A self-parametrizing partition model approach to tomographic inverse problems

T Bodin[1], **M Sambridge**[1] **and K Gallagher**[2]

[1] Research School of Earth Sciences, Australian National University, Canberra, ACT 0200, Australia
[2] Géosciences Rennes, Université de Rennes 1, Rennes 35042, France

E-mail: thomas.bodin@anu.edu.au

Abstract
Partition modelling is a statistical method for nonlinear regression and classification, and is particularly suited to dealing with spatially variable parameters. Previous applications include disease mapping in medical statistics. Here we extend this method to the seismic tomography problem. The procedure involves a dynamic parametrization for the model which is able to adapt to an uneven spatial distribution of the information on the model parameters contained in the observed data. The approach provides a stable solution with no need for explicit regularization, i.e. there is neither user supplied damping term nor tuning of trade-off parameters. The method is an ensemble inference approach within a Bayesian framework. Many potential solutions are generated, and information is extracted from the ensemble as a whole. In terms of choosing a single model, it is straightforward to perform Monte Carlo integration to produce the expected Earth model. The inherent model averaging process naturally smooths out unwarranted structure in the Earth model, but maintains local discontinuities if well constrained by the data. Calculation of uncertainty estimates is also possible using the ensemble of models, and experiments with synthetic data suggest that they are good representations of the true uncertainty.

(Some figures in this article are in colour only in the electronic version)

1. Introduction

In tomographic inversion, the aim is to recover some internal physical properties (e.g. density, composition, temperature) from a set of surface measurements. More specifically, tomographic imaging deals with reconstructing an image from data that are projections of a physical property of the body we want to image. It is widely used in a large number of different

fields from diagnostic medicine to materials sciences. The projections can represent, for example, the attenuation of x-rays through an object, as in conventional x-ray tomography, or the refractive index variations as in ultrasonic tomography (Natterer 2001). The technique consists in gathering these projection data from multiple directions and feeding the data into a tomographic reconstruction algorithm. Algorithms differ according to the number and the distribution of projections available, as well as in the nature of the relationship between the model and the data.

The 'true' model one tries to infer from observed data is often a continuous function of space: $M(\mathbf{x})$. In practice, this model is typically approximated as $m(\mathbf{x})$, which is a linear combination of a finite number, N, of basis functions B_i:

$$M(\mathbf{x}) \approx m(\mathbf{x}) = \sum_{i=1}^{N} a_i B_i(\mathbf{x}), \tag{1}$$

where a_i are the model parameters. Both N and the basis functions have to be defined. The basis function can be either locally supported, e.g. cubic cells or globally supported, e.g. spherical harmonics (see, Sambridge and Rawlinson (2005) for a discussion).

Obviously, the choice of the parametrization has a major impact on the formulation of the problem, especially in terms of the relation between the model $\mathbf{m}$ and the observed data $\mathbf{d}$:

$$\mathbf{d} = g(\mathbf{m}). \tag{2}$$

Moreover, linearity, resolution, ill-posedness and model uncertainty are concepts directly influenced by the parametrization. In practice, the choice of the model parametrization constitutes prior information about a function $M(\mathbf{x})$ which we often do not know well. A good parametrization enables one to extract the maximum information from the data, without introducing unjustified complexity to the model structure.

A common feature of many tomography problems is that the number and the distribution of projections are limited due to the irregular spatial distribution of sources and logistical constraints which control position of receivers. In seismology, for example, the sources are natural earthquakes, and the receivers (seismic recording stations) are typically restricted to continental regions. However, the problem of having incomplete data and uneven distribution of the information is also common in other fields. Electron microscopy is a typical example. In transmission electron microscopy, an electron beam passes through a planar specimen under several incidence angles. Since the beam has to traverse the specimen more or less transversally, the incidence angle is restricted to an interval less than 180°, typically 120° (Hoppe and Hegerl 1980). In other applications, the radiating sources are inside the object and it is the distribution of sources which is sought for. An example is emission computed tomography in nuclear medicine (Budinger *et al* 1979).

In this contribution, we develop an entirely new approach to the tomography problem based on the concept of partition modelling and we formulate this conceptually in terms of a self-adapting parametrization. Partition modelling (e.g. Denison *et al* (2002)) has been previously developed as a statistical analysis tool for nonlinear classification and regression. The methodology has been applied successfully, for example, in disease mapping (e.g. Denison and Holmes (2001)). More recently, it has been exploited in a variety of Earth Sciences applications including geostatistics (Stephenson *et al* 2004), thermochronology (e.g. Stephenson *et al* (2006)) and palaeoclimate inference (e.g. Hopcroft *et al* (2007)). However, the method we present has the potential to be applied in other fields than geophysical imaging and will be presented as a general tomographic inversion strategy.

In the following section, we briefly introduce the seismic tomography problem. In section 3, we give a detailed presentation of the Bayesian (seismic) tomography methodology

based on partition modelling in 2D, although the concepts generalize readily to 3D. In section 4, we present results of synthetic experiments in a situation where the ray coverage is not uniform in order to compare our approach to standard methods. We use both perfect synthetic data and data contaminated with random noise in order to test the ability of the method to infer model uncertainty in 2D problems. In the final section, we summarize the method and main results, and suggest directions for further work.

2. Seismic tomography

Tomography has been an active domain of research in seismology for 30 years (for a recent review, see Rawlinson and Sambridge (2003) and is used for investigating the internal structure and composition of our planet. The data used in tomography basically consist in arrival times of several types of elastic waves which have propagated from a seismic source to a receiver. These arrival times are read and picked on the waveforms recorded at the receiver. The speed of the seismic waves is directly linked to physical properties of the earth. Thus, a seismic wave propagating from a source to a receiver contains information about the earth structure within its volume of influence. This information obtained is strongly dependent on the location of the seismic sources (most at plate boundaries) and the positions of the receivers, which results in some regions being traversed by many seismic rays and other regions left with poor ray coverage.

In seismic tomography, most of the models are parametrized with basis function consisting of uniform local cells in 2D or 3D. These kinds of models have sharp discontinuities and usually need to be spatially smoothed. Most inversion schemes used in geophysics to date are based on optimization of a data misfit criterion and almost always contain damping procedures (Aster *et al* 2005). In seismic tomography with regular grids, it is well known that the size of the cells is chosen according to a compromise between model resolution and model uncertainty (Backus and Gilbert 1968). If the cells are large, a lot of independent information can be integrated to give a mean velocity value that will not be biased by the noise in the data. The uncertainty on the estimated velocity will be small at the expense of its 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 eventually, the solution will become non-unique.

The irregular distribution of information mentioned earlier has led some seismologists to use irregular meshes and allow for their refinement during the inversion process (see Sambridge and Rawlinson (2005) for a recent review). The use of irregular meshes in seismic tomography introduces many different implementation problems. A range of approaches have been proposed and applied to various problems. Abers and Roecker (1991) introduced a scheme where fine scale regular 3D blocks are joined to form larger irregular cells, and applied the techniques to image P-wave velocity structure beneath Papua New Guinea. Fukao *et al* (1992), Spakman and Bijwaard (1998) used non-uniform sized rectangular 3D blocks to account for uneven ray path sampling. Curtis and Snieder (1997) set up the inversion mesh adaptively to minimize the condition number of the resulting tomographic system of equations for the cross borehole tomographic problem by means of a genetic algorithm. Sambridge *et al* (1995a), Sambridge and Gudmundsson (1998) were the first to propose the use of Delaunay tetrahedra and Voronoi polyhedra for tomographic problems. These are completely unstructured meshes, e.g. not based on a cubic or other regular grid (Voronoi 1908, Okabe *et al* 1992), and have the advantage of adapting to the spatial variability of the information provided by the data. Sambridge and Faletic (2003), Zhang and Thurber (2005) developed adaptive mesh seismic tomography methods based on tetrahedral and Voronoi diagrams to
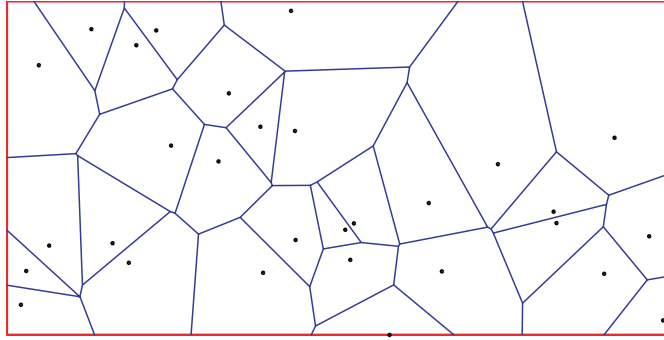
Figure 1. Voronoi cells about 30 pseudo random points on the plane. The cell nuclei have been drawn from a 2D uniform distribution over the spatial domain delimited by the red rectangle. The cell boundaries are defined as the perpendicular bisectors of pairs of nuclei. Any point inside a cell is closer to the nucleus of that cell than any other nucleus.

automatically match the inversion mesh to the data distribution. Nolet and Montelli (2005) optimized the spacing of interpolation support to fit local resolution by connecting natural neighbours with springs of length equal to the local resolving length and minimizing the potential energy of the system.

