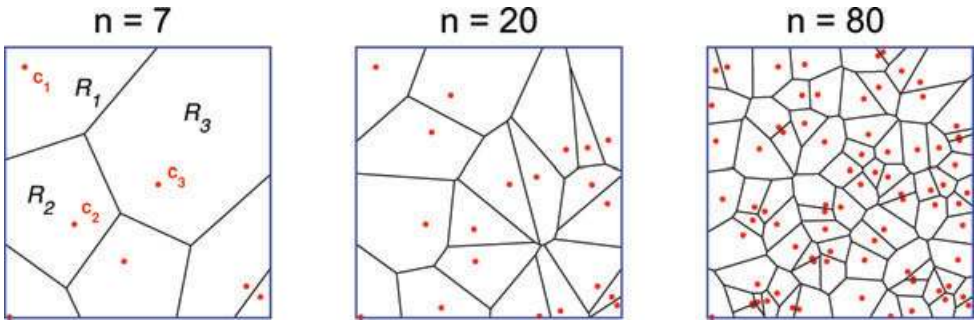


Figure 1. Examples of Voronoi diagrams (black), which form a set of irregular cells that partition the plane. Each cell contains the part of the plane closest to its nucleus (red dot) and so the shape of the parametrization is entirely defined by the location of the nuclei. The three panels show 7, 20 and 80 randomly distributed nuclei, respectively. As the number and position of the nuclei changes the Voronoi diagram corresponds to a multiscale parametrization of the velocity field.

it to the non-linear seismic tomography problem with iterative ray updates. We focus here on a 2-D problem but all components readily extend to 3-D without changes, but (of course) at increased computational cost. The inversion of arrival time data to infer a 2-D seismic velocity field is a general problem and has been addressed in a large number of different situations, such as seismic surface wave tomography (e.g. Nolet & Panza 1976; Friederich 1998; Prindle & Tanimoto 2006) or cross-hole seismic body wave tomography (e.g. Ivansson 1986).

In the next section, we give a detailed presentation of the Bayesian seismic tomography methodology using the partitioned parametrization. We show how it can be used iteratively to perform a non-linear tomography including ray bending. In Section 3, we present results of synthetic experiments in a situation where the ray coverage is far from ideal in order to compare our approach to standard methods that use regular (ideal) parametrizations. We use data contaminated with random noise in order to test the ability of the method to infer model uncertainty. In Section 4, the method has been used with ambient noise data to infer a tomographic image of Rayleigh wave group velocity for the Australian continent. In the final section, we summarize the main results, and outline directions for further study.

2 METHOD

2.1 Parametrization with Voronoi cells

As shown in Fig. 1, the 2-D seismic velocity field is discretized by a set of Voronoi polygons. Given a set $\mathbf{c}$ of n nuclei in the 2-D plane ($\mathbf{c} = \{\mathbf{c}_1, \ldots, \mathbf{c}_n\}$ where $\mathbf{c}_i \in \Re^2$), the Voronoi partition consists of n non-overlapping regions $R_1, R_2, \ldots, R_n$ such that the points within R_i are closer to $\mathbf{c}_i$ than any of the other $\mathbf{c}_j$ ($j \neq i$). To make notation clear, we have marked three nuclei $\{\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3\}$ together with corresponding Voronoi partitions R_1, R_2 and R_3 in Fig. 1. Perpendicular bisectors of each pair of neighbouring nuclei form the cell boundaries. Each cell is therefore determined by the two coordinates of its nucleus $\mathbf{c}_i$ and by a constant velocity value v_i. We represent the combined set as $\mathbf{m} = (\mathbf{c}, \mathbf{v})$, where $\mathbf{v}$ is the vector of the velocity values assigned to each cell ($\mathbf{v} = (v_1, \ldots, v_n)$ where $v_i \in \Re$). The number of unknowns, that is, the dimension of the model $\mathbf{m}$, is then $3n$. Here we use only the simplest possible representation of velocity within each partition, that is, a constant. Higher order polynomials are possible, for example, a linear gradient or quadratic, which would require additional unknowns for each cell.

During the inversion, the number and the position of the nuclei is variable so the cells can take different sizes and shapes (see Fig. 1). We will see that in the transdimensional approach, this dynamic parametrization will adapt to the spatial variability in the information provided by the data.

2.2 An iterative linearized approach

Given a complete description of a velocity field, the theory of seismic wave propagation allows us to predict traveltimes of direct phases from a point source to a receiver and compare them to observations. Here we consider the high-frequency approximation case and use ray theory (Cerveny *et al.* 1977; Cerveny & Brown 2003). We employ the fast marching method (FMM) (Sethian & Popovici 1999; Rawlinson & Sambridge 2004) to calculate traveltimes and ray paths in a laterally heterogeneous 2-D medium.

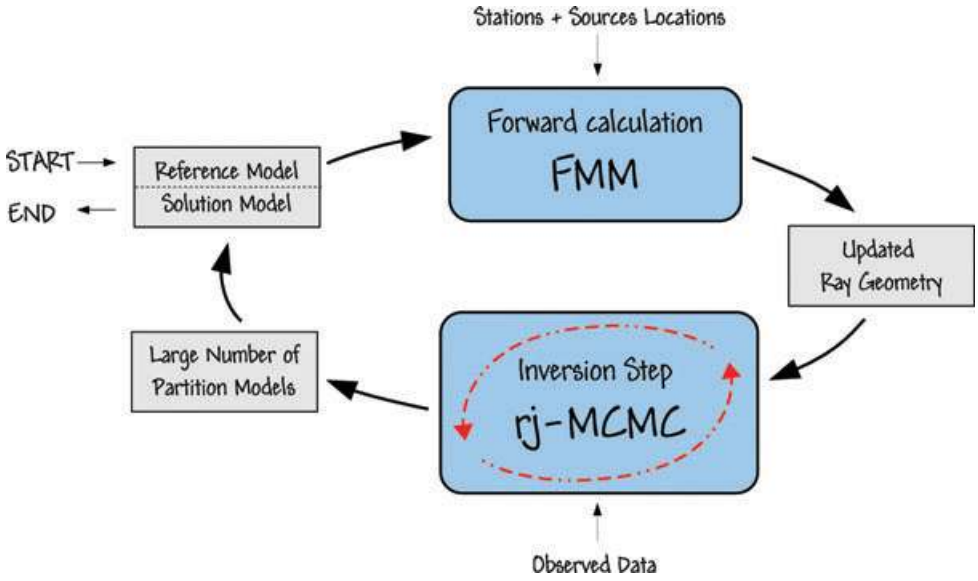


Figure 2. The transdimensional MCMC sampler is used in conjunction with fast marching eikonal solver to build an iterative linearized solution method. The inputs are an initial continuous reference velocity model, the stations and sources locations and the observed traveltimes. At each iteration, the Markov chain in the 'inner loop' (dashed red arrows) produces an ensemble of potential solutions that are spatially averaged to produce a reference model for the next iteration of the outer loop (black arrows). Only in the outer loop are ray paths updated.

Since the ray paths are dependent on the velocity model, the tomographic problem is non-linear. The development proposed here takes into account non-linearity by iteratively updating ray paths (e.g. Gorbatov *et al.* 2001). Our scheme can be described using two loops as depicted in Fig. 2. The outer loop (solid black arrows in Fig. 2) is similar to many previous tomographic schemes. At each iteration, ray paths and traveltimes are determined in the current velocity model using fast marching. (Prior to the first iteration this is a laterally homogeneous reference model.) At each successive iteration, the outer loop reference model is updated by spatially averaging the entire ensemble of models produced in the inner loop. Each inner loop model is constructed from Voronoi cells (as in Fig. 1) and will have discontinuities throughout the velocity field, but the ensemble average tends to be spatially smooth with continuously varying gradients.

The outer loop is no different from any linearized tomographic inversion scheme. The difference lies in the model update procedure within the inner loop. Rather than using a matrix inversion approach to perturb a single model, we use a reversible jump MCMC algorithm (Green 1995) to produce a chain of partitioned velocity models. The term 'chain' is used because each model generated is not independent but part of a Markov chain. With a sufficiently large number of models and possibly after some thinning of the chain, that is, retaining only every 10th or 100th model, we obtain an ensemble of solutions. For the next iteration of the outer loop, a continuous reference model is required which is a simple spatial average of each model in the ensemble. (Details of the spatially averaging procedure appear in Section 2.9.)

The Markov chain algorithm within the inner loop (dashed red arrows in Fig. 2) requires calculation of traveltimes for a large number of partitioned velocity models of the type in Fig. 1. Rather than actually calculating rays in each partition model (which would increase computation considerably) we simply integrate along the current reference ray paths using the expression

$$t = \int_{R_0} \frac{1}{v(\mathbf{x})} \mathrm{d}r, \tag{1}$$

where R_0 is the ray path corresponding to the continuous reference model (determined in the outer loop) and $v(\mathbf{x})$ is the velocity field given by the partition model (with constant velocity values in cells). Note that since this step does not involve solution of a linear system of equations there is no need to explicitly linearize the traveltime expression in (1) in terms of velocity, instead we can integrate the reciprocal of the velocity field along the reference ray. The reader will be able to verify, with some simple algebra, that since ray paths are kept fixed (1) is equivalent to the linearization in slowness commonly used in tomographic algorithms, that is, traveltimes given by (1) are the same as those obtained by evaluating

$$t = \int_{R_0} s_o(\mathbf{x})\, \mathrm{d}r + \int_{R_0} \delta s(\mathbf{x})\, \mathrm{d}r, \tag{2}$$

where $v_o(\mathbf{x})$ is the reference velocity field and $s_o(\mathbf{x}) = 1/v_o(\mathbf{x})$. Hence traveltimes given by (1), are equivalent to a first-order accurate slowness perturbation. For fixed velocities in Voronoi cells, the traveltime of the jth ray is then given by

$$t_j = \sum_{i=1}^{n} \frac{L_{ij}}{v_i}, \tag{3}$$

where L_{ij} is the length of ray j across cell i (see Fig. 3) and v_i is the velocity value assigned to cell i . If the ray j does not pass through cell i, then $L_{ij} = 0$. The ray lengths in the cells are calculated by sampling along rays at a predetermined step length (red points in Fig. 3) and by finding the cell containing the midpoint of the current ray segment. Point location within 2-D Voronoi cells is efficiently implemented with the scheme described in Sambridge & Gudmundsson (1998). Note the curvature of the ray in Fig. 3 is due to gradients in the continuous reference model. Overall the outer loop of the algorithm is quite standard. The novel features lie in the inner loop (i.e the inversion step) where both wave speeds and parametrization are updated with the reversible jump algorithm within a Bayesian framework. In the next section, we briefly introduce this approach and describe the Markov chain algorithm in detail.

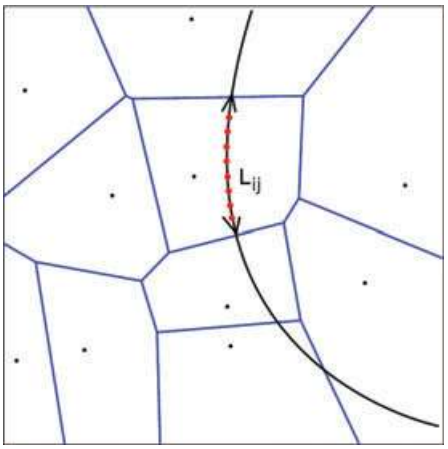


Figure 3. A seismic ray (thick black line) is bent according to a reference velocity model. For a given partition model, the estimated traveltime t_j for the jth ray is simply computed by integrating the inverse of the cell velocity v_i along the ray path. That is, using the length L_{ij} of the ray across each cell.

2.3 A Bayesian framework

In a Bayesian approach all information is represented by probability density functions (standard references are Box & Tiao 1973; Smith 1991; Gelman *et al.* 2004). Geophysical applications of Bayesian inference are described in Tarantola & Valette (1982), Duijndam (1988a,b) and Mosegaard & Tarantola (1995). The aim of Bayesian inference is to quantify the *a posteriori* probability distribution (or posterior distribution), which is the probability density of the model parameters given the observed data (Smith 1991), written as $p(\mathbf{m}|\mathbf{d}_{\text{obs}})$. Each individual model has an associated probability, conditional on the data. Therefore, the posterior distribution is a function of the unknowns. Here, the number of unknowns is not fixed so the posterior is defined in a number of spaces with different dimensions. If the model is defined by *3n* unknowns, the posterior will be a function of *3n* variables (where *n* is the number of Voronoi nuclei). This transdimensional probability distribution is taken as the complete solution of the inverse problem. In practice one tends to use computational methods to generate samples from the posterior distribution, that is, an ensemble of vectors $\mathbf{m}[= (\mathbf{c}, \mathbf{v})]$ whose density reflects that of the posterior distribution.

To define the posterior distribution, we use Bayes' theorem (Bayes 1763) that combines prior information on the model with the observed data to give

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}$$

$$p(\mathbf{m}\,|\,\mathbf{d}_{\text{obs}}) = \frac{p(\mathbf{d}_{\text{obs}}\,|\,\mathbf{m})p(\mathbf{m})}{p(\mathbf{d}_{\text{obs}})}, \tag{4}$$

where $a|b$ means a given, or conditional on, b. It is the probability of having a when b is fixed. $\mathbf{m}$ is the vector of the model parameters and $\mathbf{d}_{\text{obs}}$ is a vector defined by the set of observed data. The term $p(\mathbf{d}_{\text{obs}}|\mathbf{m})$ is the likelihood function, which is the probability of observing the measured data given a particular model. $p(\mathbf{m})$ is the *a priori* probability density of $\mathbf{m}$, that is, what we know about the model $\mathbf{m}$ before measuring the data $\mathbf{d}_{\text{obs}}$. The term, $p(\mathbf{d}_{\text{obs}})$ corresponds to the numerator on the right-hand side integrated over the model space, and can be regarded as constant since it is not a function of any particular model (Sambridge *et al.* 2006). We can therefore write (4) as a proportionality relationship

$$p(\mathbf{m}\,|\,\mathbf{d}_{\text{obs}}) \propto p(\mathbf{d}_{\text{obs}}\,|\,\mathbf{m})p(\mathbf{m}). \tag{5}$$

Hence, the posterior distribution represents how our prior knowledge of the model parameters is updated by the data. Clearly, if the prior and the posterior distributions are the same, then the data add no new information.

From an ensemble of models distributed according to the posterior, it is straightforward to determine special properties such as the best or average model, or to construct marginal probability distributions for individual model parameters. Correlation coefficients between pairs of parameters can also be extracted (Gelman *et al.* 2004).

2.4 The likelihood function

The likelihood function $p(\mathbf{d}_{\text{obs}}|\mathbf{m})$ quantifies how well a given model with a particular set of parameter values can reproduce the observed data. Using a given geometry of rays, the traveltimes are computed for a particular partition model with (3) providing an estimate of the data that would be measured for that model. Here the likelihood is based on a simple least squares misfit, which quantifies the agreement between simulated and observed data.

$$\phi(\mathbf{m}) = \left\| \frac{\mathbf{g}(\mathbf{m}) - \mathbf{d}_{\text{obs}}}{\sigma_{\mathbf{d}}} \right\|^2, \tag{6}$$

where $g(\mathbf{m})$ is the estimated data given by (3) and σ_d^2 is the estimated variance of the data noise (assumed uncorrelated).

