Transdimensional Approaches to Geophysical Inverse Problems

Thomas Bodin

October 2010

A thesis submitted for the degree of Doctor of Philosophy of The Australian National University

Except where otherwise indicated in the text, the research described in this thesis is my own original work, figurative use of the first person plural notwithstanding. The research in chapters 2 and 3 has appeared in publication in *Inverse Problems* and *Geophysical Journal International*. These articles were co-authored with Malcolm Sambridge and Kerry Gallagher. In both case I was the lead author and took prime responsability for the research. The research in chapter 4 has been combined to other work and submitted for publication to *Geochimica Cosmochimica Acta*. In this case, I am second author in the publication after Kerry Gallagher. Further details about publication are provided in the publication schedule at the end of chapter 1.

Thomas Bodin
October 15, 2010

Acknowledgments

I would like thank my supervisor Malcolm Sambridge for his advise and encouragements during this Ph.D, his taste for scientific intuition and academic rigour have been very formative, and his enthusiasm and sympathy are always enjoyable and motivating.

I am also indebted to my advisors, Hrvoje Tkalcic, Nick Rawlinson and Brian Kennett for their efforts to answer my questions on various aspects of seismology. I will always remember the delicious meals prepared by Hrvoje and Nick.

I also wish to thank Erdinc Saygin for providing ambient noise data from Australia, and Kerry Gallagher for advice and discussions on this study, and for his warm hospitality in Rennes.

Discussions with Pierre Arroucau have been extremely stimulating, and helped me a lot, notably to formulate complicated and quantitative mechanisms with simple qualitative concepts.

This research was supported under Australian Reasearch Council's Discovery Projects fundings scheme (project number DP0665111). 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.

This PhD project marks the end of my education, and I would like to thank my parents Loys and Marian, as well as my brother Pablo and my sister Agnes, for their unconditional encouragements and support during all these years. Being so far from family hasn't always been easy.

Finally, I would like to thank my beloved wife Shinta for her love, her affection, her warmth, for her support every day, for sharing with me the best and worse moments, and Jaures for all the smiles in the eyes.

Abstract

In geophysical inversion the model parameterisation, the number of unknown parameters, the level of smoothing and the required level of data fit are usually arbitrarily determined by the user prior to the inversion. These quantities are related to each other and define the formulation of the inverse problem; by definition they affect the final solution. They are often manually 'tuned' by means of subjective trial-and-error procedures, and this represents a recurring problem in geophysical inversion.

In this thesis this issue is addressed by proposing an alternative inversion strategy. Different methodologies recently developed in the area of Bayesian statistics are combined to produce a general inversion algorithm, which lets the data themself formulate the inverse problem. This is done by treating the tunable quantities as unknowns to be constrained directly by the data.

A major focus is on situations where data constrain a 2D spatially varying field, particularly seismic tomography. A variable parametrisation consisting of Voronoi cells with mobile geometry, shape and number, is treated as a set unknowns in the inversion. The reversible jump algorithm is used to sample the transdimensional model space within a Bayesian framework which avoids global damping procedures and the need to tune regularisation parameters.

The method developed in this thesis is an ensemble inference approach, where many potential solutions are generated with variable numbers of cells. Information is extracted from the ensemble as a whole by performing Monte Carlo integration to obtain an 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. As a by product, uncertainty estimates are obtained for any point in the medium.

In a transdimensional approach, the level of data uncertainty directly determines the model complexity needed to satisfy the data. Intriguingly, the Bayesian formulation can be extended to the case where data uncertainty is also uncertain. It is possible to estimate the level of data noise while at the same time controlling model complexity in an automated fashion.

For 2D seismic tomography, this novel procedure gives promising results in situations where the ray coverage is far from ideal, as it performs better compared to standard methods that use regular parameterisations. The method is also applied to the inversion of three ambient noise datasets that span the Australian continent at different scales. A multiscale tomographic image of Rayleigh wave group velocity

for the Australian continent is constructed. Experiments show that the procedure is particularly powerful when dealing with multiple data types that have different unknown levels of noise. Here it is possible to adjust the fit to different datasets and to provide a velocity map with a spatial resolution adapted to the quantity of information present in the data.

Finally, two applications of 1D problems are considered. The first is an application to a regression problem where the goal is to infer the position and number of abrupt changes in noisy geochemical records. The second is an application to receiver function waveform inversion, where both the magnitude and correlation of data noise are inverted for. Other fields where the general methodology can be applied are outlined.