The partition modelling approach to the tomography problem we propose here follows on from the idea of using irregular meshes. The model space is partitioned by a given number of non-overlapping regions defined by a Voronoi tessellation (see figure 1) (Okabe *et al* 1992). The partition is defined by a discrete set of points (or Voronoi nuclei), and each region (or Voronoi cell) encloses all the points of the space that are closer to its nucleus than to any other Voronoi nucleus. That is, the position of each cell defining the geometry of the velocity field and the velocity field itself are unknowns in the inversion. The inversion is carried out with a fully nonlinear parameter search method based on Markov chain Monte Carlo formulated in a Bayesian framework. As we shall show the mesh self-adapts to the information contained in the data in a manner quite different to the methods described in the previous paragraph.

3. Method

3.1. The model parametrization

In 2D, the seismic velocity field is discretized by a set of Voronoi polygons as shown in figure 1. Given a set $\mathbf{C}$ of n nuclei in the 2D plane ($\mathbf{C} = \{\mathbf{c}_1, \ldots, \mathbf{c}_n\}$ where $\mathbf{c}_i \epsilon \mathbb{R}^2$), the Voronoi tessellation defines 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 $\mathbf{c}_j$ ($j \neq i$). Note that the cell nuclei are not necessarily in the geometric centre, rather the cell boundaries are perpendicular bisectors of each pair of neighbouring nuclei. Each cell is therefore characterized by the two coordinates of its nucleus $\mathbf{c}_i$ and for the seismic tomography problem, by a constant seismic wave velocity value v_i. Therefore, the model parameters are encapsulated by $\mathbf{m} = (\mathbf{C}, \mathbf{V})$ where $\mathbf{V}$ is the vector of the velocity values assigned to each partition ($\mathbf{V} = (v_1, \ldots, v_n)$ where $v_i \epsilon \mathbb{R}$).

The number of unknowns, i.e. the dimension of the model, is therefore $3n$. During the inversion, we fix the dimension of the problem so that the Voronoi cells always partition the plane into n non-overlapping regions. However, the position of the nuclei is variable so the cells can take different sizes and shapes, and the velocity in each cell is allowed to vary.
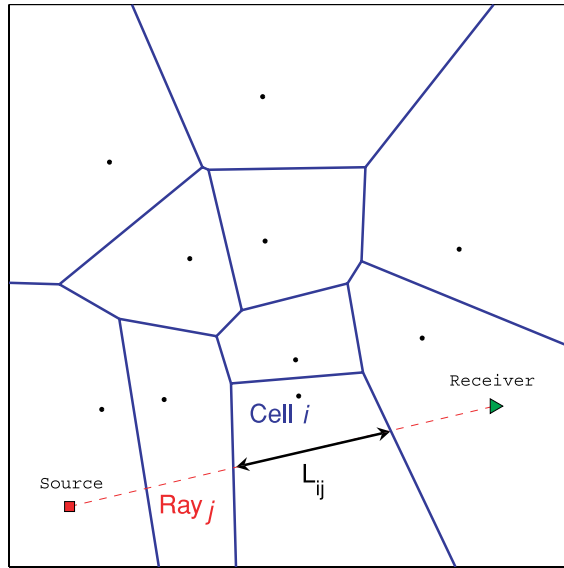
Figure 2. A seismic ray (dashed red line) joins a source to a receiver with a straight line. The travel time t_j is simply computed by integrating the inverse of the cell velocity v_i along the ray path. That is, using the length L_{ij} of the ray across each cell.

The cell geometry and the grid resolution will then be directly determined by the data. We will show that in the partition modelling approach, this dynamic parametrization will adapt to the spatial variability in the resolving power of the data.

3.2. The forward model

The physical theory of seismic wave propagation allows us to make predictions: given a complete description of a velocity field, we can calculate travel times of direct phases and compare them to actual measurements. In the high-frequency approximation case considered here we use ray theory (Cerveny *et al* 1977, Cerveny and Brown 2003). One can simulate the propagation of a wave from a point source to a receiver and predict the travel time taken. The computation of the travel times is made by integrating the slowness field, i.e. the inverse of the velocity, along the ray paths. In this work, we assume straight rays between pairs of points, although the basic methodology we present is applicable to more complex ray geometries. The travel time of the jth ray is then simply 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 figure 2) and v_i is the velocity value assigned to cell i. If ray j does not passes through cell i, then $L_{ij} = 0$. Equation (3) will be recognized by the reader as a standard linear tomographic system of equations.

  Note that in a variable velocity medium, ray paths are not straight lines and are curved according to the velocity model. However, it is valid to consider straight rays if the fluctuations of velocity on a background constant value are small. This assumption is reasonable for many seismic problems and is also relevant in some practical cases such as in x-ray tomography, ultrasonic computed tomography (Natterer 2001) or teleseismic tomography (e.g. Aki *et al*

(1977), Graeber *et al* (2002)). In the examples we subsequently consider, the rays remain straight and do not depend on the velocity field.

3.3. The data

In our problem, the observations, or data, are the first arrival travel times of seismic waves propagating across a 2D velocity field between source–receiver pairs. We use the forward model described above to generate synthetic arrival times. The time of the source is known *a priori* and the data then consist of measuring the time taken by the wave front to travel from the source to the receiver. The inversion of arrival time data is a general tomographic problem and has been addressed in a large number of different problems, such as seismic surface wave tomography (e.g. Nolet and Panza (1976), Friederich (1998), Prindle and Tanimoto (2006)) or cross-hole seismic body wave tomography (e.g. Ivansson (1986)). In examples we develop in the subsequent sections, we first consider noise-free synthetic data and later add noise to see how it propagates into model uncertainty.

3.4. Bayesian formulation

Having parametrized the velocity field and formulated the forward model to make predictions for any particular model, we now describe the inversion approach based on a Bayesian framework using partition modelling.

In a Bayesian approach all information is represented in probabilistic terms. Standard references for Bayesian inference are by Box and Tiao (1973) and useful books are by Smith (1991), Gelman *et al* (2004). Summaries within a geophysical context are given by Tarantola and Valette (1982), Duijndam (1988), Mosegaard and Tarantola (1995). The aim is to recover the posterior distribution: that is, the probability density of the model parameters given the observed data (Smith 1991). Each individual model has an associated probability, conditional on the data. The posterior distribution is therefore a function of the unknown parameters defining the model. If the model is defined by $3n$ unknowns, the posterior will be of dimension $3n$ (where n is the number of Voronoi nuclei). This multi-dimensional probability distribution is taken as the complete solution of the inverse problem.

Bayes' rule (Bayes 1763) is used to combine prior information on the model with the observed data to give the posterior probability density function:

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

where $x|y$ means x given, or conditional on, y, i.e. the probability of having x when for a given value of y. $\mathbf{d}_{\text{obs}}$ is a vector defined by the set of observed data and $\mathbf{m}$ is the vector of the model parameters. The term $p(\mathbf{d}_{\text{obs}} \mid \mathbf{m})$ is the likelihood function, which is the probability of observing the measured data given a particular model $\mathbf{m}$. $p(\mathbf{m})$ is *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, $p(\mathbf{d}_{\text{obs}})$, is often referred to as the Evidence (e.g. Sambridge *et al* (2006)) and is equivalent to the numerator on the right-hand side integrated over all possible models. In our context, this can be regarded as constant since it is not a function of any particular model $\mathbf{m}$. We therefore write (4) as a proportionality relationship:

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

Thus, the posterior distribution can be considered to represent how our prior knowledge of the model parameters is updated once we have some observed data. Clearly, if the prior and the posterior distributions are the same, then we have learnt nothing new from the data.

Once we have a reliable estimate of the posterior probability density function (in terms of an ensemble of samples), then it is straightforward to extract individual models (e.g. the best or expected average model) to construct marginal probability distributions for individual model parameters and infer credible regions or ranges on parameters. Correlations between parameters can also be examined directly (Gelman *et al* 2004).