It is well known that the least squares misfit function corresponds to a Gaussian likelihood function, that is,

$$p(\mathbf{d}_{\text{obs}}\,|\,\mathbf{m}) \propto \exp\left\{\frac{-\phi(\mathbf{m})}{\mathbf{2}}\right\}. \tag{7}$$

2.5 The prior

In a Bayesian formulation, any prior knowledge on the model can be taken into account, provided that this information can be expressed as a probability distribution $p(\mathbf{m})$ (see Scales & Snieder 1997; Gouveia & Scales 1998, for discussions). A weakness of the Bayesian formulation is that only information that can be expressed as a probability distribution can be used. All inferences from the data are relative to the prior. Here we use a simple uniform prior distribution between a fixed range. Since we have independent parameters of different physical dimension, the prior can be separated into two terms,

$$p(\mathbf{m}) = p(\mathbf{m}\,|\,n)p(n), \tag{8}$$

where $p(n)$ is the prior on the number of partitions. Here we use a uniform distribution over the interval $I = \{n \epsilon N | n_{\min} < n \leq n_{\max}\}$. Hence,

$$p(n) = \begin{cases} 1/(\Delta n) & \text{if} \quad n \epsilon I \\ 0 & \text{otherwise,} \end{cases} \tag{9}$$

where $\Delta n = (n_{\max} - n_{\min})$.

Given a number of cells n, the prior probability distributions for the $3n$ parameters, 2-D Voronoi nuclei and velocities in each cell, are independent from each other, and so can be written in separable form

$$p(\mathbf{m} \,|\, \mathbf{n}) = p(\mathbf{c} \,|\, n) p(\mathbf{v} \,|\, n). \tag{10}$$

Even though in the prior the parametrization variables **c** are independent of the velocity variables **v**, this will not be the case once the data are introduced, and hence we expect significant correlation in the posterior distribution.

For velocity, the prior is specified by a constant value over a defined velocity interval $J = \{v_i \epsilon \Re | V_{\min} < v_i < V_{\max}\}$. Hence we have

$$p(v_i \,|\, n) = \begin{cases} 1/(\Delta v) & \text{if} \quad v_i \epsilon J \\ 0 & \text{otherwise,} \end{cases} \tag{11}$$

where $\Delta v = (V_{\max} - V_{\min})$. Since the velocity in each cell is independent,

$$p(\mathbf{v} \,|\, n) = \prod_{\mathbf{i=1}}^{n} p(\mathbf{v_i} \,|\, n). \tag{12}$$

For mathematical convenience, let us assume (temporarily) that the Voronoi nuclei can only be positioned on an underlying finite grid of nodes defined by $N = n_x \times n_y$ possible positions. For n Voronoi nuclei, there are then $[\frac{N!}{n!(N-n)!}]$ possible configurations on the N possible points of the underlying grid. We give equal probability to each of these configurations, and hence the prior for the nodal positions is given by

$$p(\mathbf{c} \,|\, n) = \left[\frac{N!}{n!(N-n)!} \right]^{-1}. \tag{13}$$

Combining together (9), (11), (12) and (13), the full prior probability density function can be written as

$$p(\mathbf{m}) = \begin{cases} \dfrac{n!(N-n)!}{N!(\Delta v)^n \Delta n} & \text{if} \quad (n \epsilon I \quad \text{and} \quad \forall i \epsilon [1, n], \; v_i \epsilon J) \\ 0 & \text{otherwise.} \end{cases} \tag{14}$$

By combining (7) and (14) in (5), the posterior distribution can be evaluated for any given model **m**. The task is then to generate samples whose density follows (5).

2.6 Principle of the reversible jump Markov chain Monte Carlo

As we have seen, the aim of Bayesian inference is to characterize the complete posterior density $p(\mathbf{m}|\mathbf{d}_{\text{obs}})$. A Bayesian analysis often has two major problems. First, the posterior density cannot be expressed in a convenient analytic form, and the only way to determine the posterior is to evaluate it at different positions in the model space (as is the case here). Second, as the dimension of the model space increases, the number of models to test becomes huge, and a uniform sampling or complete enumeration of the posterior is not practical. One of the more flexible and popular techniques used to overcome these problems is the MCMC sampling method.

MCMC is an iterative stochastic approach to generate samples from a target probability density (for a concise introduction, see Sivia 1996). A sequence of models is generated in a chain, where typically each is a perturbation of the last. The starting point of the chain is selected randomly and the perturbations are governed by a proposal probability distribution that only depends on the current state of the model. The first part of the chain (called the burn-in period) is discarded, after which the random walk is assumed to be stationary and starts to produce an important sampling of the model space. Models generated by the chain are asymptotically distributed according to the posterior probability distribution (Tierney 1994). Given samples from the posterior distribution, the mean and the standard deviation can be directly determined from the distribution of the post burn-in samples.

In our problem, the dimension of the model space is itself a variable ($3n$) and the posterior becomes a transdimensional function. This can be sampled with a generalization of MCMC called reversible jump (Green 1995), which allows inference on both model parameters and model dimensionality. rj-MCMC is an extension of the well-known Metropolis–Hasting algorithm (Metropolis *et al.* 1953; Hastings 1970), and consists of a two-stage process of proposing a model probabilistically and then accepting or rejecting it. The proposal is made by drawing the new model, $\mathbf{m}'$, as a random deviate from a probability distribution $q(\mathbf{m}'|\mathbf{m})$ conditional only on the current model **m**. As before terms to the right of the bar are fixed and to the left are variable. In all expressions below, we use a prime to denote the state after the particular model step. Note that the proposed model, $\mathbf{m}'$, may be a vector of different length than the current model **m**, corresponding to partition models with differing number of Voronoi cells.

A simple example of a proposal distribution, which does not change dimension would be a Gaussian distribution with zero mean and diagonal covariance matrix $\mathbf{C} = \mathrm{diag}[\sigma_1^2, \sigma_2^2, \ldots]$

$$q(\mathbf{m}' \,|\, \mathbf{m}) \propto \exp\left\{-\frac{1}{2}(\mathbf{m} - \mathbf{m}')^{\mathbf{T}}\mathbf{C}^{-\mathbf{1}}(\mathbf{m} - \mathbf{m}')\right\}, \tag{15}$$

In practice to generate the new model $\mathbf{m}'$ from the existing model $\mathbf{m}$, one could perturb the ith component of $\mathbf{m}$ by drawing a random variable, u, from a normal distribution $N(0, 1)$ and set

$$\mathbf{m}' = \mathbf{m} + u\sigma_i \mathbf{e}_i, \tag{16}$$

where $\mathbf{e}_i$ is the unit vector in the ith direction and σ_i the variance of the proposal. This type of proposal distribution is common in fixed dimension applications of the Metropolis–Hasting algorithm. One usually cycles through the parameters perturbing each one at a time. Note in this example the proposal distribution is symmetric, because the forward proposal distribution, that is, probability of generating a perturbed model at $\mathbf{m}'$ when starting from $\mathbf{m}$ is the same as the reverse proposal distribution, that is, probability of starting from $\mathbf{m}'$ and generating a sample at $\mathbf{m}$. Hence, we have $q(\mathbf{m}'|\mathbf{m}) = q(\mathbf{m}|\mathbf{m}')$.

Once a proposed model has been drawn from the distribution $q(\mathbf{m}'|\mathbf{m})$, the new model is then accepted with a probability $\alpha(\mathbf{m}'|\mathbf{m})$, that is, a uniform random deviate, r, is generated between 0 and 1. If $r \leq \alpha$, the move is accepted, the current model $\mathbf{m}$ is replaced with $\mathbf{m}'$ and the chain moves to the next step. If $r > \alpha$, the move is rejected and the current model is retained for the next step of the chain where the process is repeated. The acceptance probability, $\alpha(\mathbf{m}'|\mathbf{m})$, is the key to ensuring that the samples will be generated according to the target density $p(\mathbf{m}|\mathbf{d}_{\mathrm{obs}})$. It can be shown (Green 1995, 2003) that the chain will converge to the transdimensional posterior distribution, $p(\mathbf{m}|\mathbf{d}_{\mathrm{obs}})$, if

$$\alpha(\mathbf{m}' \,|\, \mathbf{m}) = \min[1,\ \textbf{prior ratio} \times \textbf{likelihood ratio} \times \textbf{proposal ratio} \times |\mathbf{J}|] \tag{17}$$

$$= \min\left[1, \frac{p(\mathbf{m}')}{p(\mathbf{m})} \times \frac{p(\mathbf{d}_{\mathrm{obs}} \,|\, \mathbf{m}')}{p(\mathbf{d}_{\mathrm{obs}} \mid \mathbf{m})} \times \frac{q(\mathbf{m} \,|\, \mathbf{m}')}{q(\mathbf{m}' \,|\, \mathbf{m})} \times |\mathbf{J}|\right], \tag{18}$$

where the matrix $\mathbf{J}$ is the Jacobian of the transformation from $\mathbf{m}$ to $\mathbf{m}'$ and is needed to account for the scale changes involved when the transformation involves a jump between dimensions (Green 2003). That is, the Jacobian term 'normalizes' the difference in volume between two spaces of different dimension. In the Appendix, we show that $|\mathbf{J}| = 1$ for the problem considered here and so can conveniently be ignored. For a more detailed discussion on the reversible jump algorithm and calculation of the Jacobian, the reader is referred to Denison *et al.* (2002b) and Green (2003).

The expression for $\alpha(\mathbf{m}'|\mathbf{m})$ involves the ratio of the posterior distribution evaluated at the proposed model, $\mathbf{m}'$ to the current model $\mathbf{m}$ multiplied by the ratio of the proposal distribution for the reverse step, $q(\mathbf{m}|\mathbf{m}')$, to the forward step, $q(\mathbf{m}'|\mathbf{m})$. For symmetric proposal distributions, this ratio is one and drops out of the calculation. The likelihood function and the prior only enter into the algorithm through the acceptance probability term (18). The process of accepting or rejecting moves in this way controls the sampling of the Markov chain so that it preferentially samples regions of parameter space with high values of the target density, $p(\mathbf{m}|\mathbf{d}_{\mathrm{obs}})$. More precisely the density of the chain will asymptotically converge to that of the target density. The rate of convergence is controlled by the form of the proposal distribution. There is considerable freedom in design of the proposals. Ideally one would use a simple distribution (from which random samples could be drawn) that in some sense matches the shape of the local target density (posterior distribution) about the current model $\mathbf{m}$. It is important to note that the choice of proposal distributions only affects the convergence rate of the algorithm and not the distribution to which the algorithm will converge. From an inversion viewpoint then these choices do not affect the result of the inversion, 'merely' the practicality of the algorithm. In the next section, we describe the proposal distributions used for each type of variable in the tomographic problem.

2.7 Proposal distributions

In this study, we use four different types of perturbation to the model $\mathbf{m}$. One is a 'birth' step that adds a Voronoi cell to the existing parametrization. Another is a 'death' step that removes one of the Voronoi cells. The third is a 'move' step that is a perturbation to the position of a randomly chosen nucleus, $\mathbf{c}_i$, and the fourth is a velocity step involving a Gaussian perturbation to a velocity parameter, v_i. Note that three of the four perturbation types change the parametrization and one changes the velocity values. Together they form a randomized perturbation that is able to generate a wide range of velocity models from few to many degrees of freedom with multiple spatial scalelengths (see Fig. 4).

2.7.1 Generating new models along the Markov chain

To make the proposal distributions explicit, consider the algorithm at some point $\mathbf{m}$ in parameter space. It then proceeds as follows:

(1) *At every even step of the chain:* randomly pick one velocity parameter, say v_i and perturb its value using a Gaussian probability density $q_{v1}(v_i'|v_i)$

$$q_{v1}(v_i' \,|\, v_i) = \frac{1}{\sigma_1\sqrt{2\pi}} \exp\left\{-\frac{(v_i' - v_i)^2}{2\sigma_1^2}\right\}. \tag{19}$$

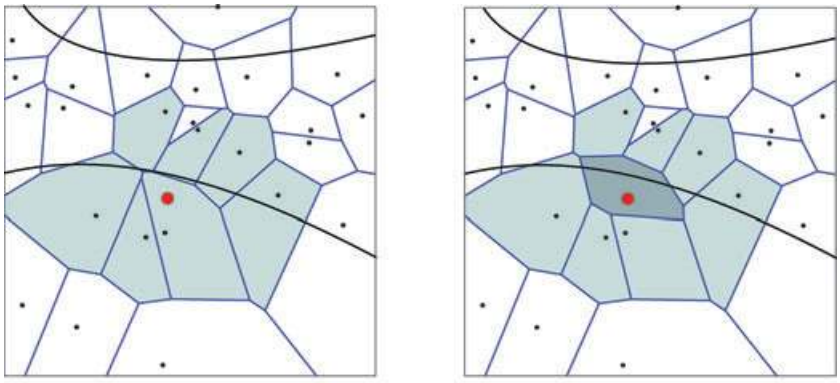


Figure 4. An example of a birth and death of a Voronoi cell. The birth step is represented by (a)→(b), where the red node is added to the cells in (a) and the resulting partition is shown in (b). Note that only the neighbouring cells (light grey) of the new born cell (dark grey) have their geometry changed during the birth. All other cells are unchanged. A death step is the reverse procedure, that is, (b)→(a). When the red node is removed, the dark grey cell disappears and the light grey cells expand to fill the gap. Two seismic rays are shown (thick black lines) corresponding to a reference velocity model. In both cases local changes in cell geometry result and only the traveltimes of rays passing through these cells need updating.

Hence we have

$$v_i' = v_i + u \times \sigma_1, \tag{20}$$

where u is a random deviate from a normal distribution $N(0, 1)$ and σ_1 is the standard deviation of the proposal. All the other model parameters are kept constant, and hence this proposal does not involve a change in dimension.

(2) *At every odd step of the chain:* perturb the cellular parametrization. This involves three possible types of change (birth, death and move) randomly selected with probability $1/3$ each.

(i) *BIRTH:* for a birth we create a new cell with the position $\mathbf{c}'_{n+1}$ by randomly selecting a point from the underlying grid that is not already occupied. For N grid points and n current cells there are $(N - n)$ discrete points to choose from. Once chosen, this becomes the nucleus of a new Voronoi cell in the parametrization (see Fig. 4). A velocity value v'_{n+1} needs to be assigned to the new cell. This is drawn from Gaussian proposal probability density $q_{v2}(v'_{n+1}|v_i)$ with the same form as (19), centred at v_i, where v_i is the current velocity value at the location $\mathbf{c}'_{n+1}$ where the birth takes place. The variance of the Gaussian function, σ_2^2 is a parameter to be chosen.

(ii) *DEATH:* delete a Voronoi centre chosen randomly from the current set of n cells. The death jump is the exact reverse of the birth jump (see Fig. 4).

(iii) *MOVE:* randomly pick a cell and perturb the position of its nucleus $\mathbf{c}_i$ according to a 2-D Gaussian proposal probability density $q_c(\mathbf{c}'_i|\mathbf{c}_i)$ centred on the current position $\mathbf{c}_i$.

$$q_c(\mathbf{c}'_i \,|\, \mathbf{c}_i) = \frac{1}{2\pi\sigma_c^2} \exp\left\{-\frac{1}{2\sigma_c^2}(\mathbf{c}'_i - \mathbf{c}_i)^T(\mathbf{c}'_i - \mathbf{c}_i)\right\}. \tag{21}$$

The covariance matrix for the 2-D Gaussian function is proportional to the identity matrix, with the constant of proportionality, σ_c^2. For this type of perturbation, the velocity parameter moves with the cell and so the velocity vector and the dimension of the model remains unchanged.

The geometrical calculations required to update the Voronoi diagram after the addition, removal or movement of a nucleus do not involve recalculation of the entire Voronoi diagram, only a local change. This can be efficiently implemented with the local Voronoi update algorithm described in Sambridge *et al.* (1995b).