Contents

Chapter 1

Introduction

1.1 Geophysical inversion and the issue of model parameterisation

A typical aim of geophysical inversion is to recover some internal physical properties of the Earth (e.g. density, composition, temperature) from a set of surface measurements. The Earth model one tries to infer is often a continuous function of space: $M(\mathbf{x})$. In practice, this model is usually approximated as $m(\mathbf{x})$, which is a linear combination of a finite number, N, of basis functions B_i:

$$M(\mathbf{x}) \approx m(\mathbf{x}) = \sum_{i=1}^{N} a_i B_i(\mathbf{x}) \tag{1.1}$$

The definition of both N and the basis functions is called the model parameterisation. It is nearly always defined before carrying out the inversion. Thus the Earth model we try to infer from the observed data is simply defined by the vector of unknown model parameters $\mathbf{m} = (a_1, ..., a_N)$ which can be thought of as weights on each basis function.

Obviously, the choice of the parameterisation 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 $\mathbf{d}$:

$$\mathbf{d} = g(\mathbf{m}) \tag{1.2}$$

Here we can see that, in mathematical terms, the data can be regarded as a projection (through the function $g(..)$) of the model. Moreover, linearity, resolution, ill-posedness, and model uncertainty are concepts directly influenced by the param-

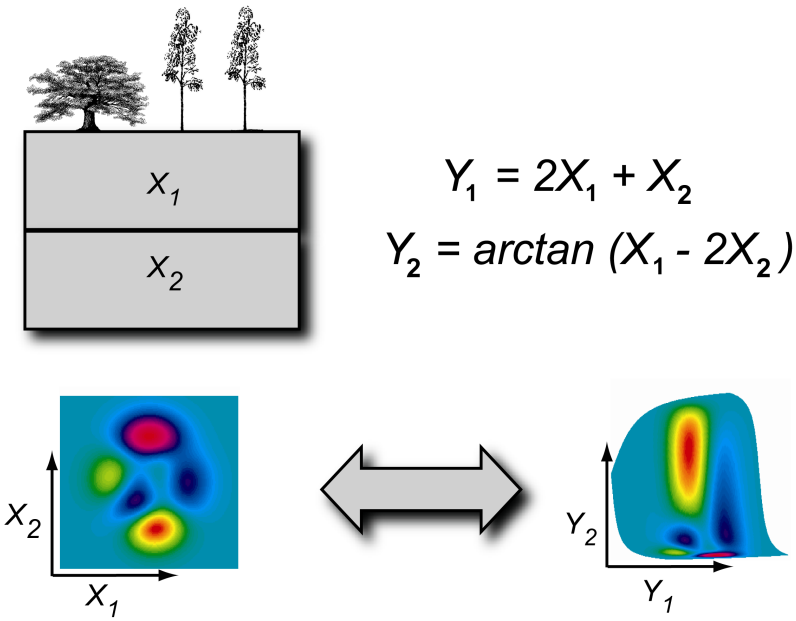


Figure 1.1: The 2D function represented here shows the data misfit. Purple areas have high misfit and orange areas have low misfit. X_1 and X_2 are physical parameters but the model in the transformed space (Y_1, Y_2) might be easier to estimate.

eterisation. For example, Hartzell and Langer (1993) demonstrated the dramatic effects of model parameterisation in the case of finite fault inversion where teleseismic long period waveforms are inverted to obtain rupture histories. In practice, the choice of the model parameterisation constitutes one form of prior information about a function $M(\mathbf{x})$ which we often do not know well. A good parameterisation enables one to extract the maximum information from the data, without introducing unjustified complexity to the model structure.

A simple example of an Earth model defined by two horizontal layers is shown in Figure 1.1. The two parameters X_1 and X_2 represent physical properties (e.g. temperature or electrical conductivity) that take a constant value within each layer. For example, the surface measurements can be gravity data and the inverse problem consists in estimating the two unknown density values X_1 and X_2. The data misfit function $\|\mathbf{d}_{obs} - g(\mathbf{m})\|$, i.e. the distance between observations and data predicted by the model, is also shown as a function of the two model parameters. The aim of the inversion is to approximate the data misfit function and eventually find the best combination of parameters that will minimise this 2D function. Here, it is obvious that a simple bijective transformation of parameters $(X_1, X_2) \longleftrightarrow (Y_1, Y_2)$ can dramatically change the "shape" of the function. In the transformed space, the

region where the data misfit takes low values is larger which may help the inversion process. The idea of parameter transformation is to change the geometry of the problem in order to make it easier to solve.