3.5. The likelihood

The likelihood $p(\mathbf{d}_{\text{obs}} \mid \mathbf{m})$ is a quantitative measure of how well a given model with a particular set of parameter values can reproduce the observed data. Using a given physical theory, in our case the propagation of seismic rays, the forward problem is solved for a particular model, providing an estimate of the data that would be measured for that model. In our problem, the likelihood is based on a least-squares misfit function, which quantifies the agreement between simulated and observed data. If the estimated data are close to the observed data, the model tested is close to the true model and the misfit is small

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

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

The data noise takes into account the precision with which arrival times are picked on the wave form (random error) but also the accuracy of the earthquake origin time (systematic error). This uncertainty is estimated by different means depending on the seismic phases (e.g. body waves or surface waves), and on the way the arrival times are measured (e.g. manual or automatic picking, stacking). See Billings *et al* (1994) for a discussion about manual picking and Diehl *et al* (2009) for a recent example of automatic picking. It is explained below how the estimated data uncertainty affects the inversion results.

As is well known, minimizing the least-squares misfit function is equivalent to maximizing the probability for a Gaussian likelihood function, i.e.

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

3.6. The prior

The Bayesian formulation enables one to account for prior knowledge, provided that this information can be expressed as a probability distribution $p(\mathbf{m})$ (Scales and Snieder 1997, Gouveia and Scales 1998). It represents information on the model. All inferences from the data are then relative to this prior. In the seismic tomography problem, this prior information is what we think is reasonable for the velocity field we want to map, according to previous studies.

In this work, we assume minimal prior knowledge and use a uniform prior distribution (note that all priors contain information and it is not possible to have a completely uninformative prior). The prior probability distributions for the $3n$ parameters, 2D Voronoi centres and velocities in each cell, are independent from each other, and so can be written in separate form:

$$p(\mathbf{m}) = p(\mathbf{C})p(\mathbf{V}). \tag{8}$$

For velocity, the prior is specified by a constant value over a defined velocity interval $\Delta V = (V_{\text{max}} - V_{\text{min}})$. Hence we have

$$p(v_i) = \begin{cases} 1/(\Delta V) & \text{if} \quad V_{\text{min}} < v_i < V_{\text{max}} \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

and since the velocity in each cell is independent,

$$p(\mathbf{V}) = \prod_{i=1}^{n} p(v_i). \tag{10}$$

For the position of the cell nuclei, we also use a uniform distribution, and so define a rectangular area A of the 2D seismic field where the nuclei of all the cells must lie with equal probability. This rectangle is represented in red in figure 1 and borders the area to map, i.e. the zone covered by seismic rays. Hence we have

$$p(\mathbf{c}_i) = \begin{cases} 1/(\Delta x \, \Delta y) & \text{if} \quad \mathbf{c}_i \in A \\ 0 & \text{otherwise,} \end{cases} \tag{11}$$

where Δx and Δy are the dimensions of the spatial domain A. Since the position of each nucleus is independent, we also have

$$p(\mathbf{C}) = \prod_{i=1}^{n} p(\mathbf{c}_i). \tag{12}$$

Clearly, the probability of any parameter lying outside the range of the relevant prior is zero.

3.7. Principles of Markov chain Monte Carlo

In the absence of convenient analytical solutions, the only practical way to determine the posterior is to evaluate it at different positions in the model space (which is the essence of Bayesian sampling). As the dimension of the model space increases, the number of models to test becomes huge due to the curse of dimensionality, and a uniform sampling or complete enumeration of the posterior is not practical.

The Markov chain Monte Carlo (MCMC) method is an iterative stochastic approach whose aim is to generate samples from the posterior probability density. Useful introductions to this methodology are given by Gilks *et al* (1996), Sivia (1996). The starting model is selected randomly, and the choice of the next model of the 'chain' is based on a proposal probability distribution and only depends on the current state of the model. After generating a number of samples, called burn-in period, the random walk starts to produce an importance sampling of the model space. That is, the models sampled by the chain are asymptotically distributed according to the posterior probability distribution (Tierney 1994), a state referred to as the chain being stationary. Given the posterior distribution, (or at least a discrete approximation to it), it is straightforward to determine the mean and the standard deviation from the distribution of the post burn-in samples.

It is important to note that an MCMC simulation of r iterations does not produce r independent samples from the posterior distribution. The samples are correlated due to the nature of the Markov chain sampling. Given that we want to generate a number of independent samples from the given distribution, it is often necessary to 'thin' the chain. This involves collecting models only every t iterations of the chain where t is the relaxation time of the random walk, i.e. the number of steps before we can expect to get a model that is roughly independent of the last model collected. The parameter t depends on the length scales of the model space and of the probability distributions used. Theoretical and practical details can be found in MacKay (2003), Neal (1993). Another way to produce independent samples is to parallelize the algorithm. Different independent chains can be run at the same time and simultaneously sample the posterior. This is also advantageous in terms of computational time (see section 3.9).
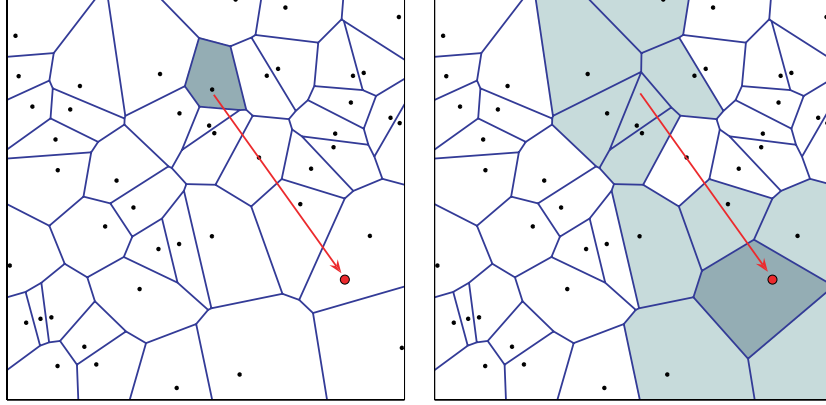
Figure 3. Move of a Voronoi nucleus. The two panels represent the Voronoi tessellation before and after the move. The new position (red circle) is drawn from a 2D Gaussian probability distribution centred at the old position of the nucleus. Note that only the cells in grey are affected by the move.

Having randomly initialized the model parameters by drawing values from the prior distribution of each parameter, the algorithm proceeds iteratively. At each iteration of the chain, we choose one Voronoi cell at random and either update its position $\mathbf{c}_i$ or its seismic velocity value v_i. Each step is divided into three stages:

(i) Randomly pick one cell (from a uniform distribution) and propose a new model by drawing from a probability distribution $q(\mathbf{m}' \mid \mathbf{m})$ such that the new proposed model $\mathbf{m}'$ is conditional only on the current model $\mathbf{m}$.

  • At every even iteration: Randomly change the velocity value of the cell according to a Gaussian proposal probability density $q(v_i' \mid v_i)$ centred at the current value v_i. The variance of the Gaussian function is a parameter to be chosen.
  • At every odd iteration: Randomly change the position of the cell nucleus according to a 2D Gaussian proposal probability density $q(\mathbf{c}'_i \mid \mathbf{c}_i)$ centred at the current position $\mathbf{c}_i$ (see figure 3). The covariance matrix for the 2D Gaussian function is proportional to the identity matrix, with the constant of proportionality a parameter to be chosen.

(ii) Solve the forward problem: Sort the rays passing through the cells, whose geometry or velocity has been modified, and with (3) update the estimated travel times for only the rays affected. The new estimated travel times are compared with the observations to build the misfit (6), the likelihood (7) and the posterior value of the proposed model $p(\mathbf{m}' \mid \mathbf{d}_{\mathrm{obs}})$.

(iii) Decide whether or not to replace or update the current model with the proposed model by drawing from a uniform random variate, U, (between 0 and 1) and using an acceptance criterion which takes the form (Gilks *et al* 1996)

$$p(\mathrm{accept}) = \min\left(1, \frac{p(\mathbf{m}' \mid \mathbf{d}_{\mathrm{obs}})}{p(\mathbf{m} \mid \mathbf{d}_{\mathrm{obs}})} \cdot \frac{q(\mathbf{m} \mid \mathbf{m}')}{q(\mathbf{m}' \mid \mathbf{m})}\right). \tag{13}$$

Assuming symmetric proposal distributions (i.e. $q(\mathbf{m}' \mid \mathbf{m}) = q(\mathbf{m} \mid \mathbf{m}')$), we have

$$p(\mathrm{accept}) = \min\left(1, \frac{p(\mathbf{m}' \mid \mathbf{d}_{\mathrm{obs}})}{p(\mathbf{m} \mid \mathbf{d}_{\mathrm{obs}})}\right). \tag{14}$$

This form is the well-known Metropolis–Hastings algorithm (Metropolis *et al* 1953, Hastings 1970). If the proposed model has a higher posterior value, it is always accepted and becomes the current model. If the proposed model has a lower posterior value, it is accepted with probability equal to the ratio of the posteriors. In practice this means if

9

the posterior ratio is greater than the random variate U, we accept the proposed model. If the priors are always the same, the acceptance condition is based on the ratio of the likelihoods (which is the original Metropolis algorithm). When the proposed model is rejected, the current model is retained for the next iteration and also added again to the samples collected.