2.7.2 Proposal ratios

Having described the four types of model perturbation, we now need to evaluate the proposal ratio of forward and reverse moves so that the acceptance probability in (18) can be calculated in each case. For the proposal types that do not involve a change of dimension (i.e. a velocity update and a nucleus move) the distributions are symmetrical. That is, the probability to go from $\mathbf{m}$ to $\mathbf{m}'$ is equal to the probability to go from $\mathbf{m}'$ to $\mathbf{m}$. Hence

$$\begin{aligned} q_{v1}(v'_i \,|\, v_i) &= q_{v1}(v_i \,|\, v'_i) \\ q_c(\mathbf{c}'_i \,|\, \mathbf{c}_i) &= q_c(\mathbf{c}_i \,|\, \mathbf{c}'_i) \end{aligned} \tag{22}$$

and so in both cases the proposal ratio is one

$$\frac{q(\mathbf{m} \,|\, \mathbf{m}')}{q(\mathbf{m}' \,|\, \mathbf{m})} = 1, \tag{23}$$

which simplifies the acceptance probability expression (18). The situation is different for the birth and death proposal steps. Here the dimension of the model does change and the proposals are not symmetric. In these cases, we must determine expressions for the proposal ratios to be inserted into (18). For a birth step, the algorithm jumps between a model $\mathbf{m}$ with n cells to a model $\mathbf{m}'$ with $(n+1)$ cells. Since the new nucleus $\mathbf{c}'_{n+1}$ is generated independently from the velocity value v'_{n+1} then proposal distributions can be separated and we write

$$\frac{q(\mathbf{m}\,|\,\mathbf{m}')}{q(\mathbf{m}'\,|\,\mathbf{m})} = \frac{q(\mathbf{c}|\ \mathbf{m}')}{q(\mathbf{c}'\,|\,\mathbf{m})}.\frac{q(\mathbf{v}\,|\,\mathbf{m}')}{q(\mathbf{v}'\,|\,\mathbf{m})}, \tag{24}$$

where each term on the right-hand side follows from the definitions above. Specifically, we have the probability of a birth at position $\mathbf{c}'_{n+1}$, which is given by

$$q(\mathbf{c}'\,|\,\mathbf{m}) = 1/(N-n), \tag{25}$$

the probability of generating a new velocity value at v'_{n+1}

$$q(\mathbf{v}'\,|\,\mathbf{m}) = \frac{1}{\sigma_2\sqrt{2\pi}}\exp\left\{-\frac{(v'_{n+1}-v_i)^2}{2\sigma_2^2}\right\}, \tag{26}$$

the probability of deleting the cell at position $\mathbf{c}'_{n+1}$ (reverse step)

$$q(\mathbf{c}\,|\,\mathbf{m}') = \frac{1}{(n+1)} \tag{27}$$

and the probability of removing a velocity when cell is deleted (reverse step)

$$q(\mathbf{v}\,|\,\mathbf{m}') = 1. \tag{28}$$

Substituting these expressions in (24), we obtain

$$\left[\frac{q(\mathbf{m}\,|\,\mathbf{m}')}{q(\mathbf{m}'\,|\,\mathbf{m})}\right]_{\text{birth}} = \frac{\sqrt{2\pi}(N-n)}{(n+1)}\sigma_2\exp\left\{\frac{(v'_{n+1}-v_i)^2}{2\sigma_2^2}\right\}. \tag{29}$$

We see then that as a new cell is created, the probability distribution increases exponentially as the new cell's velocity, v'_{n+1}, departs from the velocity that was in the same position in the unperturbed model, v_i. Hence the birth process encourages changes in velocity as well as parametrization. This exponentially increasing probability density is ultimately restrained when combined with the likelihood ratio term in (18), which would penalize large velocity perturbations that did not lead to an improvement in data fit.

For the death of a randomly chosen nucleus, we move from n to $(n-1)$ cells (Fig. 4). Suppose that nucleus, $\mathbf{c}_i$ with velocity v_i is removed. In this case, a similar reasoning to the birth case above leads us to a proposal ratio (reverse to forward) of

$$\left[\frac{q(\mathbf{m}\,|\,\mathbf{m}')}{q(\mathbf{m}'\,|\,\mathbf{m})}\right]_{\text{death}} = \frac{n}{\sigma_2\sqrt{2\pi}(N-n+1)}\exp\left\{-\frac{(v'_j-v_i)^2}{2\sigma_2^2}\right\}, \tag{30}$$

where v'_j is the velocity at the point $\mathbf{c}_i$ in the new tessellation, $\mathbf{c}'$, after removal of the ith cell.

We see then that for all four types of step we are able to (i) generate new samples easily and (ii) determine the proposal ratio for insertion into the acceptance probability expression (18). To complete the evaluation of the acceptance probability $\alpha(\mathbf{m}'|\mathbf{m})$, we need the likelihood ratio (7) and prior ratio (14). The likelihood evaluation involves calculation of traveltimes in the proposed model $\mathbf{m}'$. In the linearized formulation used here, we simply integrate along rays calculated in the current reference model although in a fully non-linear approach we would update the rays as well.

The Markov chain will converge for a wide range of proposal distributions, and hence there is freedom in choosing the parameters (σ_1, σ_2, σ_c). In practice, poor choices of variance lead to slow movement around the model space, such that convergence of the chain can depend exponentially on the number of steps, an undesirable situation. For example as perturbations in velocity variables become larger (σ_1 is increased) then more velocity steps would tend to be rejected because the data fit (and likelihood ratio) would tend to decrease. Hence the chain would sample the space less. Conversely, as the velocity perturbations decrease (σ_1 is decreased) the acceptance ratio would increase but the Markov chain would take much smaller steps around model space. At both extremes, convergence would be inhibited (Hopcroft *et al.* 2007). Ideally the proposal distribution should be similar in shape to the local posterior probability function about the current model. In the ideal case, the proposal and posterior distribution were the same then $\alpha(\mathbf{m}'|\mathbf{m}) = 1$ and all steps would be accepted, but this could not happen as the proposal distribution must be one where we can generate samples using some simple method, such as those for a Gaussian. Design of suitable proposal distributions that adapt to the shape of the posterior distribution is a central issue in the development of transdimensional MCMC algorithms and the subject of much research (Stephens 2000; Brooks *et al.* 2003; Green 2003; Al-Awadhi *et al.* 2004).

One sign of inefficiency easily detected in experiments is a high rejection rate for the proposed changes. In fixed dimensions, small changes usually have higher acceptance rates than large ones, and proposal mechanisms can be scaled to achieve a desired acceptance rate (e.g. Gelman *et al.* 1996; Mosegaard 1998; Tierney & Mira 1999). Brooks *et al.* (2003) point out that this option is not always available for moves between dimensions as there may be no natural distance measure between states of different dimensions. Therefore, failure to

achieve acceptable performance can be considered merely a result of poorly constructed between-dimension transitions (see Sisson 2005, for a discussion). A problem that arises in our partition model is that the 'size' of a jump between two dimensions varies with the dimension itself. Adding a new cell to a 3-cell model will represent a larger change compared to adding 1 to a 100-cell model (Although in principle this could be corrected by having our Gaussian proposal function q_{v2} dependent on the dimension).

In an attempt to locally scale the variance of our Gaussian proposal distributions for the fixed dimension moves, we have implemented the delayed rejection scheme proposed by Tierney & Mira (1999). The basic idea is that, upon rejection, instead of advancing time and retaining the same position, a second move with lower variance is proposed. The acceptance probability of the second-stage candidate is computed so the convergence of the chain towards the posterior distribution is preserved. Details are given in Appendix A. In this study, delayed rejection was only used for moves that do not jump between dimensions although Green & Mira (2001) showed that the method can be extended to transdimensional moves. The delayed rejection scheme is particularly useful in increasing convergence rate and robustness of the Markov chain, in effect removing the need to carefully tune the proposal distributions.

2.8 The acceptance probability

To complete our description of the algorithm, we now substitute expressions for each proposal ratio into (18) to get final expressions for the acceptance probability in each case. For the velocity update and nucleus move steps, we have seen that the proposal ratio and Jacobian terms become unity. Hence for both cases the acceptance term is simply given by the ratio of the posteriors

$$\alpha(\mathbf{m}' \,|\, \mathbf{m}) = \min\left[1, \frac{p(\mathbf{m}')}{p(\mathbf{m})} . \frac{p(\mathbf{d}_{\rm obs} \,|\, \mathbf{m}')}{p(\mathbf{d}_{\rm obs} \,|\, \mathbf{m})}\right]. \tag{31}$$

Since the dimension of the model does not change, according to (14), the prior ratio is either null or unity and we have

$$\alpha(\mathbf{m}', \mathbf{m}) = \begin{cases} \min\left[1, \dfrac{p(\mathbf{d}_{\rm obs} \,|\, \mathbf{m}')}{p(\mathbf{d}_{\rm obs} \,|\, \mathbf{m})}\right] & \text{if} \quad \forall i \in [1, n],\ v_i \in J) \\ 0 & \text{otherwise.} \end{cases} \tag{32}$$

For both the move and velocity update steps, this only requires the ratio of the likelihoods and hence calculation of traveltimes at the proposed model. We see then that perturbations that improve data fit are always accepted and those which decrease it are accepted with probability equal to the ratios of the likelihoods.

As mentioned above, we use a delayed rejection scheme for fixed dimension moves. If the candidate $\mathbf{m}'$ is rejected, a second try $\mathbf{m}''$ is made by drawing from a similar proposal distribution but with smaller variance. The acceptance term for the second candidate $\alpha_2(\mathbf{m}''|\mathbf{m})$ is more complicated to determine, an expression is given in Appendix A. Note that by reducing the variance of the second proposal, we attempt a less ambitious move that is more likely to be accepted. In principal, this process can be repeated every time a rejection occurs thereby increasing the overall acceptance rate and hence efficiency of the algorithm.

For a birth step, according to (14), the prior ratio takes the form

$$\left[\frac{p(\mathbf{m}')}{p(\mathbf{m})}\right]_{\rm birth} = \begin{cases} \left[\frac{(n+1)!(N-n-1)!}{N!(\Delta v)^{n+1}\Delta n}\right]\left[\frac{n!(N-n)!}{N!(\Delta v)^{n}\Delta n}\right]^{-1} & \text{if} \quad ((n+1) \in I \quad \text{and} \quad v'_{n+1} \in J) \\ 0 & \text{otherwise} \end{cases} \tag{33}$$

$$\left[\frac{p(\mathbf{m}')}{p(\mathbf{m})}\right]_{\rm birth} = \begin{cases} \frac{n+1}{(N-n)\Delta v} & \text{if} \quad [(n+1) \in I \quad \text{and} \quad v'_{n+1} \in J] \\ 0 & \text{otherwise.} \end{cases} \tag{34}$$

After substituting (7), (29) and (34) into (18), the acceptance term reduces to

$$\alpha(\mathbf{m}', \mathbf{m}) = \begin{cases} \min\left[1, \frac{\sigma_2\sqrt{2\pi}}{\Delta v} . \exp\left\{\frac{(v'_{n+1}-v_i)^2}{2\sigma_2^2} - \frac{\phi(\mathbf{m}')-\phi(\mathbf{m})}{2}\right\}\right] & \text{if} \quad [(n+1) \in I \quad \text{and} \quad v'_{n+1} \in J] \\ 0 & \text{otherwise,} \end{cases} \tag{35}$$

where i is the cell in the current tessellation $\mathbf{c}$ that contains the point $\mathbf{c}'_{n+1}$ where the birth takes place. For the birth step then we see the acceptance probability is a balance between the proposal probability (which encourages velocities to change) and the difference in data misfit that penalizes velocities if they change so much that they degrade fit to data.

For the death step, the prior ratio in (34) must be inverted. After substituting this with (7) and (30) into (18), and after simplification we get an the acceptance probability

$$\alpha(\mathbf{m}', \mathbf{m}) = \begin{cases} \min\left[1, \frac{\Delta v}{\sigma_2\sqrt{2\pi}} . \exp\left\{-\frac{(v'_j-v_i)^2}{2\sigma_2^2} - \frac{[\phi(\mathbf{m}')-\phi(\mathbf{m})]}{2}\right\}\right] & \text{if} \quad (n-1) \in I \\ 0 & \text{otherwise,} \end{cases} \tag{36}$$

where i indicates the cell that we remove from the current tessellation $\mathbf{c}$ and j indicates the cell in the proposed tessellation c' that contains the deleted point $\mathbf{c}_i$. Unsurprisingly the death acceptance probability has a similar form to that of the birth, with proposal and data terms opposing each other. We see from these expressions that the variable N, that is, the number of candidate positions for the nuclei, cancels out. This mean that there is no need to use an actual discrete grid in generating nuclei positions. In fact, it was only ever a mathematical convenience that

ensures that the acceptance expressions have the correct analytic form. In practice, we are at liberty to generate the nuclei using a continuous distribution over the region of the model (which is tantamount to $N \to \infty$).

Given that the choice of adding in more unknowns to the problem is left to the algorithm itself, one might ask whether there will be a tendency to simply improve data fit by continually adding in more cells. It turns out that this is not the case. In proposing an increase in the number of cells (a birth step), the likelihood function will tend to encourage acceptance when fit is improved, however the prior ratio will tend to discourage acceptance due to the increased dimensionality of the space. Loosely speaking, the algorithm always prefers a large cell rather than two small cells with similar velocity values. This is an example of a property of Bayesian inference referred to as 'natural parsimony', which means that given a choice between a simple and complex models that provide similar fits to data, the simpler one will be favoured (see MacKay 2003, for a discussion).

A simple way to check that the form of the acceptance term is correct is to set the likelihood to a uniform distribution (i.e. remove the data). In this case, the posterior is directly proportional to the prior and the Markov chain should sample the known prior distribution. We verified that this was the case by performing numerical experiments with the likelihood set to unity. We recovered a uniform distribution of the number of cells in the resulting ensemble, which is what was specified for the prior $p(n)$ in (9). (This experiment was also repeated for other choices of $p(n)$ and in each case, histograms of the number of cells for models in the posterior ensemble closely resembled those of the assigned prior.)

2.9 Extracting a reference solution and error map from the ensemble

After employing the reversible jump algorithm to sample the transdimensional posterior, we end up with an ensemble of velocity models with varying numbers of cells. If convergence has been achieved then these will reflect the posterior density. In our linearized scheme described by Fig. 2, we need to extract a reference model for use in the outer loop of the algorithm. A single model is also a useful aid for interpretation. One possibility is to take the velocity model with the maximum posterior value (often called the MAP in Bayesian terminology). However, this is not of much use in our case as it often corresponds to a single partitioned model with relatively crude parametrization (an example is seen in Section 3).

Instead we look at the spatially averaged model defined by taking the mean of the distribution of velocity values at each point across the 2-D region. That is, we project the partition models into the spatial domain and then average all the sampled images at a fine grid of positions across the model. The underlying grid structure can be as fine as needed for visualizing the reference model. At each iteration of the outer loop in Fig. 2, we take this pointwise average velocity field as the continuous reference model extracted from the ensemble of solutions, and use it to update ray geometries in the next iteration of the outer loop.

An estimated error map can be obtained in a similar manner, that is, by calculating the standard deviation of velocities as a function of position. In this way, a large number of models with different parametrizations are stacked together. As can be seen in the examples to follow, the continuous reference model contains features common to the entire family of models and considerably more information than any single Voronoi partition.

2.10 Convergence assessment

It is important to collect enough samples so that the solution maps are stationary and represent well the posterior mean and variance. The issue for assessing convergence of the algorithm, that is, when to start collecting the sample of models and how many to collect, is the subject of current research in Bayesian statistics. (e.g. Brooks & Giudici 1999; Brooks *et al.* 2003). To date, there have been relatively few convergence diagnostics designed specifically for transdimensional samplers. Current technology seems to be insufficiently advanced to permit a rigorous assessment of stationarity. Although the potential benefits of transdimensional Markov chains seem to be large, the practical importance of ensuring chain convergence is often overlooked by practitioners (see Sisson 2005, for a discussion).

Conventional convergence diagnostics for a Markov chain, (see, for example, Cowles & Carlin 1996; Gelman & Rubin 1992; Robert 1995) rely on showing that deviations from stationarity are not present in individual parameters throughout the run of the sampler (i.e. these are not 'drifting' in any direction). In practice, the population of samples for a model parameter plotted as a function of iteration should resemble for example a white noise process, with no trends or obvious structure.