In linear problems, i.e. where Equation (1.2) can be expressed as a linear system of equations, a very common technique that uses parameter transformations is SVD decomposition (Aster *et al.*, 2005). Parameter transformations are used in order to have the model and data vectors expressed in a coordinate system such that each transformed datum is only sensitive to one of the transformed model parameter. This approach enables us to ignore the transformed data that are almost not sensitive to the model (or changes in the model) and which can destabilise the inverse problem. In SVD decomposition, the numerical stability can be improved by truncating the SVD representation by removing the terms equating to small eigenvalues. The threshold for truncation is specified by the user. If too many transformed parameters (i.e. combinations of physical parameters) are estimated, the problem will still be numerically unstable; if too few parameters are estimated, the model fit may be unnecessarily poor and data misfit may be larger than an optimally parameterised model. This is the well known trade off between model resolution and data fit.

Previous authors have recognised the potential of using adaptive parameterisation in inverse problems. Dynamic parameterisation methods enable the parameterisation to be adjusted during the model building procedure. Nolte and Fraser (1994) constructed a Genetic Algorithm where the problem was reparameterised after each run, so that the new unknown parameters were more independent than the old ones. The inversion technique was applied to a vertical seismic profile (VSP) inversion problem where the goal is to recover slowness and impedance profiles. We can also simplify or even linearise a problem by changing its parameterisation. Vasco (1995) used a set of coordinate transformations to linearise a non-linear inverse problem. Statistical and analytical techniques were employed to estimate the parameters of such linearisation transformations. In the transformed space, techniques for linear inverse theory were used.

In this thesis, we develop a general self-parameterising inversion method for the tomographic problem. The overall philosophy is to let the data decide of the parameterisation and thus avoid arbitrary choices that have to be made before the inversion and which influence the form of the solution. This work is carried out in the context of a probabilistic sampling framework where the solution is a large ensemble of Earth models that fully quantify the degree of knowledge we have about seismic structure (i.e. constraints, resolution, a trade-offs). Many interesting and useful

features emerge when a sampling based framework is applied to seismic tomography. These include transdimensionality (letting the number of degrees of freedom in the model vary), data driven regularization (letting the data determine the complexity of the model during the inversion) and super-resolution (where fine scale detail in an image is achieved by combining an ensemble of more crudely parameterised models). In addition the trade-off between data fit and model complexity is handled in a consistent manner driven by the data itself rather than by potentially subjective decisions. We also show that the methodology is general and can be applied to other geophysical inverse problems such as non-linear regression of geochemical time series or receiver function inversions.

1.2 Tomographic imaging

The word *tomography* literally means slice picture (from the Greek word *tomos* meaning *slice*). Tomographic imaging deals with reconstructing an image from its projections where the data 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 science. 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.

A common feature of many tomography problems is that the number and the distribution of projections is 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 spatial 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 that is sought. An example is

proton emission computed tomography in nuclear medicine (Budinger *et al.*, 1979).

1.3 Seismic tomography

Tomography has been an active domain of research in seismology for 30 years (for a recent review, see Rawlinson *et al.*, 2009) and consists of using data from seismograms to infer a map of seismic wave velocity. 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.

In most cases Earth models are parameterised with basis functions of uniform local cells in 2D or 3D 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 (Nolet, 2008). 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. Some form of model smoothing or regularisation is often imposed to obtain a single model, which biases linearised uncertainty estimates.

As mentioned earlier, the information obtained in seismic tomography strongly depends on the location of the sources (mostly at plate boundaries) and the positions of the receivers, which are not evenly distributed on the Earth surface. This leads to having some regions 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. The usual way of dealing with ill-constrained parts of a model is to apply some spatial smoothing, norm damping, or simply to coarsen the parameterisation, e.g., increase cell sizes. Traditionally these forms of 'regularisation' 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 over-smoothed and hence information may be lost.

In order to deal with the irregular distribution of information, some seismologists have used irregular meshes and allowed for their refinement during the inversion process (See Sambridge and Rawlinson, 2005, for a 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. Michelini (1995) adjusted the velocity and position of cubic B-spline vertices in 2D cross-hole tomography. Â Fukao *et al.* (1992) and Spakman and Bijwaard (1998) used non-uniform sized rectangular 3D blocks to account for uneven ray path sampling in regional models. Curtis and Snieder (1997) set up the inversion mesh adaptively to minimise the condition number of the resulting tomographic system of equations for the cross borehole tomographic problem by means of a genetic algorithm. Bijwaard *et al.* (1998), Bijwaard and Spakman (2000) and Spakman and Bijwaard (2001) performed global P-wave traveltime tomography using an approach similar to Abers and Roecker (1991) in which the 3D mesh is matched to the ray path density prior to inversion.