The move of a cell nucleus can be efficiently implemented by adding and removing points from the existing Voronoi diagram without having to recalculate the entire structure. This can be done with the local Voronoi update algorithm described in Sambridge *et al* (1995b).

A burn-in period is needed to ensure that the chain has converged before samples start to be collected (Cowles and Carlin 1996, Brooks and Roberts 1998, 1999) presented practical tools to detect convergence in a running simulation). After many post burn-in iterations, this procedure asymptotically converges to an ensemble of models whose density is proportional to the posterior distribution (see Cowles and Carlin (1996), Brooks and Roberts (1998) for detailed proofs of convergence). Then, models can be collected every t steps of the walk. This thinning of the chain is to ensure the independence of the samples collected. The value for t is a parameter to be chosen according to the length scale ratio of the posterior to the proposal densities (MacKay 2003).

The great advantage of this method is that we do not need to know the normalizing constant in (5) to sample the posterior (as the constant cancels in the acceptance ratio). Depending on the nature of the priors and proposal distributions, the sampler will generally accept better data-fitting (i.e. higher likelihood) models. However, it can also accept what are considered 'worse' models (in that the data fit is lower than the previous model in the chain). This is required to obtain a distribution of samples that converges to the posterior and achieves a representative approximation of the probability distribution of the model parameters, i.e. we need to sample the tails as well as the modes.

Although the choice of proposal distributions is essentially arbitrary, poor choices lead to very slow movement around the model space, such that convergence of the chain can take a very long time. It is, therefore, desirable to choose proposal distributions carefully such that the model space search is as efficient as possible (e.g. Hopcroft *et al* (2007)). This is a central issue in the development of MCMC algorithms and the subject of much research (Brooks *et al* 2003, Al-Awadhi *et al* 2004, Raggi 2005). Monitoring the acceptance rate of the chain is useful to tune the variance of the proposal probability functions. Practically, it is advisable to choose the largest possible variance that maintains a high acceptance rate (Mosegaard 1998). Experience has also shown that a frequency of accepted models (after the burn-in period) of 25%–50% indicates that the algorithm performs well (Gelman *et al* 1996). 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.

The definition of the prior also has a direct impact on the efficiency of the inverse scheme. Overly precise prior information (e.g. a Gaussian distribution with a relatively narrow variance), if incorrect, can bias the inversion process such that the full range of possible solutions (i.e. models that fit the data) may not be properly sampled. Otherwise, if the prior is too loose or uninformative (e.g. a wide uniform distribution on the model space), the chain may sample many models far from the mode(s) of the distribution and never effectively converge.

3.8. The solution model and its error estimation

Once the posterior has been sampled from, the question that remains is how to interpret it, i.e. how do we extract an appropriate solution model? A very easy way would be to take the

10

model sampled that best fits the data, i.e. the model that provides a maximum posterior value. However, a Bayesian framework encourages one to think in terms of an ensemble solution, i.e. to look at the properties of an ensemble instead of a single model. Importance sampling provides the possibility of getting any statistical information about the posterior. Indeed, one can evaluate the posterior expectation of any function $f(\mathbf{m})$ of the model such as the mean or the covariance (Gilks *et al* 1996). The expected value is given by the expression

$$E[f(\mathbf{m})] = \int f(\mathbf{m}) p(\mathbf{m} \mid d_{obs}) \, d\mathbf{m}, \tag{15}$$

where $p(\mathbf{m} \mid \mathbf{d}_{obs})$ is the posterior distribution and $f(\mathbf{m})$ is some function of interest. The samples generated by the (post burn-in) Markov chain are distributed according to the posterior and can be used to calculate this expectation value (Gilks *et al* 1996),

$$E[f(\mathbf{m})] = \frac{1}{S} \sum_{s=1}^{S} f(\mathbf{m}_s), \tag{16}$$

where S is the number of models collected in the post burn-in period (one model collected every t steps of the chain).

In the case that $f(\mathbf{m})$ is the velocity model itself ($f(\mathbf{m}) = \mathbf{m}$), the expectation E is just the mean of the samples collected. This is equivalent to integrating across all models weighted by their posterior probability (as the collected models are sampled with frequency proportional to their posterior probability density).

In our problem, the posterior distribution is far from being a simple multidimensional Gaussian and has a large number of maxima. This can be seen from a qualitative consideration. Let us assume that we are at the maximum of the posterior. If we swap the location of two nuclei with the velocity assigned to them, the velocity model does not change at all but we are at a different point in the parameter space. This means the geometry of the posterior is highly redundant and the posterior expectation $E(\mathbf{m})$ will not correspond to a simple 'average' velocity field. To obtain a 'solution map' for analysis purpose, we project the partition models into the spatial domain (i.e. where the parameters represent the velocities in each pixel of the velocity field), and then simply average all the sampled images, that is, by taking the mean of the distribution of velocity values at each point across the 2D model (Denison *et al* 2002). Similarly we generate an error estimation map by calculating the standard deviation of velocities at each point.

All the models sampled have particular parametrization defined by their cell geometry. When a large number of models with different parametrizations are stacked, the Voronoi polyhedra overlap so the spatial average model is effectively a continuous function of the plane. As can be seen in the examples to follow, this continuous map preserves features common to a family of models and provides a higher spatial resolution than any single sampled model.

It is important to collect enough independent samples so that the solution maps are stationary and represent well the solution mean and variance. Integrating the information coming from different models with different parametrizations can be seen as a self-smoothing process that automatically removes unwarranted high frequency features of the models. In this way, the approach to be described here has an inherent smoothing character without the need to define an explicit external smoothing function. We will show that the method does not need a predefined damping procedure and can be viewed in a sense as an example of a self-regularizing inversion algorithm.

3.9. Computational cost

The major limitation of Monte Carlo methods is that the dimension of the model space, e.g. the number of unknowns, has to be restricted in size. If the model is defined by too many parameters, the number of samples needed to explore the whole model space becomes huge. This is known as the curse of dimensionality (Tarantola 2005). Each time a model is proposed, the predicted data have to be computed, e.g. the forward problem has to be solved. This can be expensive 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 a good resolution. This is the reason why Monte Carlo methods have not been routinely used in tomographic imaging. Nevertheless, our methodology enables the use of an MCMC method for a 2D imaging problem for three reasons.

First, we will show that the resolution obtained by taking an average of the ensemble of models with fixed dimensions but spatially varying parametrizations is higher than the resolution obtained with a higher dimensional single model having a fixed set of basis functions. This allows the number of unknowns to be smaller than for a standard tomographic inversion.

Second, our method has been optimized such that only a part of the forward problem needs to be recomputed each time a new model is tested. At each step of the chain, a new model is proposed by locally perturbing a model previously tested. This small local change in the model implies a change in only a few of the predicted measurements. For example, figure 3 shows the geometry of the problem before and after a Voronoi nuclei is moved. Only the cells in grey change during the move. Therefore, only the travel times of the rays crossing the grey cells need to be updated to compute the data misfit of the new model. The new ray lengths $\mathbf{L}_{ij}$ in the grey cells are found by sampling along the concerned rays at a predetermined step length and by finding the cell containing the midpoint of the current ray segment. This is done efficiently with the point location algorithm described in Sambridge and Gudmundsson (1998). This fast way of solving the forward problem enables one to rapidly explore the model space. It turns out to save a significant amount of time and allows the number of unknowns to be larger than for a standard MC approach.

The last reason is that the MCMC algorithm is easy to parallelize. In effect, the same algorithm can be run on many different machines so multiple chains can sample the model space independently of each other. Rosenthal (2000) showed that 'parallel Markov chain Monte Carlo' essentially gives linear speedup over a single-computer simulation (aside from needing to discard the first burn-in samples from each computer). Furthermore, the resulting estimator $E(\mathbf{m})$ has linearly-reduced variance, compared to the original single-processor estimate. See Rosenthal (2000) for details and for an argument to suggest that MCMC algorithms are ideally suited to parallel computing.