However, for transdimensional posterior distributions, the most widely used convergence diagnostics are not applicable. Parameters for models that change dimension have little interpretation from one model to the next. For example, the location of the 'tenth' nucleus, $\mathbf{c}_{10}$, does not have the same meaning across all the models. When there are less than 10 cells, it is not even present in the model. This makes tracking the position of this particular nuclei along the chain meaningless. Therefore, in this study we assess convergence (or non convergence) without using parameters ($\mathbf{c}$, $\mathbf{v}$) defining the partitioned models. Instead, we look for convergence in terms of numbers such as the velocity value at a given point in the 2-D field or the model dimension.

2.11 Computational cost

When performing a Monte Carlo (MC) integration, the dimension of the model space, for example, the number of unknowns, has to be restricted in size. If too many parameters define the model, the number of samples needed to explore the whole model space becomes huge

(Tarantola 2005). The predicted data have to be computed each time a model is proposed, and if too many models need to be generated, all algorithms become computationally prohibitive. Tomographic models use a large amount of data and often require a large number of cells to image heterogeneous structures with adequate resolution. This is a reason why MC methods have not been habitually used in tomographic imaging. Nevertheless, several features of the proposed methodology make the algorithm feasible.

The first, as mentioned above, is that we only compute the ray geometries in the outer loop of the algorithm, thereby saving considerable computation. The second is that each time a new partitioned velocity model is tested in the inner loop, only a part of the traveltimes are recomputed in the evaluation of (3). For example, Fig. 4 shows the geometry of the problem before and after a Voronoi nucleus is added. Only the cells in grey change during this birth jump. Therefore, only the traveltimes of the rays crossing the grey cells need to be updated to compute the data misfit of the new model. This also turns out to be a significant saving of compute time and allows the number of unknowns to be larger than for a standard MC approach.

Another computational consideration is that the algorithm is straightforward to parallelize in the sense that multiple chains (i.e. inner loops) can sample the model space independently of each other. For example, each chain may be conveniently placed on an independent processor of a parallel computer system (Rosenthal 2000). To give some measure of computational cost, the synthetic examples presented in Section 3 can all be performed without parallelization on a standard desktop workstation (Intel core 2 duo with CPU running at 2.1 GHz) using about 150 min CPU time.

3 SYNTHETIC DATA EXAMPLES

3.1 Experimental setup

A synthetic data set is constructed by using the FMM to compute traveltime for seismic energy propagating across a spherical surface between 17 sources and 20 receivers in the presence of severe velocity heterogeneity. This simplistic setup also contains highly irregular distribution of rays and is motivated by surface wave experiments in regions of limited data coverage.

The synthetic velocity field is shown in Fig. 5. The areas in red have a velocity of 5 km s^{-1} and the blue areas are of 4 km s^{-1}. The velocity field presents high contrast discontinuities. The blue heterogeneity represents a velocity anomaly of -20 per cent of the red background and the red heterogeneity is $+25$ per cent of the blue background. Hence this 'simple' problem is reasonably non-linear and serves to illustrate the algorithm.

The ray geometry associated with the synthetic velocity field is shown in Fig. 6. All the calculations are performed in 2-D spherical coordinates (Hence, straight rays become great circles on a sphere). As expected, the rays avoid the red low velocity heterogeneity in the lower-right part of the velocity field and are attracted towards the blue high velocity heterogeneity in the upper-left part. Overall, the lower-left part of the model is covered by many ray paths, whereas the upper-left part is barely sampled.

Here we see the difficulty of choosing an appropriate cell size for a regular mesh. A constant cell size across the entire model will most likely result in the problem becoming underdetermined in the upper-left part (not enough rays crossing the cells) and overdetermined in the lower-right part (large number of rays crossing each cell). A second problem is that in the upper-left quarter, all the rays are in similar directions (i.e. NW to SE), indicating that the resolution in this direction will be poor. This effect is typically associated with smearing in tomographic reconstruction algorithms.

We choose to compare the reversible jump tomography to a subspace method (e.g. Kennett *et al.* 1988), which is a convenient inversion scheme based on a fixed regular parametrization. Both approaches are linearized methods and use the same forward modelling to update the geometry of rays. The subspace scheme uses a matrix inversion approach including implicit regularization to solve the linearized tomographic equations at each iteration.

We observe and compare the propagation of data error into model uncertainty. For the true model and geometry of rays, the average observed traveltime is 473 s. For a homogeneous initial model with velocities equal to 4.5 km s^{-1}, without noise in the observed data, the average difference between observed and estimated traveltimes is about 35 s. Some random Gaussian noise has been added to the observed traveltimes with a standard deviation of 9.5 s (i.e. 2 per cent of the average observed traveltime).

3.2 Fixed parametrization tomography with the subspace method

3.2.1 The regularization process

Most of the methods using a predefined fixed parametrization formulate the tomography problem with a linear system of algebraic equations represented by a matrix $\mathbf{G}$. In the example considered here, the ray coverage is quite sparse and with a uniform grid of sufficiently small cell sizes, the problem becomes non-unique. Regularization procedures must be used to choose a solution among all the acceptable possibilities. This resulting solution will have properties reflecting the particular choice of regularization.

In order to discard the models that are unrealistic and make the solution unique, criteria other than the misfit can be minimized such as the distance to a reference model or the norm of the first or second spatial derivative of the velocity field. The inversion scheme consists then

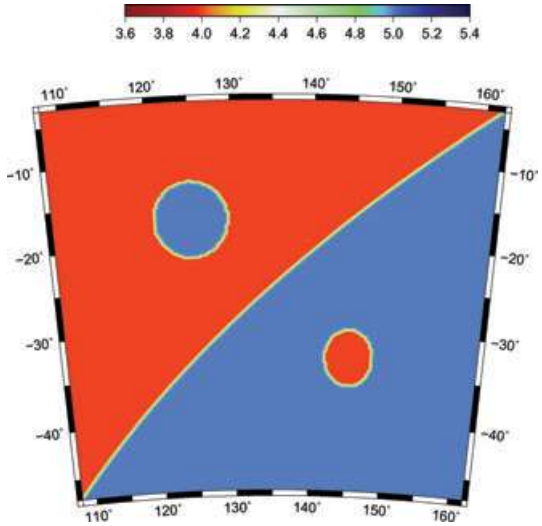


Figure 5. True velocity field (km s^{-1}).

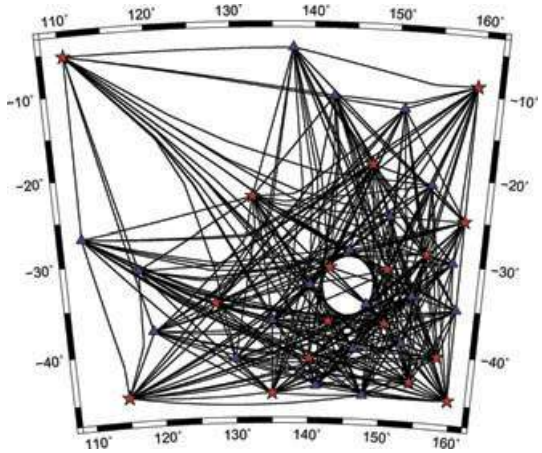


Figure 6. True ray paths. 340 rays join 17 sources (red stars) to 20 receivers (blue triangles). Due to the principle of least time, the rays seem to avoid the low-velocity anomaly and to be attracted towards the high-velocity anomaly.

in minimizing an objective function that is a linear combination of different criteria

$$\Phi(\mathbf{m}) = \left\| \frac{\mathbf{Gm} - \mathbf{d}_{\text{obs}}}{\sigma_d} \right\|^2 + \varepsilon \|\mathbf{m} - \mathbf{m}_0\|^2 + \eta \|\mathbf{Dm}\|^2, \tag{37}$$

where the first term is the data misfit, the second quantifies the distance to a reference model $\mathbf{m}_0$ and in the third, the vector $\mathbf{Dm}$ is a finite difference approximation proportional to either the first or second derivative of the model. By minimizing the semi-norm $\|\mathbf{Dm}\|^2$, the regularization techniques favour models that are relatively flat (first-order regularization) or smooth (second-order regularization). The damping factor ε effectively prevents the solution model from staying too far from the reference model $\mathbf{m}_0$, while the smoothing factor η constrains the smoothness of the solution model.

3.2.2 Fixed parametrization and B-spline interpolation

Here, the velocity field is defined by a uniform grid of nodes with bicubic B-spline interpolation (Virieux & Farra 1991). Once a velocity value has been assigned to each node, the grid is interpolated with spline patches to produce a continuous, smooth and locally controlled velocity field. These nodes constitute the inversion grid, that is, the velocity values of these nodes are adjusted by the inversion scheme in order to satisfy the data. The nodes are evenly distributed and do not move during the inversion process. This way of parametrizing the velocity field is common in surface wave tomographic studies (e.g. Yoshizawa & Kennett 2004; Fishwick *et al.* 2005). However, note that the choice of a B-spline interpolation is purely arbitrary. One could equally well have chosen a triangle-based linear interpolation or any other type of 2-D interpolation.

3.2.3 The subspace method

Here, the unknown which is sought for during the inversion step is a perturbation of the reference field (which causes only a perturbation to the geometry of rays). The problem is then locally linearized around the reference model. The perturbed solution becomes the reference model for the next iteration. The process is stopped when, for example, the data are satisfied, that is when the normalized χ^2 misfit measure (corresponding to the first term in (37) divided by the number of data) equals one (Rawlinson *et al.* 2006).

The method is a subspace inversion because at each iteration, it projects the full linearized inverse problem onto a smaller m-dimensional model space to reduce computational effort. Details of the subspace method are given in Kennett *et al.* (1988), Rawlinson & Sambridge (2003) and Rawlinson *et al.* (2006). The advantage of this approach is that the optimization of (37) can proceed with only the inversion of an $m \times m$ matrix at each iteration. The set of vectors that span the m-dimensional subspace are computed based on the gradient vector and Hessian matrix in model space. In our experiments, we set m to 15, which results in having the m vectors strongly linearly dependent. Singular value decomposition is used to orthogonalize subspace vectors, and remove those directions that are redundant. In obtaining the results presented in the next section, we have experimented with the number of subspace vectors and found that the velocity models obtained are not strongly dependent on the choice of subspace dimension.

3.2.4 Results

Fig. 7 shows the results obtained after six iterations for a grid of 20×20 nodes for different values of ε and η with a semi-norm defined by the second derivative of the model. When we changed the two regularization parameters, we observed the classic trade-off between smooth models with poor spatial resolution 7(d) and instability 7(a) (Menke 1989). The solution shown in 7(a) has been obtained with relatively small values of ε and η. It has strong amplitudes and shows features not present in the true model. The map in 7(b) is damped and 7(c) is a smoothed solution. The solution model in 7(d) has been produced with relatively large values for both regularization parameters.

Here, the user has to choose the number of nodes in the grid, and we experienced the difficulty of manually finding an optimal value. When the number of nodes is decreased, the instabilities are removed but the spatial resolution is not good enough to map the heterogeneities. Fig. 7 represents the best results we have been able to obtain with the regularization framework. Of course, automated procedures methods also exist for selecting optimal values of the regularization parameters such as L-curve, cross-validation or the discrepancy principle (Aster *et al.* 2005). Each has their limitations. In this synthetic problem, we know the true solution and are able to judge the performance of each pair of regularization parameters by comparing the result to the known true model. Of course this is not possible in a real data case but in our synthetic problem we are more interested in obtaining the best possible pair (ε, η) for comparison with the reversible jump tomography. The model in 7(c) seems to be the closest to the true model and hence will be used as the best solution from the subspace inversion for comparison.

3.3 Reversible jump tomography

3.3.1 The average model: a naturally smooth solution

Posterior inference was made using an ensemble of 5000 models. We ran the rj-MCMC algorithm for 560 000 steps in total. The first 60 000 steps were discarded as burn-in steps, only after which the sampling algorithm was judged to have converged. Then, every 100th model visited in the last 500 000 steps was taken in the ensemble. The prior on the number of cells $p(n)$ was set uniform with $n_{\min} = 0$ and $n_{\max} = 500$. Four passes were made around the outer loop of the algorithm (Fig. 2) with an update of the ray geometry for each pass. The results presented here are obtained from the ensemble of samples collected during the last iteration only.

The best partitioned velocity model obtained in terms of posterior value is shown in Fig. 8(a). It would appear to be a rather poor recovery of the true model in Fig. 5. However, the spatial average of the post burn-in samples collected shown in Fig. 8(b) seems to recover much closer the features of the true velocity field. Each individual partitioned model consists of a different configuration of a finite number of Voronoi cells as in Fig. 8(a), but the average solution taken pointwise is smooth, except across the true discontinuities, where a rapid change is seen. Since the variability of the individual models in the ensemble represents the posterior distribution then by averaging them spatially we have a form of 'data-driven' smoothing, that is, without the need to impose an explicit smoothing function, choose regularization parameters or interpolation procedure. Average velocity maps such as Fig. 8(b) are in a sense self-regularized solutions.

The average solution is clearly quite different from the models obtained with a fixed grid. The artefacts of the later are not present and the discontinuities have been recovered with better accuracy. Furthermore, the fictitious gradients that are evident in Fig. 7 are removed in Fig. 8(b) giving a more faithful recovery of the true model in Fig. 5. The normalized χ^2 misfit measure for the average solution model is 0.86, which is of the same order as for the solution in Fig. 7(c) (0.92).

In this transdimensional approach, the number of cells n needed to construct the model becomes a parameter itself in the inversion and it is possible to make posterior inference on it. Fig. 8(c) shows $p(n \mid \mathbf{d}_{\text{obs}})$, that is, a histogram of the number of cells in the output ensemble of velocity models. No models with more than 30 cells have been sampled, which suggests that the choice of an upper limit of 500 on the prior for the number of cells $p(n)$ was rather large. This choice in the prior therefore does not affect the solution.

It must be remembered that the parameter n is not a 'physical' parameter, it does not have any geological interpretation as a wave speed, or a layer thickness. This may seem somewhat awkward to interpret. However, we see that it is an unknown in the problem that can be constrained by data. From a Bayesian point of view, the distribution $p(n \mid \mathbf{d}_{\text{obs}})$ gives information on the complexity of the problem, that is on the level of support in the data for the number of degrees of freedom in the model. Model dimension parameters such as this are used extensively in the Bayesian computation literature (Sisson 2005).

Fig. 8(c) shows that this inversion only uses an average of about 13 mobile cells whereas the subspace inversion scheme uses 400 fixed cells. Hence the reversible jump approach achieves a finer representation of the velocity field with fewer model parameters that results, as expected, from averaging many overlapping Voronoi cells in different configurations. The solution in the regularization framework is obtained with chosen values for ε and η, and inevitably this represents a compromise across the entire model. In the transdimensional approach, there is no global damping parameter, but instead the algorithm has smoothed the model locally in response to the data.

It appears that the averaging process has removed unwarranted discontinuities in individual models (i.e. the large number of velocity discontinuities at every cell boundary) but constructively reinforced the well-constrained ones about the anomalies and the diagonal step. The dynamic parametrization looks to have adapted to the structural features of the underlying model.