4. Synthetic data examples

4.1. Experimental set-up

To illustrate the new algorithm in the case of an uneven spatial distribution of information, we present results for a 2D problem where the ray coverage is irregular. This situation is common in seismic tomography where the sources (earthquakes) occur mostly at plate boundaries. The ray geometry is shown in figure 4. The upper left part of the model is covered by many ray paths, whereas the lower right part is left with relatively few.
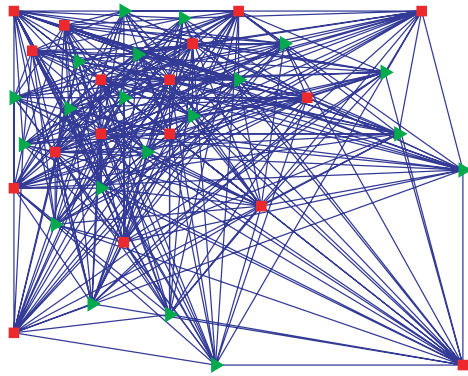
Figure 4. Geometry of rays. 340 rays join 17 sources (red squares) to 20 receivers (green triangles).
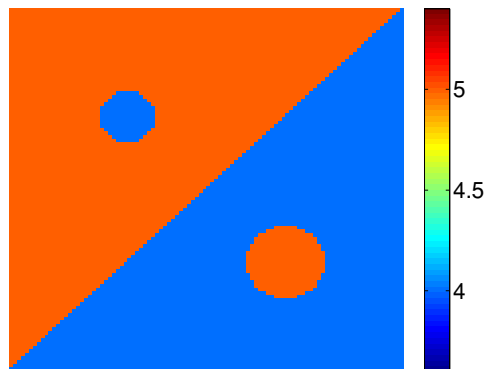

Figure 5. True velocity field (km s^{-1}).

One sees here the difficulty of choosing an appropriate cell size for a regular mesh. The problem may become overdetermined in the upper left part (large number of rays crossing each cell) and underdetermined in the lower right part (not enough rays crossing the cells). Note also that in the lower right quarter, all the rays are in similar directions indicating that resolution in this direction will be poor. In this area, a small error in the data will likely map into a large error in the velocity estimation of the cells, a property known as ill-posedness. Figure 5 shows the true velocity field that we try to recover. The areas in red have a velocity of 5 km s^{-1} and the velocity in the blue areas is of 4 km s^{-1}. The velocity field contains high contrast discontinuities. It will be shown that the dynamic parametrization can recover these interfaces and adapt to the geometry of the underlying seismic earth model. The small blue heterogeneity represents a velocity anomaly of -20% of the red background and the red heterogeneity is $+25\%$ of the blue background. Here we consider only a linear problem, where it is assumed that the ray paths are independent of the velocity structure. While this is only a first-order approximation for the seismic problem, it serves well to demonstrate the approach we advocate.

We compare our Monte Carlo sampler to some widely used regularization techniques for linear inversion. We first consider an ideal noise-free case and later add random noise to the

synthetic data. In this way, we investigate the potential of partition modelling to deal with a large variability in the range of possible solutions and to predict model uncertainty.

4.2. Noise-free experiment: Tikhonov regularization versus partition modelling

4.2.1. The regularization process in linear inversion. Most of the methods using a predefined fixed parametrization formulate the tomography problem with a linear system of algebraic equations represented by a matrix **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. A common choice of regularization is to use the norm of the first or the second derivative of the model (Aster *et al* 2005). The inversion scheme consists then in minimizing a linear combination of two criteria and takes the form

$$\min[\|\mathbf{Gm} - \mathbf{d}\|^2 + \alpha^2 \|\mathbf{Dm}\|^2], \tag{17}$$

where the first term is the data misfit and **Dm** is a finite difference approximation that is proportional to the first or the second derivative of **m**. By minimizing the semi-norm $\|\mathbf{Dm}\|^2$, the regularization techniques will favour models that are relatively flat (first-order regularization) or smooth (second-order regularization). The parameter α gives relative weight to the two terms and is called the regularization factor. Determining a solution through solving the problem (17) is known as a Tikhonov regularization. For any given value of α, the solution to (17) is unique and can easily be found with least-squares optimization techniques. The issue is then to find an appropriate α. If α is too large, the solution is too damped: the model is smooth but the fit to the data is poor. If α is too small, the fit to the data is good but there are strong instabilities due to the non-uniqueness of the solution. Different methods can be used to choose the regularization factor. Here we adopt the general cross validation (GCV) method (see Aster *et al* (2005) for details).

The GCV function obtained with a second-order Tikhonov regularization for a uniform grid of 400 cells is plotted in figure 6(a). The red dashed line shows the minimum of the GCV function which corresponds to an α of 2.6. The corresponding solution model after solving (17) with this value of α is shown in figure 6(b). In order to examine details and show the result, a (triangle-based) linear interpolation has been performed. The interpolated model is shown in figure 6(c).

The cells in the upper left part are well determined and give velocities close to the true value. The upper-left slow velocity anomaly is well imaged given the grid resolution. The lower-right fast velocity anomaly is poorly recovered, as we expect given the low ray path density in this region.

In an attempt to improve the method, we tried to manually adjust the regularization parameter α. We observed the classic trade-off between smooth models with poor spatial resolution and instability (Menke 1989) (see figure 7). We have also experimented with changing the number of cells in the grid. When the number of cells is decreased, the instabilities are removed but the resolution is not good enough to map the heterogeneities. Figures 6 and 7 represent the best results obtained from a range of experiments within a regularization framework.

4.2.2. Tomography with partition modelling. As mentioned above, our Monte Carlo sampler does not need explicit regularization to produce a single smooth velocity field. We select a representative 'solution' as the pointwise spatial average of an ensemble of models. Note that
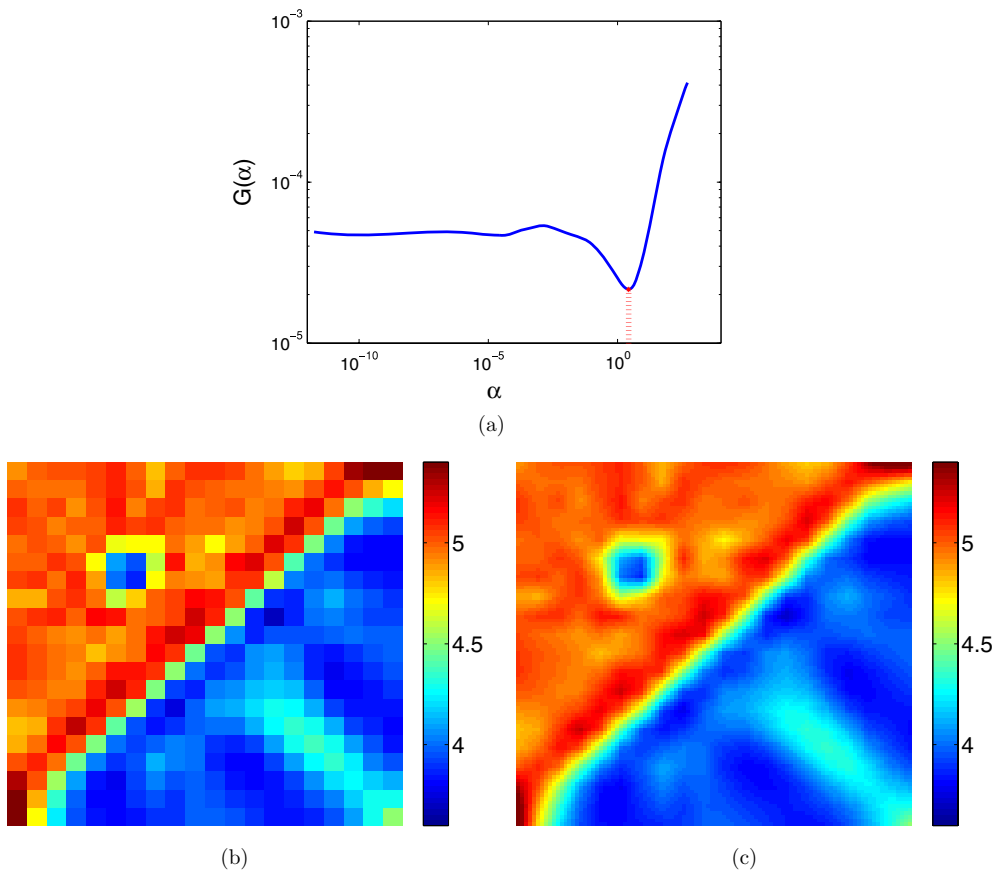
14

Figure 6. Noise-free experiment. Linear inversion after linear interpolation over a regular grid: second-order Tikhonov regularization with general cross validation. (a) GCV function, minimum at $\alpha = 2.6$. (b) Results for a grid of $20 \times 20 = 400$ cells (km s^{-1})). (c) Solution Model after Linear Interpolation (km s^{-1}). The colour scales are the same as for the true model.

each individual model consists of a different configuration of a finite number of Voronoi cells (see figure 8(a)), but the average solution taken pointwise is smooth without need to impose an explicit interpolation procedure. The MCMC sampling has been carried out with 45 Voronoi cells. It took 300 s on a standard desktop computer (Intel core 2 duo with CPU running at 2.1 GHz) to collect and store 50 000 samples (2000 of which are burn-in samples). The best model obtained in terms of data misfit (which is an estimate of the maximum posterior model) is shown in figure 8(a). The spatial average of the post burn-in samples collected (i.e. the solution map) is shown in figure 8(a). The scales are the same as in the previous figures.