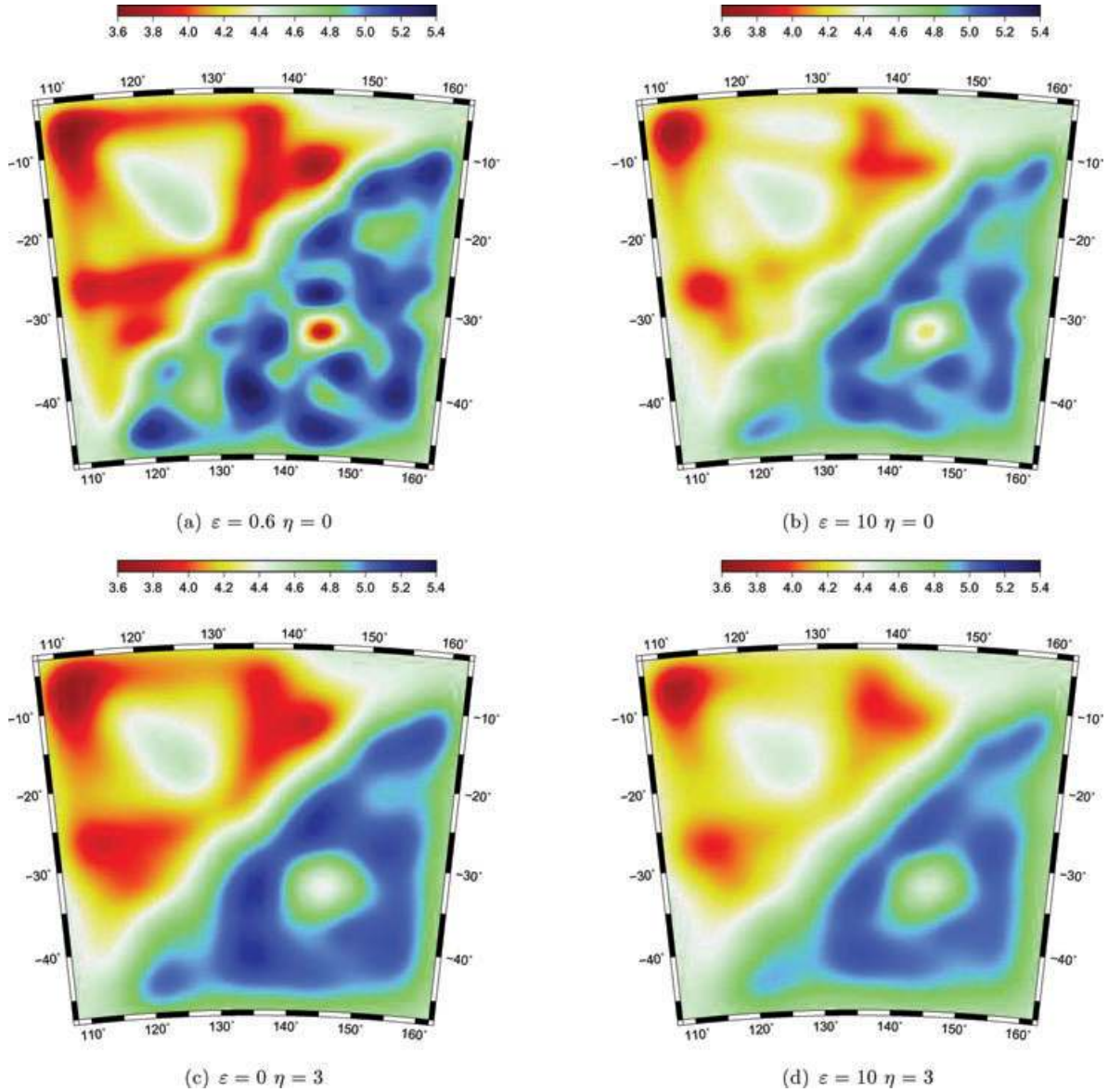


Figure 7. Subspace inversion with a regular grid (20 × 20 nodes) and B-spline interpolation. Results after six iterations for different values of damping and smoothing (km s^{-1}). A random Gaussian noise has been added to the data with a standard deviation equal to 2 per cent of the average observed traveltime. The colour scales are the same as for the true model.

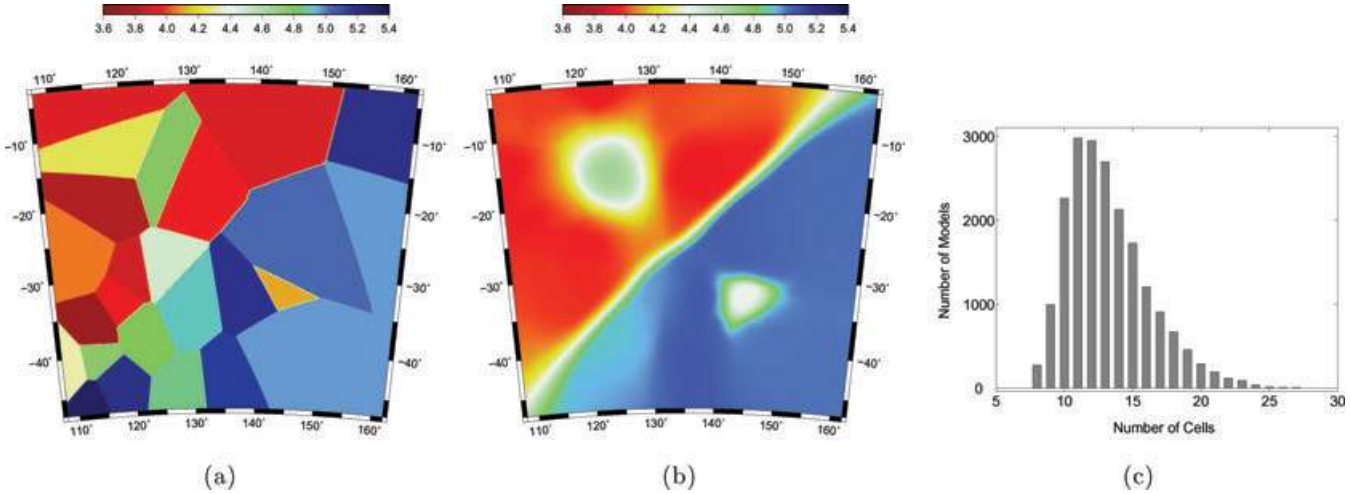


Figure 8. Reversible jump tomography. Results after four iterations (km s^{-1}). (a) best model sampled (i.e posterior maximum). (b) Average solution map. The scales are the same as for previous figures. (c) Information on the number of cells in the Voronoi tessellation. Posterior probability density for the parameter n, $p(n \mid \mathbf{d}_{\text{obs}})$.

3.3.2 The variance map: an estimate of model uncertainty

A well-known problem with regularized inversion algorithms that construct a single optimal model is that it is impossible to assess the distance between the estimated and true model. Regularization or damping helps to stabilize the inversion of linear systems equations and suppresses propagation of data noise into the solution, but this is at the cost of biasing the solution in a statistical sense (Aster *et al.* 2005). In practice if noise is added to data, the random variability estimated in the model can be much less than the true errors (For a recent example see Bodin *et al.* 2009). Technically the distance between the estimated and true model can only be constrained if additional information is available on the regularity of the true solution. Although recently alternate ways around this problem have been suggested (Rawlinson *et al.* 2008).

In contrast to regularization methods, the reversible jump algorithm enables one to perform an ensemble inference, that is to capture the variability in the range of possible solutions. The standard deviation of the family of models provides a smooth map that can be interpreted as an error map for the velocity model. Fig. 9 shows the results. The model uncertainty map obtained in this way (Fig. 9a) appears to be similar to the actual error in Fig. 9(c), but with higher amplitude. Indeed, as can be seen in 9(b), the one standard deviation estimated error (green dashed lines) includes the true model (black solid line) for the entire length of the upper profile and for more than 90 per cent of the lower profile. In these experiments, the MC sampler seems to provide a reliable estimation of the model uncertainty both in terms of amplitudes and lateral variations. We are unaware of any alternative approaches that can reproduce this kind of error estimation.

The true error map in 9(c) can be also compared to the true error map for the subspace solution in 9(d). Values are clearly smaller in 9(c), and hence the average solution obtained with the RJ-MCMC approach is closer to the true model than the subspace solution in 7(c). This can be quantified with the 'norm' for 9(c) being almost half of the 'norm' for 9(d).

It has been shown that the reversible jump tomography is particularly suited for recovering earth models with sharp features. It might be argued that this comparison suits the MCMC approach as it is often difficult for a uniform grid scheme to recover sharp gradients. To further examine these issues, we present a second synthetic example without discontinuities and a more complex spatial pattern of anomalies.

3.4 Example with a Gaussian random model

3.4.1 Synthetic model

In this example, the model is constructed from a uniform grid (14 × 14 nodes). The velocity value assigned to each node is drawn from a Gaussian distribution. The true synthetic model in Fig. 10 is obtained with a cubic B-spline interpolation between the nodes. Hence, the basis functions used to construct the model are the same as the basis functions used by the subspace inversion. The sources and receivers locations (Fig. 11) are the same as previous example. Some random Gaussian noise has been added to the observed traveltimes with a standard deviation of 5 s (i.e. about 1 per cent of the average observed traveltime).

3.4.2 Comparing regularized and reversible jump solutions

In the previous example, results with the subspace inversion were shown for the optimal grid size and different solutions corresponded to different values of regularization parameters. Here we show solutions obtained for different grid sizes (Fig. 12) and for each case, the regularization parameters are chosen with an 'L-curve' technique (Aster *et al.* 2005).

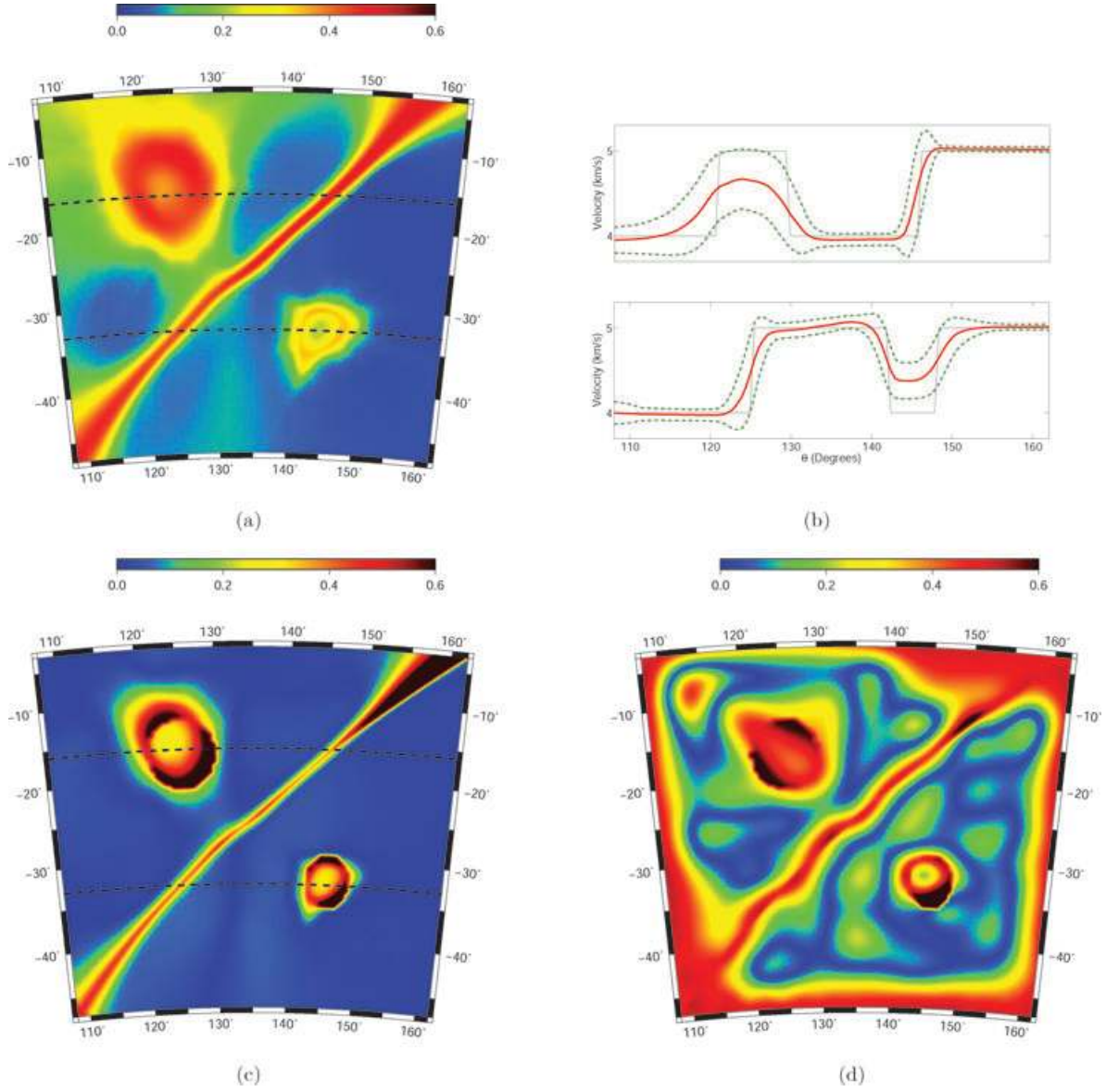


Figure 9. (a) Estimated error map (km s^{-1}). The two dashed lines show the cross-sections presented on 9(b). (b) Cross-sections showing the true model (black), the solution model (red) and +/− one standard deviation (dashed green). (c) Actual error for the reversible jump tomography. Absolute difference between the true model in ()5 and the estimated model in 8(b) (km s^{-1}). (d) Actual error for best subspace solution. Absolute difference between the true model in (5) and the estimated model in Fig. 7(c) (km s^{-1}).

For each grid size, the complete inversion process was run a number of times with different values of ε and η. This was done systematically by first setting the damping parameter to $\varepsilon = 1$ and varying η. The upper panel of Fig. 12(d) shows a plot of the resultant trade-off between the fit to data and roughness of the solution model for a 35×35 nodes grid. The lower panel shows the curvature for this curve that is maximized at the corner of the 'L-curve'. The corner obtained at $\eta = 5$ offers a compromise between minimizing the data misfit and producing the smoothest model. In the next step, the smoothing parameter was set to $\eta = 5$ and ε was varied providing a new 'L-curve'. As in Rawlinson *et al.* (2006), the process was iterated one more time (with ε fixed and varying η) yielding an 'optimum' value for η and ε. This scheme was used to produce an optimal regularized solution for different gird sizes and results are shown in Fig. 12.

Results clearly shows that solutions are acutely dependent on the grid size. Some features are missed if the grid is too coarse as in 12(a) whereas gradients not present in the true model may appear with a too fine grid as in 12(c). Note that the solution in 12(b) has been produced with the 'perfect' grid size, as it is the same node spacing used to construct the true model in 10.

In general, the average solution obtained with the reversible jump tomography in Fig. 13(a) seems to recover the velocity anomalies with a better amplitude than any of the solutions obtained with the regularized method in Fig. 12. Interestingly, the recovered velocity map seems to be an improvement over the case where the regularized scheme uses the actual parametrization of the true model (Fig. 12b). We speculate that this may be due to the beneficial effect of sampling and averaging many solutions in the ensemble. As in the discontinuity example the Bayesian scheme looks to have detected and adapted to the local scale of the velocity anomalies without any imposed information about the

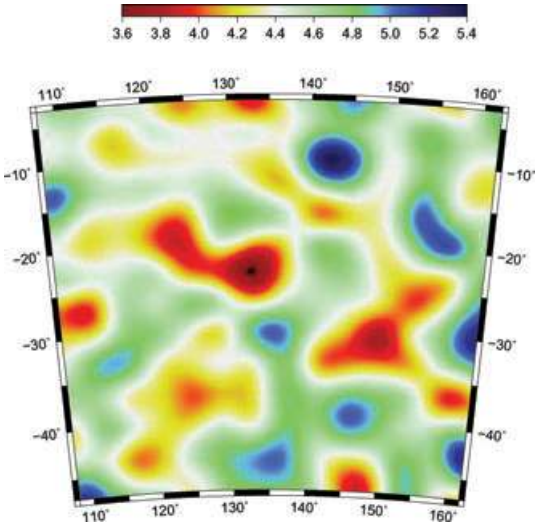


Figure 10. True velocity field (km s^{-1}). Grid of 14 × 14 nodes that is B-spline interpolated.