Clearly, the average model is closer to the true solution than the models obtained with a fixed grid. The instabilities are not present and the heterogeneities have been recovered with improved accuracy. Note that this inversion only uses 45 Voronoi cells, whereas the Tikhonov inversion scheme uses 400 fixed cells. Therefore, the Monte Carlo sampler achieves a better resolution with fewer cells which results, as expected, from averaging many overlapping Voronoi cells in different configurations. The best solution in the regularization framework (figure 7(c)) is obtained with $\alpha = 10$, and this represents a compromise across the entire
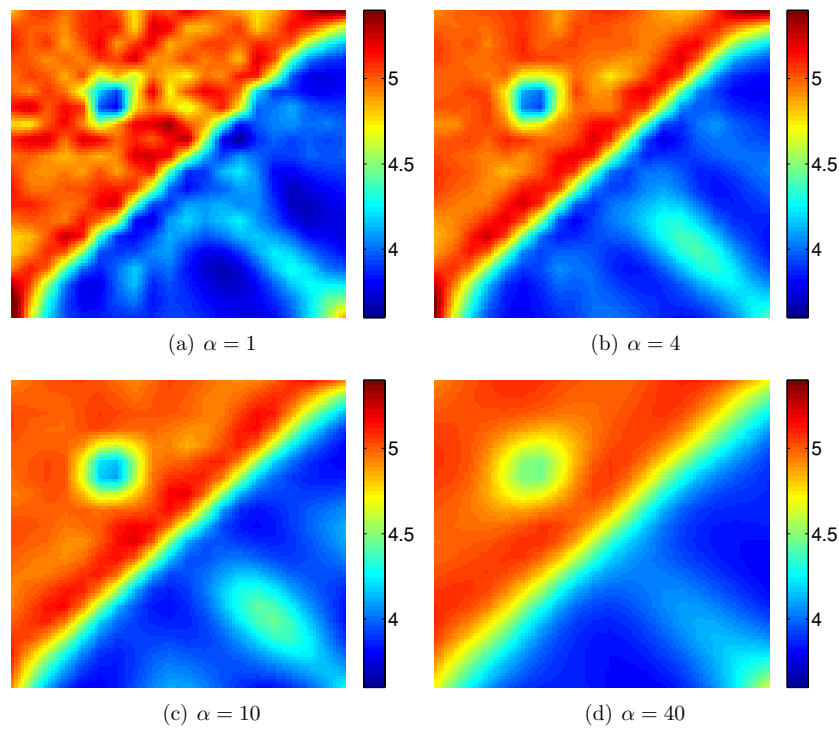
(a) $\alpha = 1$

(b) $\alpha = 4$

(c) $\alpha = 10$

(d) $\alpha = 40$

Figure 7. Linear inversion with a regular grid (20×20 cells). Noise-free experiment. Results for different values of the regularization parameter (km s^{-1}).
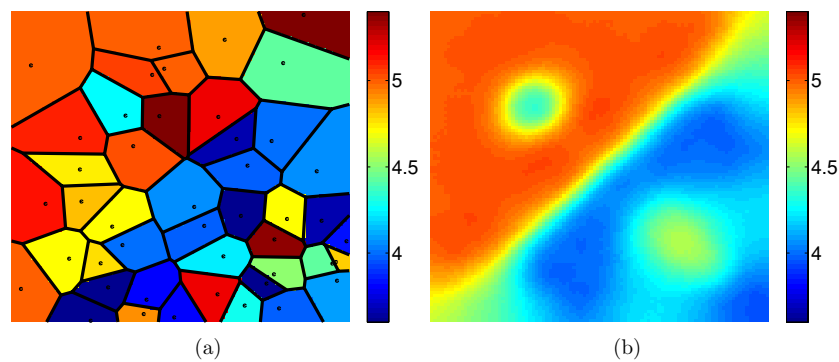

(a)

(b)

Figure 8. Partition modelling with 45 mobile Voronoi cells. Results for the noise-free experiment. (a) Best model sampled: posterior maximum (km s^{-1}). (b) Average solution map (km s^{-1}).

model. In the partition modelling approach there is no global damping parameter, but instead the algorithm has smoothed the model locally in response to the data.

The averaging process removes unwarranted discontinuities in individual models due to the parametrization but constructively reinforces the well constrained ones. That is, real discontinuities in the velocity field are preserved because the dynamic parametrization has been

able to adapt to the structural features of the unknown model (i.e. many models approximate the significant features well).

In the work presented here, the model parametrization is adaptive but its dimension (i.e. the number of cells) is fixed and a parameter to be chosen by the user. In the absence of a method that produces Bayesian information about the dimension, the number of cells has to be chosen from criteria based on visual inspection. It is possible to monitor how the solution map changes with the model dimension. In experiments with different number of cells, we observed that, for small number of cells, only large anomalies are mapped and velocity amplitudes and gradients are low. As the model space dimension increases, small anomalies start to appear, velocity amplitudes and gradients increase until the solution map becomes stable. That is, the solution does not change anymore with the dimension. In the examples presented here, the dimension was chosen in order to achieve a stable solution with the minimum number of cells.

4.3. Noise propagation and model uncertainty

The error on the model depends on data errors. In order to test the ability of the sampler to predict model uncertainty, random Gaussian noise has been added to the data. Our 2D velocity field is defined on a square of 100 km by 100 km. For our true model and geometry of rays, the average observed travel time is 10 s. For a homogeneous initial model with velocities equal to 4.5 km s^{-1}, if no noise is added to the observed data, the average difference between observed travel times and estimated travel times will be around 0.4 s. The standard deviation of the added noise is 1.2 s (i.e. 12% of the average observed travel time), which is quite large in practice.

4.3.1. Linear inversion with noise. As we have seen, the regular grid and the uneven distribution of rays make the problem ill-posed. A small amount of noise in the data will imply a large variability in the solution. When noise is added, it becomes more difficult to find a satisfactory regularization scheme. As in the noise-free case, we tried to use the first- and second-order Tikhonov regularizations with a damping factor given by either the GCV or the L-curve method (Aster *et al* 2005). In both cases, a satisfactory solution was not obtained. The solution is either too damped and the anomalies are not resolved or there are unrealistic instabilities in the solution. The damping factor had to be tuned manually, and figure 9 shows the results obtained with different values of α for a semi-norm defined by the second derivative of the model. Note that the values of α here are higher than those in the noise-free experiment. Figure 9(d) shows two cross sections of the solution map for $\alpha = 40$: one at $y = 70$ km and a second at $y = 30$ km. The black line is the amplitude of the true velocity model, the red line is the solution model and the dashed green lines are the solution model plus and minus one standard deviation. The standard deviation of the estimated model is given by the diagonal elements of the posterior covariance matrix which are interpolated over the 2D field (Aster *et al* 2005). It is well known that in a regularization or damping approach, error estimation procedures often lead to unrealistically small errors in the model (i.e., if noise is added to data, the variability estimated in the model can be much less than the true errors). This is because regularization stabilizes the model construction process at the cost of biasing the solution in a statistical sense (Aster *et al* 2005). This effect can be seen in figure 9(d). The errors on the solution model are clearly underestimated.

4.3.2. Partition modelling with noise. In contrast to regularization procedures, MCMC allows us to perform ensemble inference, that is, to capture the variability in the range of possible solutions consistent with the data. The standard deviation of the family of models
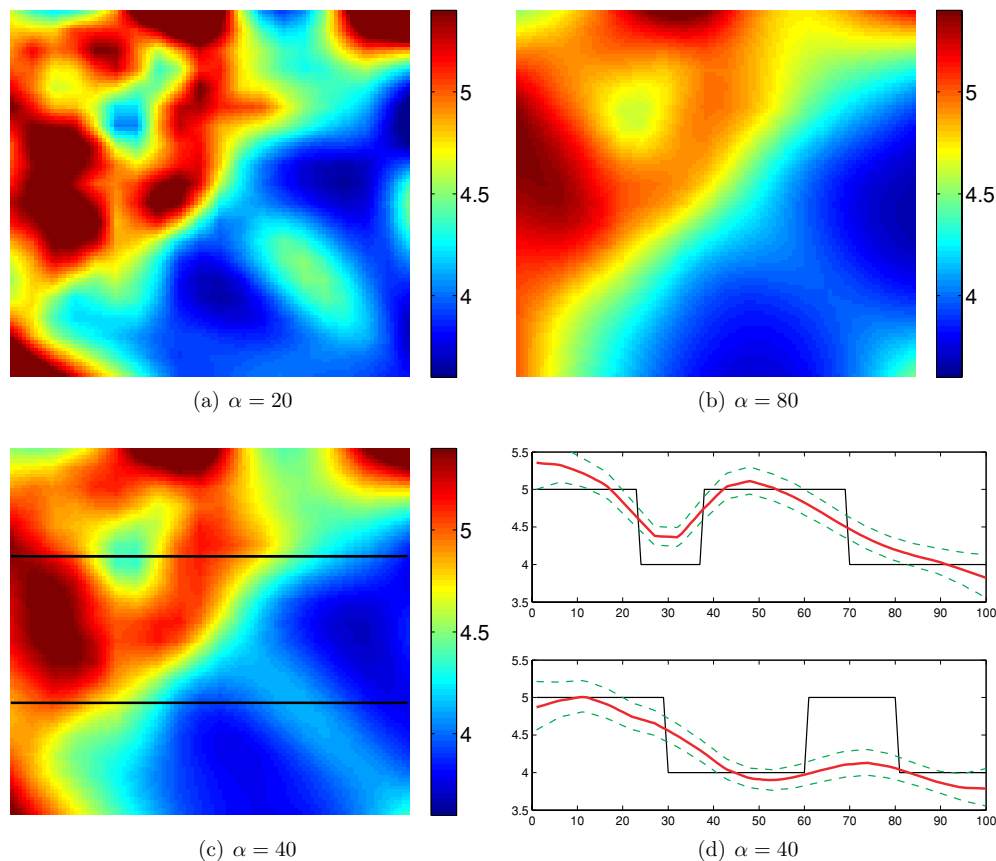
(a) $\alpha = 20$