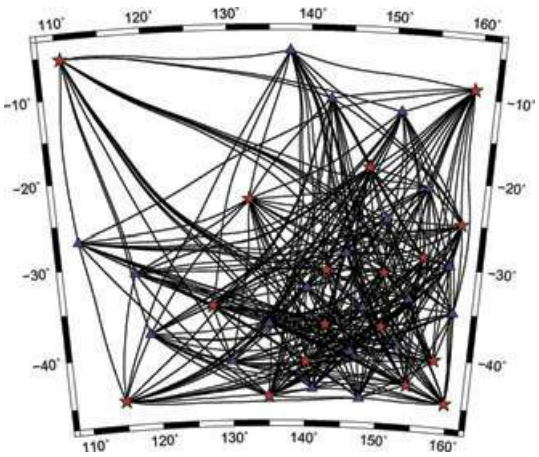


Figure 11. Geometry of rays. The sources (red stars) and receivers (blue squares) are located as in previous example. However, rays take different path due to the different velocity heterogenities.

cell sizes. Fig. 13(b) shows the posterior histogram for the number of cells recovered by the variable dimension sampler. Here the amount of detail required is much larger than the previous example. There appears to be support in the data for up to 120 cells with the mode near 70 cells. This would be consistent with the increased complexity of the true model.

4 AMBIENT NOISE DATA EXAMPLE

The use of ambient seismic noise to recover the traveltimes of surface waves between pairs of stations is rapidly becoming popular (Shapiro & Campillo 2004). There are various causes of seismic noise. The most energetic component is the oceanic microseism that is a result of the interaction of atmosphere, ocean and coast. Perturbations in the atmosphere due to strong storms impact on the ocean to set up standing wave patterns that create continuous pressure on the sea bottom, with variable intensity. The disturbance of the sea bottom results in the emergence of the elastic waves as for an earthquake (Saygin 2007). Using noise for exploration is not new. Aki (1957) and Toksöz (1964) proposed using noise records on an array to evaluate the phase velocity of the predominant surface waves. For a complete review, see Larose *et al.* (2006). Recent developments in acoustics (e.g. Derode *et al.* 2003) and seismology (Campillo & Paul 2003) showed that it is possible to perform the cross-correlation between signals recorded at two stations and extract Green's function (the signal that would be recorded at one station if an impulsive force was applied at another station). A simple demonstration of this property is based on a modal representation of a diffuse wavefield inside an elastic body (the Earth in our case) (Lobkis & Weaver 2001). Shapiro & Campillo (2004) showed that coherent Rayleigh waves can be extracted from the ambient seismic noise and that their dispersion characteristics can be measured in a broad range of periods.

This new idea created an opportunity to use an important part of the recorded wavefield of the dense networks on the Earth that is normally neglected. New measurements can be obtained for paths that could not be sampled with the ballistic waves and therefore, can significantly

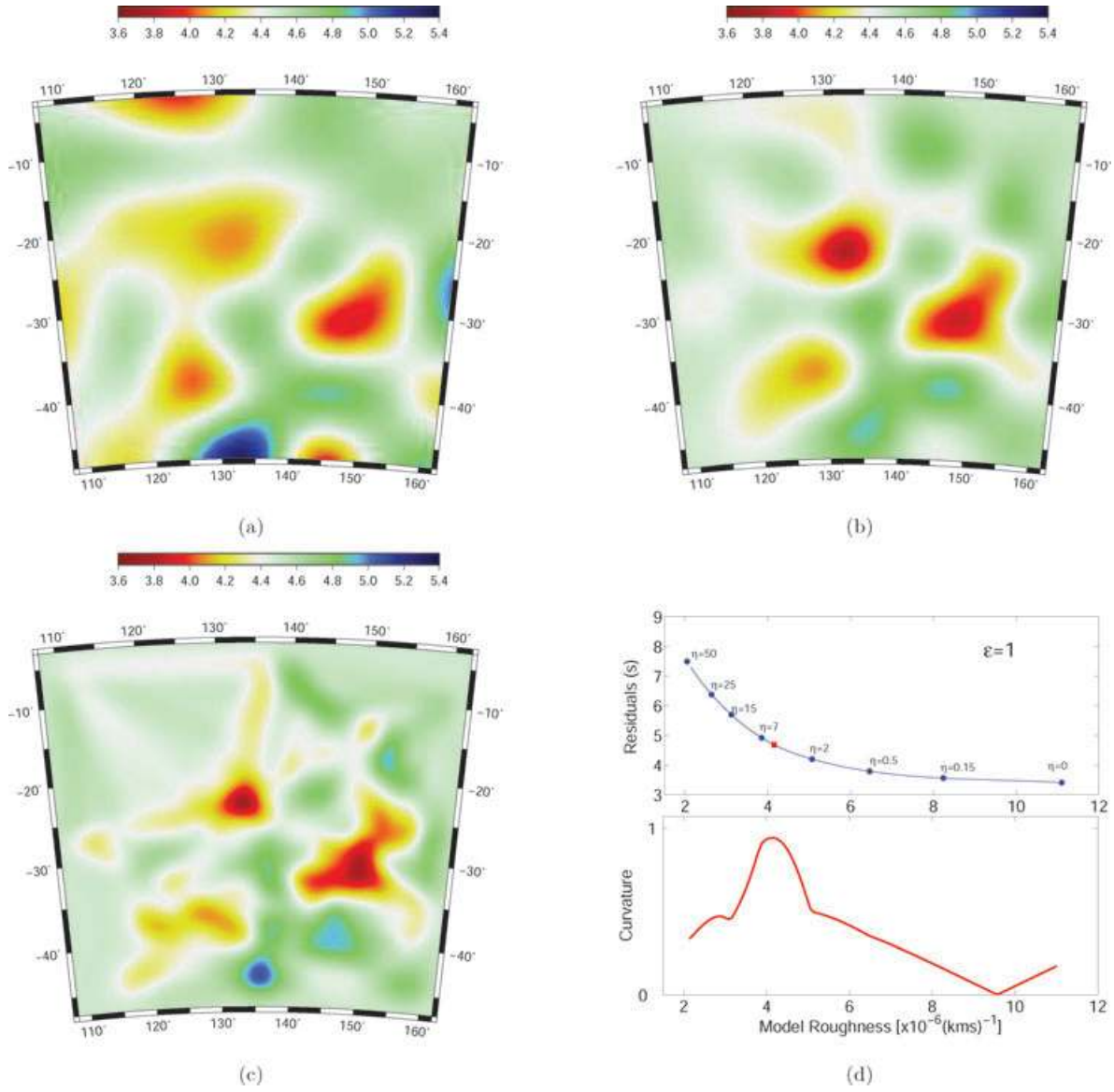


Figure 12. Results with the subspace inversion for different grid sizes. Each solution is obtained after finding the regularization parameters with an L-curve technique. (a) Grid size = 7 × 7 nodes. (b) Grid size = 14 × 14 nodes. Grid size used to produce the synthetic model. (c) Grid size = 35 × 35 nodes. (d) Upper panel: 'L-curve' for the 35 × 35 nodes grid. ε is kept constant and η is changed. Lower panel: curvature of the 'L-curve'. The maximum curvature gives the corner of the L-curve and provides the optimum η.

improve the resolution of seismic images. Saygin (2007) compiled all the seismic broad-band data from temporary and permanent stations across the Australian continent from 1992 to 2006. The data were used to calculate Green's functions between each possible station pairs that resulted in a coverage of the continent as in earthquake tomography studies with over 1000 individual ray paths. All of the available data were used for the calculation scheme from a lower duration limit of 15 d up to several years. Due to the interstation distance and spectrum characteristics of the noise field, the extracted signal was mainly Green's function of Rayleigh-type surface wave for vertical components. For each frequency, the Rayleigh wave arrival time could be picked on the envelop of the bandpass filtered seismogram. The extracted traveltimes were used to build a tomographic image of the group velocity for the Australian crust with frequency dependency.

We propose here to use the same set of measured traveltimes for a period of 5 s (0.2 Hz) and test the reversible jump tomography. Apart from dealing with observational data, this example differs from the synthetic problem in the sense that all the receivers are considered as virtual sources. The 208 stations used in the experiment are represented by blue triangles in Fig. 14. The station plays both the role of source and receiver. Despite the very large number of couples of stations (21 632), the actual number of measured traveltime is only 1158, due to only a relatively small proportion of the total number of receivers being deployed at once. The 1158 ray paths are shown in Fig. 14. They have been computed on a homogeneous velocity model so they follow great circles joining couples of stations that were deployed at the same time.

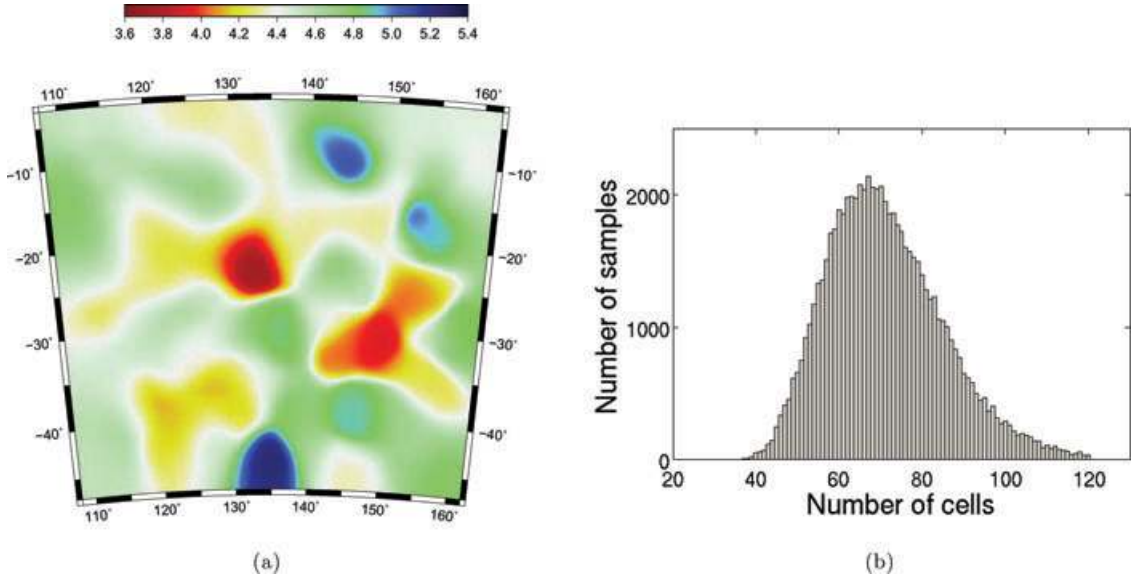


Figure 13. Reversible jump tomography. Results after three iterations (km s^{-1}). (a) Solution map obtained by averaging 10 000 post burn-in samples. The scales are the same as for previous figures. (b) Information on the number of cells in the Voronoi tessellation. Posterior probability density for the parameter n, $p(n \mid \mathbf{d}_{\text{obs}})$.

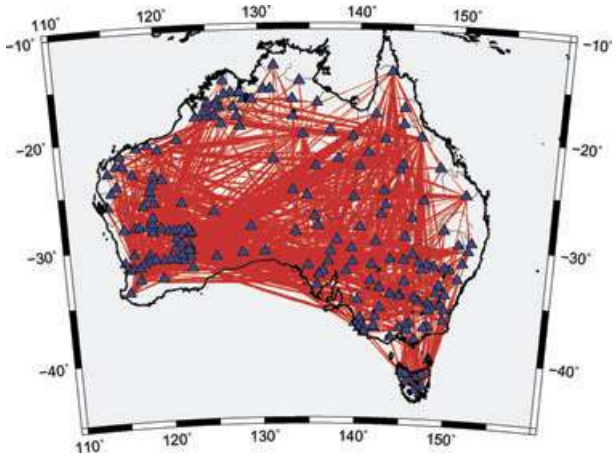
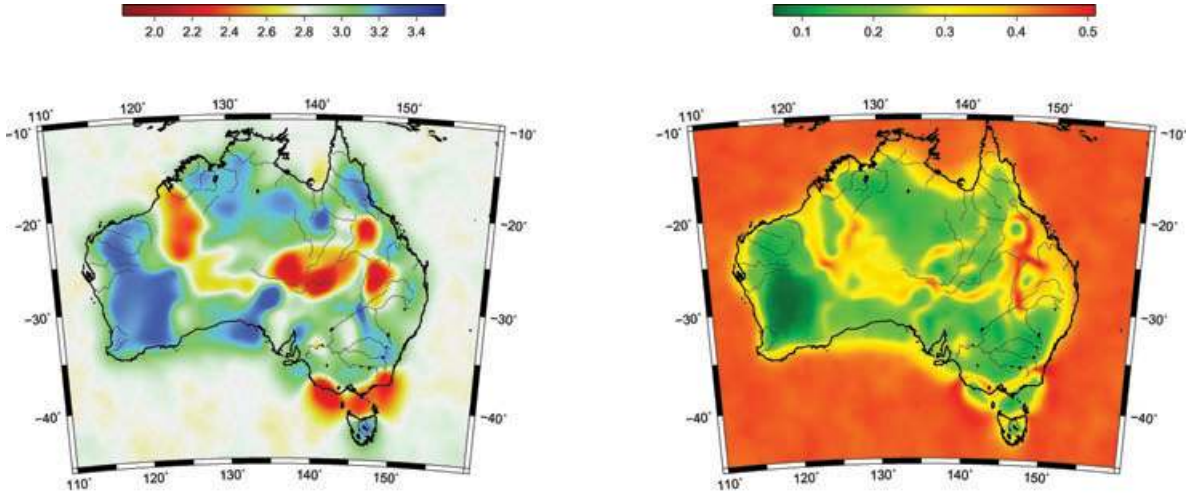


Figure 14. Map of connecting ray paths between couples of stations for 0.2 Hz. The blue triangles show the locations of the stations.

4.1 Results

Figs 15(a) and (b) show the results obtained after three iterations with the new scheme. Posterior inference was made using an ensemble of 6600 models. The rj-MCMC algorithm was run for 1.2×10^6 steps in total. The first 200 000 steps were discarded as burn-in. Then, every 150th model visited was taken in the ensemble. There are no data in the ocean areas and according to Bayesian principles, we recover the mean and the variance of the prior probability density function. Fig. 16 shows the results of the posterior on the number of cells, $p(n \mid \mathbf{d}_{\text{obs}})$. Here there are more rays and hence more information than in the synthetic example. As a consequence, the algorithm has automatically chosen to parametrize the model with many more Voronoi cells in order to fit the data.

From the modelling of the Rayleigh wave derivatives at a period of 5 s, the traveltime used here are mostly sensitive to the structure in the first 3 km of he crust (Saygin 2007). The average model in Fig. 15(a) reveals some features that can be correlated with the surface geology of the Australian continent in Fig. 17. The zones of elevated wave speed in western Australia correspond with the Pilbara and Yilgarn cratons that are fragments of ancient Archean lithosphere (Betts *et al.* 2002; Fishwick *et al.* 2005). The main Proterozoic units of the continent (i.e. Kimberley craton, Mt Isa block and George town Inler) are also visible. They represent a basement layer at the surface with no overlying soft sediment and give fast group velocities of around 3.2 km s^{-1} (Betts *et al.* 2002; Clitheroe *et al.* 2000a). Along the east coast and in Victoria, the phanerozoic orogens also show a signature on the tomographic image and give elevated wave speeds. In Central Australia, the north to south pattern of slow-fast-slow anomalies correlates closely with the presence of the Officer Basin, Musgrave Block (preserved Proterozoic orogen) and Amadeus Basin. However, it should be noted that this sector of the model is not very well constrained by the data set as can be seen in the error map [15(b)]. The short period tomographic image shows multiple low-velocity zones with velocities lower than 2.4 km s^{-1},



(a) Average solution map (km/s)

(b) Estimated error map (km/s).

Figure 15. Reversible jump tomography. Results after one iterations of 4×10^5 Markov samples.