(b) $\alpha = 80$

(c) $\alpha = 40$

(d) $\alpha = 40$

Figure 9. Results with 12% random Gaussian noise. Linear inversion (after interpolation) with a regular grid of 400 cells (km s^{-1}). (d) Cross section for $y = 30$ km and $y = 70$ km (km s^{-1}). Black: true model; red: solution model. The dashed green lines are the average model plus and minus one value of the interpolated standard deviation.

provides a map of the posterior standard deviation which can be taken as a measure of the error for the velocity model.

The results obtained are shown in figure 10. The map in figure 10(a) shows the average solution map with the same scale as in previous figures. It is clear that the heterogeneities have been recovered here with improved accuracy than in the linear inversion results. The map in figure 10(c) shows the estimated error map. The map in figure 10(d) shows the true error for the average solution (i.e., the absolute value of the difference between the true model (figure 5) and the average solution (figure 10(a))). Figure 10(b) shows the cross sections of the map as in figure 9.

The model uncertainty map obtained in this way (figure 10(c)) appears to be strikingly similar to the actual error (figure 10(d)). Note that these two maps are at the same scale. The true error map has lower amplitudes than the variance map. Indeed, one sees in figure 10(b) that the error for the solution model is actually smaller than the estimated error. That is, the true error is virtually always within the interval defined by plus and minus one standard deviation.
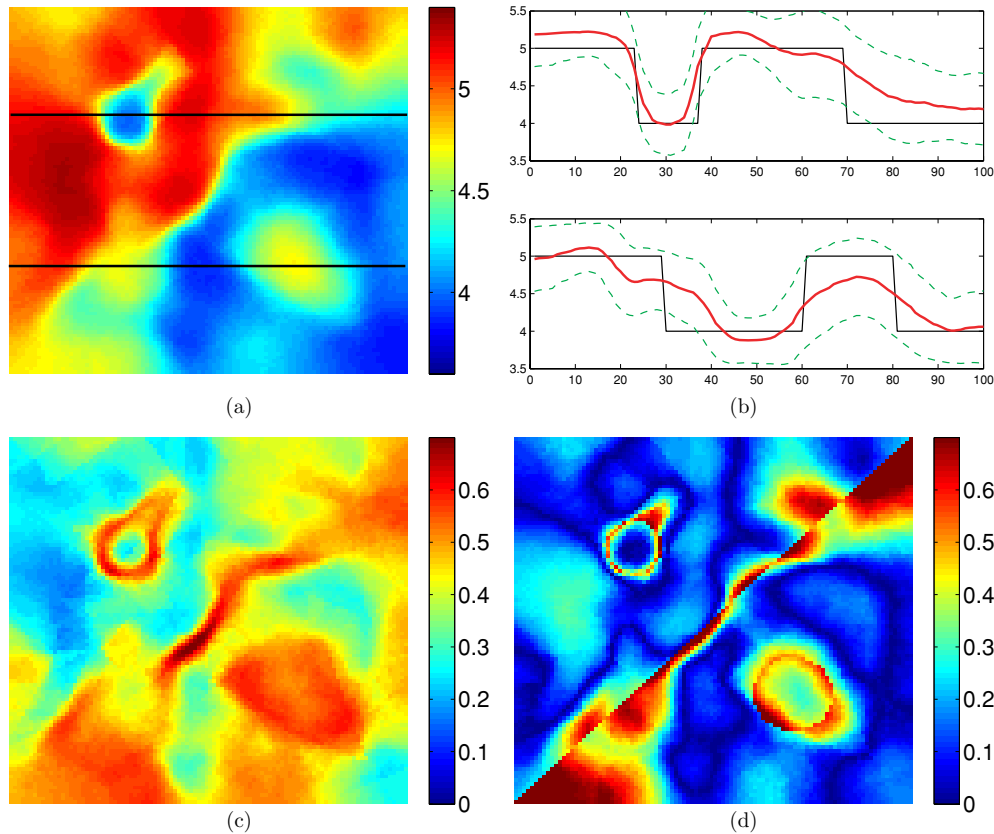
Figure 10. Results with 12% random Gaussian Noise. Partition modelling with 45 mobile Voronoi cells. (a) Average solution map (km s^{-1}). (b) Cross section for $y = 30$ km and $y = 70$ km (km s^{-1}). The black line is the amplitude of the true model and the red line is the estimated average model. The dashed green lines are the average model plus and minus one value of the local standard deviation. (c) Estimated error map (km s^{-1}). (d) True error. Absolute difference between the true and the estimated model (km s^{-1}).

In these experiments, the Monte Carlo sampler provides a reliable estimation of the model uncertainty both in terms of amplitudes and lateral variations. This is a considerable advantage over the linear methods using a fixed grid.

It is important to note that in above synthetic experiments, the variance of the data noise σ_d^2 is known, whereas, in practice, it has to be estimated when measuring the data. If the data noise is overestimated, the likelihood function in (7) becomes broad and the Markov chain will explore more extensively the model space, including models with a bad fit, and resulting in a solution map with low amplitudes and smooth gradients. If the data noise is underestimated, the Markov chain will tend to 'over-fit' the data by keeping in regions of the model space providing a very good data fit, and resulting in a solution map with unnecessary strong amplitudes and gradients.

Since the cells are able to adapt their position, size and shape, one might expect that the size of the cells would adapt to the density of rays, i.e., small cells would gather in the areas of maximum information, leaving the zones with less ray density covered with large cells.

Surprisingly, this is not the case. The Markov sampler does not tend to change the size of cells significantly, but instead it automatically adapts the acceptance rate during the Monte Carlo Markov chain. The acceptance rate is higher for cells located in low ray density areas. In other words, the proposed moves (both in velocity and nuclei location) are more often accepted when the cells are in a poor ray density area, i.e. in regions where a lot of different models fit equally the data. This can easily be seen on the map of the estimated error in figure 10(c). The upper left part shows a lower variance than the lower right part.

The size of the error estimation interval obtained from the Monte Carlo method (i.e. the posterior standard deviation) increases in areas where the problem is ill-posed. In a situation where a lot of models almost equally satisfy the data, regularization chooses one model according to particular criteria and discards the others. In contrast, Bayesian inference takes into account the variability in the models and therefore looks to have a considerable advantage over choosing a single 'best' model obtained by optimization methods.

5. Conclusion and future work

The methodology we have presented here is a general tomographic inversion strategy. The problem is formulated in a probabilistic framework and is able to both exploit and infer the spatial variability of the information provided by data. We have utilized partition modelling which allows efficient exploration of the model space by sampling models of spatially varying parametrizations. Although the numerical experiments reported here are in relatively simple test problems, some notable features emerge.

First, the advantages of this approach are that the inverse problem can be treated with a fully nonlinear parameter search method and that explicit regularization of the model parameters is not required, thus avoiding global damping procedures and the often subjective process of finding an optimal regularization value.

Second, the MCMC approach lets us consider the issue of discretizing the 2D model. The velocity field has been parametrized by Voronoi polyhedra with mobile geometry throughout the inversion process. A posterior probability distribution for the velocity field has been defined. When the posterior expectations (mean and variance) are computed, models with different cell geometry overlap providing a smooth solution map that has a resolution better than any single model. This dynamic parametrization enables automatic self smoothing and thus avoids the need to impose a global level of spatial smoothing, e.g., through a damping procedure. Furthermore, the model parametrization involves fewer parameters to achieve a better resolution than a fixed grid.

Another advantage is that the spatial average over all models in the ensemble seems to better capture the variability in the range of possible solutions than a single (e.g. best data fitting) model (e.g. Hopcroft *et al* (2007)). The discontinuities of individual models are smoothed out in the ensemble solution, but the discontinuities required by the data are constructively reinforced. We view the construction of a continuous smooth map giving an accurate estimation of the velocity uncertainty as a novel result from this study.

Moreover, the optimization of the MCMC sampler has proved to be efficient in terms of computational costs, and our preliminary results are encouraging for applications to larger data sets. All the routines presented here for the 2D case already exists in 3D, and hence the algorithm can easily be implemented in 3D, although it is difficult at present to asses the extra computational cost. The method has been tested for straight ray linear tomography, but if used iteratively, it can be extended to nonlinear tomography with bent rays, again at increased computational cost.

The cell shapes and sizes themselves do not adapt to structure but rather the acceptance rate of the Metropolis simulation does. The adaptive approach to parametrization used here is highly novel in geophysical applications and would appear to have considerable potential for similar problems elsewhere. Note that a similar MCMC algorithm in a Bayesain framework could easily be implemented for the fixed regular cell case, producing smooth results with good spatial resolution. However, in order to obtain the same resolution as with Voronoi cells, a much larger number of cells would have to be used, making the inversion computationally expensive and potentially impractical.

One can argue that with MCMC used in a Bayesian framework, the solution is influenced by subjective choices, such as the form of the prior distributions or the proposal probability densities. In the examples shown here, the final models are clearly dominated by the data rather than by prior information and so we do not consider this to be a major criticism.

Future work will focus on allowing the number of cells to be an unknown, i.e., allowing the sampler to explore models of different dimensions. This will result in a trans-dimensional problem, which can be addressed using a birth–death MCMC (Geyer and Møller 1994, Stephens 2000a, 2000b) or a Reversible Jump Markov chain Monte Carlo algorithm (Green 1995). We anticipate that a trans-dimensional approach will be more efficient than the fixed dimension method presented here, and this is the subject of future work.

Acknowledgments

References

Abers G and Roecker S 1991 *J. Geophys. Res.* **96** 6379–401
Aki K, Christoffersson A and Husebye E 1977 *J. Geophys. Res.* **82** 277–96
Al-Awadhi F, Hurn M and Jennison C 2004 *Stat. Probab. Lett.* **69** 189–98
Aster R, Borchers B and Thurber C 2005 *Parameter Estimation and Inverse Problems* (New York: Academic)
Backus G and Gilbert F 1968 *Geophys. J. Astrophys. Soc.* **16** 169–205
Bayes T 1763 An essay towards solving a problem in the doctrine of chances *Philos. Trans. R Soc. London* **53** 370–418
Billings S, Sambridge M and Kennett B 1994 *Bull. Seismol. Soc. Am.* **84** 1978–90
Box G and Tiao G 1973 *Bayesian Inference in Statistical Inference*
Brooks S, Giudici P and Roberts G 2003 *J. R. Stat. Soc.* B **65** 3–39
Brooks S P and Roberts G O 1998 *Stat. Comput.* **8** 319–35
Brooks S and Roberts G 1999 *Biometrika* **86** 710–7
Budinger T F, Gullberg G T and Huesman R H 1979 *Image Reconstruction from Projection: Implementation and Applications* (Berlin: Springer) p 147
Cerveny V and Brown M 2003 *J. Acoust. Soc. Am.* **113** 14
Cerveny V, Molotkov I and Psencik I 1977 (Prague: Charles University)
Cowles M and Carlin B 1996 *J. Am. Stat. Assoc.* **91** 883–904
Curtis A and Snieder R 1997 *Geophysics* **62** 1524–32
Denison D and Holmes C 2001 *Biometrics* **57** 143–9
Denison D, Holmes C, Mallik B and Smith A 2002 *Bayesian Nonlinear Methods for Classification and Regression* (Chichester: Wiley)
Diehl T, Kissling E, Husen S and Aldersons F 2009 Simulation procedures and liklihood inference for spatial point processes *Geophys. J. Int.* **176** 542–54
Duijndam A 1988 *Geophys. Prospect.* **36** 878–98
Friederich W 1998 *Geophys. J. Int.* **132** 203–25
Fukao Y, Obayashi M, Inoue H and Nembai M 1992 *J. Geophys. Res.* **97** 4809–22

Gelman A, Carlin J, Stern H and Rubin D 2004 Bayesian Data Analysis. Text In Statistical Science (Boca Raton, FL: Chapman and Hall) vol 25 p 668 (ISBN 0 412(03991), 5)

Gelman A, Roberts G and Gilks W 1996 *Bayesian Stat.* **5** 599–607

Geyer C J and Møller J 1994 Inference of past climate from borehole temperature data using Bayesian reversible Jump Markov chain Monte Carlo *Scand. J. Stat.* **21** 359–73

Gilks W, Richardson S and Spiegelhalter D 1996 *Markov Chain Monte Carlo in Practice* (London: Chapman and Hall)

Gouveia W and Scales J 1998 *J. Geophys. Res.* **103** 2759–80

Graeber F, Houseman G and Greenhalgh S 2002 *Geophys. J. Int.* **149** 249–66

Green P 1995 *Biometrika* **82** 711–32

Hastings W 1970 *Biometrika* **57** 97–109

Hopcroft P O, Gallagher K and Pain C C 2007 *Geophys. J. Int.* **171** 1430–9

Hoppe W and Hegerl R 1980 *Computer Processing of Electron Microscope Images* (Berlin: Springer) pp 127–85

Ivansson S 1986 *Proc. IEEE* **74** 328–38

MacKay D J C 2003 Information theory, inference and learning algorithms

Menke W 1989 *Geophysical Data Analysis: Discrete Inverse Theory* (New York: Academic)

Metropolis N *et al* 1953 *J. Chem. Phys.* **21** 1087–91

Mosegaard K 1998 *Inverse Problems* **14** 405–26

Mosegaard K and Tarantola A 1995 *J. Geophys. Res.* **100** 12–431

Natterer F 2001 *The Mathematics of Computerized Tomography (Society for Industrial Mathematics)*

Neal R 1993 *Probabilistic Inference Using Markov Chain Monte Carlo Methods* (Department of Computer Science, University of Toronto)

Nolet G and Montelli R 2005 *Geophys. J. Int.* **161** 365–72

Nolet G and Panza G 1976 *Pure Appl. Geophys.* **114** 775–90

Okabe A, Boots B and Sugihara K 1992 *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams* (New York: Wiley)

Prindle K and Tanimoto T 2006 *Geophys. J. Int.* **166** 601–21

Raggi D 2005 *Econometrics J.* **8** 235–50

Rawlinson N and Sambridge M 2003 Seismic braveltime tomography of the crust and lithosphere *Advances in Geophysics* **46** 81–199

Rosenthal J 2000 *Far East J. Theor. Stat.* **4** 207–36

Sambridge M, Braun J and Mcqueen H 1995a *Geophys. J. Int.* **122** 837–57

Sambridge M, Braun J and McQueen H 1995b *Geophys. J. Int.* **122** 837–57

Sambridge M and Faletic R 2003 *Geochem. Geophys. Geosyst.* **4** 1022

Sambridge M, Gallagher K, Jackson A and Rickwood P 2006 *Geophys. J. Int.* **167** 528–42

Sambridge M and Gudmundsson O 1998 *J. Geophys. Res.* **103** 773–82

Sambridge M and Rawlinson N 2005 *Geophys. Monogr.* **157** 49–65

Scales J and Snieder R 1997 *Geophysics* **62** 1045–6

Sivia D 1996 *Data Analysis: A Bayesian Tutorial* (Oxford: Oxford University Press)

Smith A 1991 *Phil. Trans.: Phys. Sci. Eng.* **337** 369–86

Spakman W and Bijwaard H 1998 *Ann. Geophys.* **16** 18

Stephens M 2000a *Ann. Stat.* **28** 40–74

Stephens M 2000b *J. R. Stat. Soc.* B **62** 795–809

Stephenson J, Gallagher K and Holmes C 2004 *Geol. Soc. London Special Publications* **239** 195

Stephenson J, Gallagher K and Holmes C 2006 *Earth Planet. Sci. Lett.* **241** 557–70

Tarantola A 2005 *Inverse Problem Theory and Methods for Model Parameter Estimation* (Philadelphia, PA: SIAM)

Tarantola A and Valette B 1982 *J. Geophys.* **50** 150–70

Tierney L 1994 *Ann. Stat.* **22** 1701–28

Voronoi G 1908 *J. Reine Angew. Math.* **134** 198–287

Zhang H and Thurber C 2005 *J. Geophys. Res.* **110**