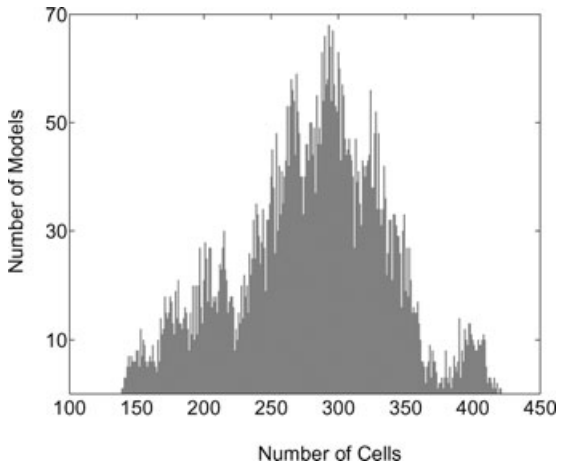


Figure 16. Posterior probability density for the parameter n.

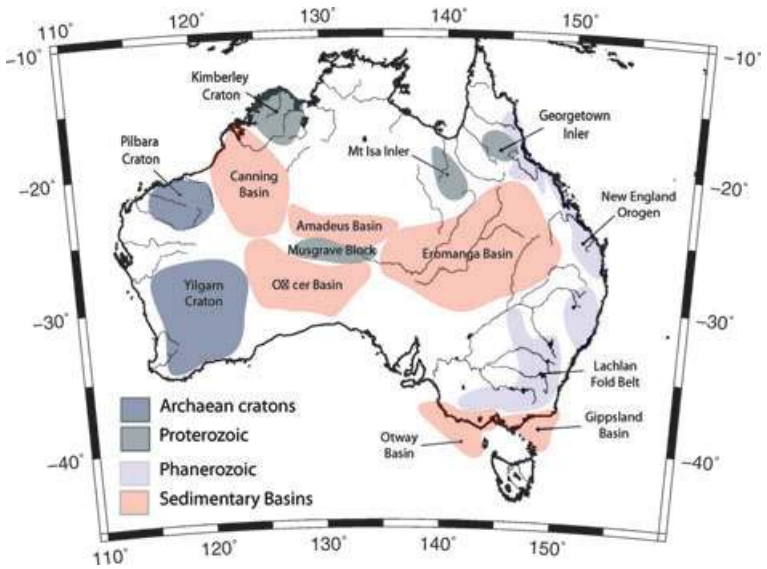


Figure 17. Surface geology of Australia.

which have a clear correspondence to regions of thick sedimentary cover. There is a general agreement on these reduced wave speed regions and for the sediment thickness map given by Clitheroe *et al.* (2000b).

Although our analysis of the geological implications of Fig. 15(a) is rather limited, the aim here is to argue that the average solution model produced by the transdimensional approach is able to recover the main geological features of the Australian continent.

Saygin (2007) inverted the same data set with the subspace method described above and obtained results consistent with ours. The results shown here are also supported by Rawlinson *et al.* (2008) where the same data were inverted with a dynamic objective function technique. The appeal of the variable dimension scheme used here is that there is no need to choose any explicit parametrization or any regularization procedure. The standard deviation map in Fig. 15(b), appears to indicate that the average solution model is optimally constrained in southeast Australia and the western half of Western Australia. This result is verified by the synthetic checkerboard resolution test given in Rawlinson *et al.* (2008) where the same source and receiver geometry was used.

5 DISCUSSION AND CONCLUSIONS

We have described a new algorithm for seismic imaging that uses the reversible jump technique (Green 1995, 2003) to adjust the underlying parametrization of the model while recovering a velocity field in 2-D. The model space is efficiently explored by sampling models of varying dimension. The output of the scheme is an ensemble of models from which properties such as a spatial average and variance can be extracted. Some notable features emerge from the results.

First, the Bayesian inference approach provides a stable solution with no need for explicit regularization, that is, there is no user-supplied damping term nor tuning of trade-off parameters. Although the overall framework is one of iterative linearized inversion, the process involves no matrix inversion, and hence no numerical procedures to stabilize matrix inverses.

Second, the procedure involves a dynamic parametrization for the model that is able to consider the spatial variability of the information provided by data. The output is an ensemble of Voronoi partitioned velocity models that are distributed according to a posterior probability distribution. When the spatial expectations (mean and variance) of the ensemble are computed, models with different cell geometry overlap providing a continuous velocity map that can be taken as a reference solution. The spatially averaged velocity map has an effective parametrization much finer than any single model, and hence, seems to better capture the variability in the range of possible solutions than a single (e.g. best) model. The averaging process naturally smooths out unwarranted structure in the Earth model, but maintains local discontinuities in wave speeds if well constrained by the data. Furthermore, the model parametrization typically involves fewer parameters and achieves a better representation of the velocity field than a fixed grid. Another feature is the ability to construct a continuous error map. Experiments suggest that it gives a reasonable estimation of the velocity uncertainty.

Application to ambient noise data from Australia shows that we are able to recover the map for Rayleigh wave group velocities without the need to impose an explicit regularization or fixed grid parametrization. Moreover, the optimization of the MCMC sampler has proved to be efficient in terms of computational costs and these preliminary results are encouraging for applications to larger data sets, different problems and extension to 3-D.

A drawback is that the design of ways to perform probabilistic transitions between models (within and between dimensions) in the Markov chain has to be implemented with the some care in order to avoid inefficiency of the algorithm. This is a common complaint with MCMC algorithms. (see Han & Carlin 2001, for an argument to suggest that transdimensional sampling may have a detrimental effect on efficiency). Traditionally, the user has to choose *a priori* the proposal probability density that will remain fixed during the sampling process. The task of manually tuning transition variables via repeated pilot runs of the chain can become laborious and quickly prohibitive. For the fixed dimension moves, we have implemented the delayed rejection scheme proposed by Tierney & Mira (1999), which helps to locally scale the proposal distributions and make the overall efficiency much less dependent on the choice of proposal distributions. An issue for future study would be to extend this scheme to transdimensional moves as shown in Green & Mira (2001) or to use recent development of assisted or automated proposal generation for transdimensional sampling scheme (e.g. Tierney & Mira 1999; Haario *et al.* 2001; Al-Awadhi *et al.* 2004; Haario *et al.* 2006).

It can also be difficult to rigorously assess convergence of the transdimensional Markov chain and hence to decide when to start collecting the sample of models and how many to collect. In our study this was not a major factor but within the Bayesian statistics literature it is an active area of research. Our results show that the reversible jump tomography represents a new and potentially powerful alternative to optimization based approaches with fixed grids and globally constrained regularization schemes.

ACKNOWLEDGMENTS

We would like to thank Kerry Gallagher for advice and discussions on this study, Alberto Malinverno, Albert Tarantola and Cliff Thurber for useful feedback, Erdinc Saygin for providing data from Australia, and Nick Rawlinson for helpful discussions during this study in particular for advice on the geological interpretation. Some calculations were performed on the Terrawulf II cluster, a computational facility supported through AuScope. AuScope Ltd is funded under the National Collaborative Research Infrastructure Strategy (NCRIS), an Australian Commonwealth Government Programme. This project was also supported by a French-Australian Science and Technology travel grant, FR090051, under the International Science Linkages program from the Department of Innovation, Industry, Science and Research, and ARC-Discovery grant 665111.

REFERENCES

Abers, G. & Roecker, S., 1991. Deep structure of an arc-continent collision: earthquake relocation and inversion for upper mantle P and S wave velocities beneath Papua New Guinea, *J. geophys. Res.,* **96**(B4), 6379–6401.

Aki, K., 1957. Space and time spectra of stationary stochastic waves, with special reference to microtremors, *Bull. Earthq. Res. Inst.,* **35,** 415–456.

Al-Awadhi, F., Hurn, M. & Jennison, C., 2004. Improving the acceptance rate of reversible jump MCMC proposals, *Stat. Probab. Lett.,* **69**(2), 189–198.

Aster, R., Borchers, B. & Thurber, C., 2005. *Parameter Estimation and Inverse Problems,* Academic Press, San Diego, CA.

Bayes, T., 1763. *An Essay Towards Solving a Problem in the Doctrine of Chances,* ed. C. Davis, Royal Society of London.

Betts, P., Giles, D., Lister, G. & Frick, L., 2002. Evolution of the Australian lithosphere, *Aust. J. Earth Sci.,* **49**(4), 661–695.

Bodin, T., Sambridge, M. & Gallagher, K., 2009. A Self-parameterising partition model approach to tomographic inverse problems, *Inverse Problems,* **25**, 055009.

Box, G.E.P. & Tiao, G.C., 1973. *Bayesian Inference in Statistical Inference*, Addison–Wesley, Reading, MA.

Brooks, S. & Giudici, P., 1999. Convergence assessment for reversible jump MCMC simulations, *Bayesian Stat.,* **6,** 733–742.

Brooks, S., Giudici, P. & Philippe, A., 2003. Nonparametric convergence assessment for MCMC model selection, *J. Comput. Graphic. Stat.,* **12**(1), 1–22.

Brooks, S., Giudici, P. & Roberts, G., 2003. Efficient construction of reversible jump Markov chain Monte Carlo proposal distributions, *J. Roy. Stat. Soc.: Ser. B (Stat. Methodol.),* **65**(1), 3–39.

Campillo, M. & Paul, A., 2003. Long-Range Correlations in the Diffuse Seismic Coda, *Science*, **299,** 547–549.

Cerveny, V. & Brown, M., 2003. Seismic ray theory, *J. Acoust. Soc. Am.,* **113,** 14.

Cerveny, V., Molotkov, I. & Psencik, I., 1977. *Ray Methods in Seismology,* Charles University, Prague.

Clitheroe, G., Gudmundsson, O. & Kennett, B., 2000a. The crustal thickness of Australia, *J. geophys. Res.,* **105**(B6), 13 697–13 713.

Clitheroe, G., Gudmundsson, O. & Kennett, B., 2000b. Sedimentary and upper crustal structure of Australia from receiver functions, *Aust. J. Earth Sci.,* **47**(2), 209–216.

Cowles, M. & Carlin, B., 1996. Markov Chain Monte Carlo convergence diagnostics: a comparative review., *J. Am. Stat. Assoc.,* **91**(434), 883–904.

Curtis, A. & Snieder, R., 1997. Reconditioning inverse problems using the genetic algorithm and revised parameterization, *Geophys.,* **62**(5), 1524–1532.

Denison, D. & Holmes, C., 2001. Bayesian partitioning for estimating disease risk, *Biometrics,* **57**(1), 143–149.

Denison, D., Adams, N., Holmes, C. & Hand, D., 2002a. Bayesian partition modelling, *Comput. Stat. Data Anal.,* **38**(4), 475–485.

Denison, D., Holmes, C., Mallik, B. & Smith, A., 2002b. *Bayesian Nonlinear Methods for Classification and Regression,* John Wiley, Chichester.

Derode, A., Larose, E., Tanter, M., de Rosny, J., Tourin, A., Campillo, M. & Fink, M., 2003. Recovering the Green's function from field-field correlations in an open scattering medium (L), *J. Acoust. Soc. Am.,* **113,** 2973.

Duijndam, A., 1988a. Bayesian estimation in seismic inversion. Part I: principles, *Geophys. Prospect.,* **36**(8), 878–898.

Duijndam, A., 1988b. Bayesian estimation in seismic inversion. Part II: uncertainty analysis, *Geophys. Prospect.,* **36,** 899–918.

Fishwick, S., Kennett, B. & Reading, A., 2005. Contrasts in lithospheric structure within the Australian craton-insights from surface wave tomography, *Earth planet. Sci. Lett.,* **231**(3-4), 163–176.

Friederich, W., 1998. Wave-theoretical inversion of teleseismic surface waves in a regional network: phase-velocity maps and a three-dimensional upper-mantle shear-wave-velocity model for southern Germany, *Geophys. J. Int.,* **132**(1), 203–225.

Fukao, Y., Obayashi, M., Inoue, H. & Nembai, M., 1992. Subducting slabs stagnant in the mantle transition zone, *J. geophys. Res.,* **97**(B4), 4809–4822.

Gelman, A. & Rubin, D., 1992. Inference from iterative simulation using multiple sequences, *Stat. Sci.,* **7,** 457–457.

Gelman, A., Roberts, G. & Gilks, W., 1996. Efficient Metropolis jumping rules, *Bayesian Stat.,* **5,** 599–607.

Gelman, A., Carlin, J., Stern, H. & Rubin, D., 2004. Bayesian Data Analysis. *Texts in Statistical Science*, Vol. 25, p. 668, Chapman & Hall, Boca Raton, FL.

Gorbatov, A., Fukao, Y., Widiyantoro, S. & Gordeev, E., 2001. Seismic evidence for a mantle plume oceanwards of the Kamchatka-Aleutian trench junction, *Geophys. J. Int.,* **146**(2), 282–288.

Gouveia, W. & Scales, J., 1998. Bayesian seismic waveform inversion- Parameter estimation and uncertainty analysis, *J. geophys. Res.,* **103**(B2), 2759–2780.

Green, P., 1995. Reversible jump MCMC computation and Bayesian model selection, *Biometrika,* **82,** 711–732.

Green, P., 2003. Trans-dimensional Markov chain Monte Carlo, *High. Struct. Stoch. Syst.,* **27,** 179–98.

Green, P. & Mira, A., 2001. Delayed Rejection in Reversible Jump Metropolis-Hastings, *Biometrika,* **88**(4), 1035–1053.

Haario, H., Saksman, E. & Tamminen, J., 2001. An adaptive Metropolis algorithm, *Bernoulli,* **7**(2), 223–242.

Haario, H., Laine, M., Mira, A. & Saksman, E., 2006. DRAM: efficient adaptive MCMC, *Stat. Comput.,* **16**(4), 339–354.

Han, C. & Carlin, B., 2001. Markov Chain Monte Carlo methods for computing bayes factors: a comparative review, *J. Am. Stat. Assoc.,* **96**(455), 1122–1132.

Hastings, W., 1970. Monte Carlo simulation methods using Markov chains and their applications, *Biometrika,* **57,** 97–109.

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.

Ivansson, S., 1986. Seismic borehole tomography-Theory and computational methods, *Proceedings of the IEEE,* **74**(2), 328–338.

Kennett, B., Sambridge, M. & Williamson, P., 1988. Subspace methods for large inverse problems with multiple parameter classes, *Geophys. J. Int.,* **94**(2), 237–247.

Larose, E. *et al.*, 2006. Correlation of random wavefields: an interdisciplinary review.

Lobkis, O. & Weaver, R., 2001. On the emergence of the Green's function in the correlations of a diffuse field, *J. Acous. Soc. Am.,* **110,** 3011.

MacKay, D., 2003. *Information theory, inference, and learning algorithms*, Cambridge Univ. Press, Cambridge.

Malinverno, A., 2002. Parsimonious Bayesian Markov chain Monte Carlo inversion in a nonlinear geophysical problem, *Geophys. J. Int.,* **151**(3), 675–688.

Malinverno, A. & Leaney, W., 2000. A Monte Carlo method to quantify uncertainty in the inversion of zero-offset VSP data, in *SEG 70th Annual Meeting,* Calgary, Alberta, The Society of Exploration Geophysicists (Expanded Abstracts).

Malinverno, A. & Leaney, W., 2005. Monte-Carlo Bayesian look-ahead inversion of walkaway vertical seismic profiles, *Geophys. Prospect.,* **53**(5), 689–703.

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

Metropolis, N. et al., 1953. Equations of state calculations by fast computational machine, *J. Chem. Phys.,* **21**(6), 1087–1091.

Mira, A., 2001. On Metropolis-Hastings algorithms with delayed rejection, *Metron,* **59**(3-4), 231–241.

Mosegaard, K., 1998. Resolution analysis of general inverse problems through inverse Monte Carlo sampling, *Inverse Problems,* **14**(3), 405–426.

Mosegaard, K. & Tarantola, A., 1995. Monte Carlo sampling of solutions to inverse problems, *J. geophys. Res.,* **100**(B7), 12–431.

Nolet, G. & Montelli, R., 2005. Optimal parametrization of tomographic models, *Geophys. J. Int.,* **161**(2), 365–372.

Nolet, G. & Panza, G., 1976. Array analysis of seismic surface waves: limits and possibilities, *Pure appl. Geophys.,* **114**(5), 775–790.

Okabe, A., Boots, B. & Sugihara, K., 1992. *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams,* John Wiley & Sons, Inc., New York, NY, USA.

Prindle, K. & Tanimoto, T., 2006. Teleseismic surface wave study for S-wave velocity structure under an array: Southern California, *Geophys. J. Int.,* **166**(2), 601–621.

Rawlinson, N. & Sambridge, M., 2003. Seismic traveltime tomography of the crust and lithosphere, *Adv. Geophys.,* **46,** 81–199.

Rawlinson, N. & Sambridge, M., 2004. Wave front evolution in strongly heterogeneous layered media using the fast marching method, *Geophys. J. Int.,* **156**(3), 631–647.

Rawlinson, N., Reading, A. & Kennett, B., 2006. Lithospheric structure of Tasmania from a novel form of teleseismic tomography, *J. Geophys. Res.-Solid Earth,* **111**(B2), B02301, doi:10.1029/2005JB003803.

Rawlinson, N., Sambridge, M. & Saygin, E., 2008. A dynamic objective function technique for generating multiple solution models in seismic tomography, *Geophys. J. Int.,* **174,** 295–308.

Robert, C., 1995. Convergence control methods for Markov Chain Monte Carlo algorithms, *Stat. Sci.,* **10,** 231–253.

Rosenthal, J., 2000. Parallel computing and Monte Carlo algorithms, *Far East J. Theoret. Stat.,* **4**(2), 207–236.

Sambridge, M. & Faletic, R., 2003. Adaptive whole Earth tomography, *Geochem. Geophys. Geosyst.,* **4**(3), 1022.

Sambridge, M. & Gudmundsson, O., 1998. Tomographic systems of equations with irregular cells, *J. geophys. Res.,* **103**(B1), 773–782.

Sambridge, M. & Rawlinson, N., 2005. Seismic tomography with irregular meshes, *Geophys. Monogr.,* **157,** 49–65.

Sambridge, M., Braun, J. & Mcqueen, H., 1995a. Geophysical parametrization and interpolation of irregular data using natural neighbours, *Geophys. J. Int.,* **122**(3), 837–857.

Sambridge, M., Braun, J. & McQueen, H., 1995b. Geophysical parametrization and interpolation of irregular data using natural neighbours, *Geophys. J. Int.,* **122**(3), 837–857.

Sambridge, M., Gallagher, K., Jackson, A. & Rickwood, P., 2006. Transdimensional inverse problems, model comparison and the evidence, *Geophys. J. Int.,* **167**(2), 528–542.

Saygin, E., 2007. Seismic receiver and noise correlation based studies in Australia, *Doctor of Philosophy thesis*, The Australian National University.

Scales, J. & Snieder, R., 1997. To Bayes or not to Bayes, *Geophysics,* **62**(4), 1045–1046.

Sethian, J. & Popovici, A., 1999. 3-D traveltime computation using the fast marching method, *Geophysics,* **64**(2), 516–523.

Shapiro, N. & Campillo, M., 2004. Emergence of broadband Rayleigh waves from correlations of the ambient seismic noise, *Geophys. Res. Lett.,* **31**(7), 1615–1619.

Sisson, S., 2005. Transdimensional Markov Chains: a decade of progress and future perspectives, *J. Am. Stat. Assoc.,* **100**(471), 1077–1090.

Sivia, D., 1996. *Data Analysis: A Bayesian Tutorial,* Oxford University Press, USA.

Smith, A., 1991. Bayesian computational methods, *Philos. Trans.: Phys. Sci. Eng.,* **337**(1647), 369–386.

Spakman, W. & Bijwaard, H., 1998. Irregular cell parameterization of tomographic problems, *Ann. Geophys,* **16,** 18.

Stephens, M., 2000. Bayesian analysis of mixture models with an unknown number of components-an alternative to reversible jump methods, *Ann. Stat.,* **28**(1), 40–74.

Stephenson, J., Gallagher, K. & Holmes, C., 2004. Beyond kriging: dealing with discontinuous spatial data fields using adaptive prior information and Bayesian partition modelling, *Geol. Soc. Lond. Special Publications,* **239**(1), 195.

Stephenson, J., Gallagher, K. & Holmes, C., 2006. Low temperature thermochronology and strategies for multiple samples 2: partition modelling for 2D/3D distributions with discontinuities, *Earth planet. Sci. Lett.,* **241**(3-4), 557–570.

Tarantola, A., 2005. *Inverse Problem Theory and Methods for Model Parameter Estimation,* Society for Industrial Mathematics.

Tarantola, A. & Valette, B., 1982. Inverse problems= quest for information, *J. Geophys.,* **50**(3), 150–170.

Tierney, L., 1994. Markov Chains for exploring posterior distributions, *Ann. Stat.,* **22**(4), 1701–1728.

Tierney, L. & Mira, A., 1999. Some adaptive Monte Carlo methods for Bayesian inference, *Stat. Med.,* **18**(1718), 2507–2515.

Toksöz, M., 1964. Microseisms and an attempted application to exploration, *Geophysics,* **29,** 154.

Virieux, J. & Farra, V., 1991. Ray tracing in 3-D complex isotropic media: an analysis of the problem, *Geophysics,* **56,** 2057.

Voronoi, G., 1908. Nouvelles applications des parametres continus a la theorie des formes quadratiques, *J. Reine Angew. Math.,* **134,** 198–287.

Yoshizawa, K. & Kennett, B.L.N., 2004. Multimode surface wave tomography for the Australian region using a three-stage approach incorporating finite frequency effects, *J. geophys. Res.,* **109,** B02310, doi:10.1029/2002JB002254.

Zhang, H. & Thurber, C., 2005. Adaptive mesh seismic tomography based on tetrahedral and Voronoi diagrams: application to Parkfield, California, *J. geophys. Res.,* **110,** B04303, doi:10.1029/2004JB003186.

APPENDIX A: DELAYED REJECTION

In a Metropolis–Hastings algorithm, rejection of proposed moves is an intrinsic part of ensuring that the chain converges to the intended target distribution. However, persistent rejection, perhaps in particular parts of the model space, may indicate that locally the proposal distribution is badly calibrated to the target posterior (Green & Mira 2001). Tierney & Mira (1999) and Mira (2001) showed that the basic algorithm can be modified so that, on rejection, a second attempt to move is made. A different proposal can be generated from a new distribution that is allowed to depend on the previously rejected proposal.

For example, let us consider the 1-D target distribution $\pi(x)$ shown in Fig. A1. The shape of this distribution is such that the 'optimal' spread of the proposal (e.g. the variance of a Gaussian proposal distribution) depends on the current position of the chain. When x takes low values, the spread should be quite small, otherwise proposals are likely to be rejected. On the other hand, using the same small spread when x is large will give high acceptance rate, but the chain is going to explore this portion very slowly. Every time we find ourselves in such situations, and this happens quite often for multidimensional target distributions, the delayed rejection algorithm can be of great help (Green & Mira 2001).

When at x, we propose a new state y_1 with density $q_1(y_1|x)$. As in Hastings (1970), this is accepted with probability

$$\alpha_1(y_1\,|\,x) = \min\left[1, \frac{\pi(y_1)}{\pi(x)}\frac{q_1(x\,|\,y_1)}{q_1(y_1\,|\,x)}\right]. \tag{A1}$$

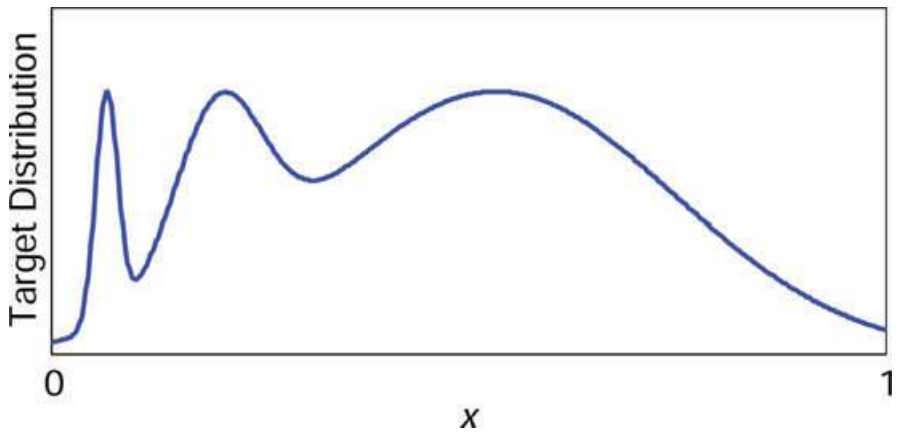


Figure A1. An example of a 1-D probability density function with variable 'length scale'.

Tierney & Mira (1999) propose that, if the move to y_1 is rejected, a second proposal y_2, say, is made, with density $q_2(y_2|x, y_1)$. The acceptance probability of the new candidate has to be determined in order to preserve the stationary distribution. Tierney & Mira (1999) use

$$\alpha_2(y_2\,|\,x, y_1) = \min\left[1, \frac{\pi(y_2)}{\pi(x)}\frac{q_1(y_1\,|\,y_2)}{q_1(y_1\,|\,x)}\frac{q_2(x\,|\,y_2, y_1)}{q_2(y_2\,|\,x, y_1)}\frac{\{1-\alpha_1(y_1\,|\,y_2)\}}{\{1-\alpha_1(y_1\,|\,x)\}}\right] \tag{A2}$$

that preserves detailed balance at the second stage, as shown in Mira (2001). If the candidate at the second stage is also rejected we could either stay in the current state x or move on to a third stage, and so on. Since detailed balance is imposed on each stage separately, it is valid to make any fixed or random number of attempts. If q_i denotes the proposal at the ith stage, the acceptance probability at that stage is, following Mira (2001),

$$\alpha_i(y_i\,|\,x, y_1, \ldots, y_{i-1}) = \min\left[1, \frac{\pi(y_i)}{\pi(x)}\frac{q_1(y_{i-1}\,|\,y_i)q_2(y_{i-2}\,|\,y_i, y_{i-1})\ldots q_i(x\,|\,y_i, y_{i-1}, \ldots y_1)}{q_1(y_1\,|\,x)q_2(y_2\,|\,x, y_1)\ldots q_i(y_i\,|\,x, y_1, \ldots, y_i)}\right.$$
$$\left.\frac{\{1-\alpha_1(y_{i-1}\,|\,y_i)\}\{1-\alpha_2(y_{i-2}\,|\,y_i, y_{i-1})\}\ldots\{1-\alpha_{i-1}(y_1\,|\,y_i, y_{i-1}, \ldots, y_2)\}}{\{1-\alpha_1(y_1\,|\,x)\}\{1-\alpha_2(y_2\,|\,x, y_1)\}\ldots\{1-\alpha_{i-1}(y_{i-1}\,|\,x, y_1, \ldots, y_{i-2})\}}\right]. \tag{A3}$$

The implementation of a general delayed rejection algorithm using (A3) appears to be non-trivial due to its recursive nature.

In the reversible- jump tomography, we have used delayed rejection for the fixed dimension moves (i.e. for a velocity value update or for a nucleus move). We start with a first-stage Gaussian proposal $q_1(\mathbf{m}'|\mathbf{m})$ with large variance. The proposed model is accepted with probability

$$\alpha_1(\mathbf{m}'\,|\,\mathbf{m}) = \min\left[1, \frac{p(\mathbf{m}'\,|\,\mathbf{d}_{\rm obs})}{p(\mathbf{m}\,|\,\mathbf{d}_{\rm obs})}\right]. \tag{A4}$$

If the model $\mathbf{m}'$ is rejected, instead of going back to $\mathbf{m}$ and counting it twice in the chain, we propose a second model $\mathbf{m}''$ drawn form a second-stage Gaussian proposal $q_2(\mathbf{m}''|\mathbf{m})$ centred at $\mathbf{m}$ but with a reduced variance. Note that here, the second proposal does not depend on the rejected model $\mathbf{m}'$ and the acceptance term for the second try in (A2) is simplified

$$\alpha_2(\mathbf{m}''\,|\,\mathbf{m}) = \min\left[1, \frac{p(\mathbf{m}''\,|\,\mathbf{d}_{\rm obs})}{p(\mathbf{m}\,|\,\mathbf{d}_{\rm obs})}\frac{q_1(\mathbf{m}'\,|\,\mathbf{m}'')}{q_1(\mathbf{m}'\,|\,\mathbf{m})}\frac{\{1-\alpha_1(\mathbf{m}'\,|\,\mathbf{m}'')\}}{\{1-\alpha_1(\mathbf{m}'\,|\,\mathbf{m})\}}\right]. \tag{A5}$$

The advantage of this strategy is apparent in our problem. Due to the irregular distribution of the information, the optimal variances of the proposals are not constant throughout the velocity field. Proposed moves that imply Voronoi cells in well-constrained regions need to be smaller compare to moves that take pace in low ray density areas where model uncertainty is higher.

APPENDIX B: THE JACOBIAN

By definition, the Jacobian only needs to be calculated when there is a jump between two models of different dimensions, that is when a birth or death is proposed. If the current and proposed model have the same dimension, the Jacobian term is 1, and can be ignored.

For a birth step, the bijective transformation h used to go from $\mathbf{m}$ to $\mathbf{m}'$ writes

$$(\mathbf{c}, \mathbf{v}, \mathbf{u}_c, \mathbf{u}_v) \longleftrightarrow (\mathbf{c}, \mathbf{v}, \mathbf{c}'_{n+1}, v'_{n+1}) = \mathbf{m}'. \tag{B1}$$

The random variable $\mathbf{u}_c$ used to propose a new nucleus $\mathbf{c}_{n+1}$ is drawn from a discrete distribution defined on the integers $[0, 1, \ldots, N-n]$. The random number $\mathbf{u}_v$ is drawn from a Gaussian distribution centred at 0 and the velocity assigned to the new cell is given by

$$v'_{n+1} = v_i + \mathbf{u}_v, \tag{B2}$$

where v_i is the current velocity value where the birth takes place.

Note that the model space is divided into a discrete space (nuclei position) and a continuous space (velocities). $\mathbf{u}_c$ is a discrete variable used for the transformation between discrete spaces and $\mathbf{u}_v$ is a continuous variable used for the transformation between continuous spaces. Denison *et al.* (2002a) showed that the Jacobian term is always unity for discrete transformations. Therefore, the Jacobian term only accounts for the change in variables from

$$(\mathbf{v}, \mathbf{u}_v) \longleftrightarrow (\mathbf{v}, v'_{n+1}) = \mathbf{v}'. \tag{B3}$$

Hence, we have

$$|\mathbf{J}|_{birth} = \left| \frac{\delta(\mathbf{v}')}{\delta(\mathbf{v}, \mathbf{u}_v)} \right| = \left| \frac{\delta(\mathbf{v}, v'_{n+1})}{\delta(\mathbf{v}, \mathbf{u}_v)} \right| = \left| \frac{\delta(v_i, v'_{n+1})}{\delta(v_i, \mathbf{u}_v)} \right| = \begin{vmatrix} 1 & 0 \\ 1 & 1 \end{vmatrix} = 1 \tag{B4}$$

So it turns out that for this style of birth proposal the Jacobian is also unity. Since the Jacobian for a death move is $|\mathbf{J}|_{\text{death}} = |\mathbf{J}^{-1}|_{\text{birth}}$, this is also equal to one. Conveniently, then the Jacobian is unity for each case and can be ignored.