1.4 Voronoi cells

In this work, we propose to parameterise the velocity field for seismic tomography by a variable number of non-overlapping regions defined by Voronoi cells (see Figure 1.2) (Voronoi, 1908; Okabe *et al.*, 1992). These are completely unstructured meshes, e.g. not based on a cubic or other regular grid, and have the advantage of adapting to the spatial variability of the information provided by the data. The parameterisation 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. See Aurenhammer (1991) for more details on Voronoi diagrams, and see Du and Foulger (1999) for a broad review of applications such as image compression, finite difference methods, quadrature, distribution of resources, cellular biology or territorial behaviour of animals.

Sambridge *et al.* (1995) and Sambridge and Gudmundsson (1998) were the first to propose the use of Voronoi polyhedra for tomographic problems. In 3D reflection tomography, Vesnaver *et al.* (2000) and Bohm *et al.* (2000) used Voronoi cells and adapted the local resolution iteratively, by means of a singular value analysis of the tomographic matrix. Sambridge and Faletic (2003) and Zhang and Thurber (2005) developed adaptive mesh seismic tomography methods based on tetrahedral and Voronoi diagrams to automatically match the inversion mesh to the data distribution. Nolet and Montelli (2005) optimised the spacing of interpolation support to

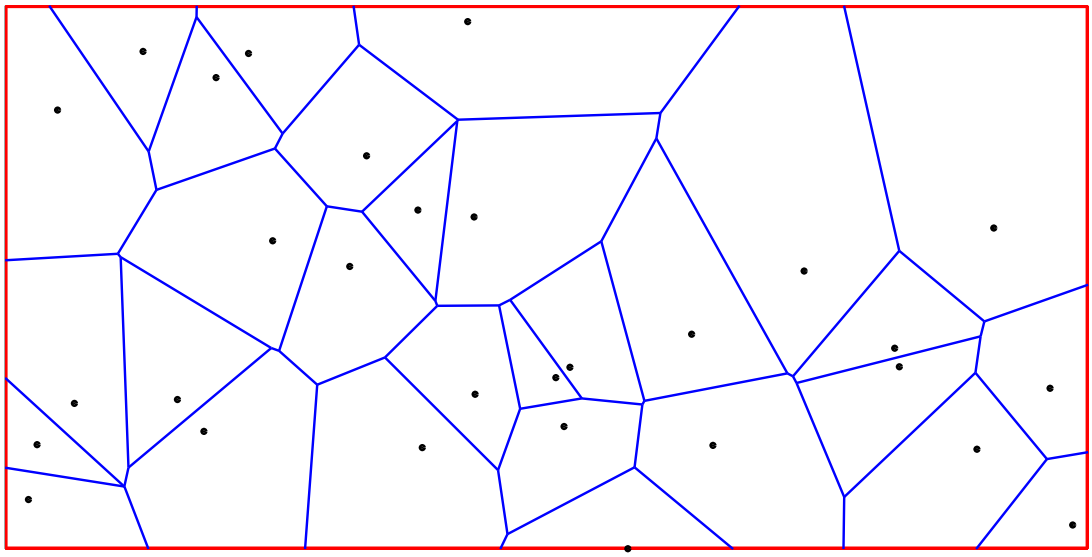


Figure 1.2: Voronoi cells about 30 pseudo random points on the plane. The cell nuclei (black dots) have been drawn from a 2D uniform distribution over the spatial domain delimited by the red rectangle. Each cell contains the part of the plane closest to its nucleus and so the shape of the parameterisation is entirely defined by the location of the nuclei. The boundaries of each cell are defined by the perpendicular bisector of each pair of points.

fit local resolution by connecting natural neighbours with springs of length equal to the local resolving length and minimising the potential energy of the system. In our case, we shall show that the Voronoi mesh self-adapts to the information contained in the data in a manner quite different to the methods described above.

1.5 Transdimensional models and Bayesian inference

Most of the studies that use irregular meshes have a fixed number of unknowns decided before hand, e.g. the number of layers or cells. Since the data fit can always be improved by adding more unknowns into the problem, there is a clear trade-off between explicative and predictive models or between complexity and accuracy. Our approach to problem solving can be considered as finding a solution that is, as Albert Einstein observed "as simple as possible but not simpler". Although this philosophy appears to be a solid principle, it is still not clear to geophysicists how it should be applied. A range of statistical techniques have been developed for

judging whether the choice of the model dimension is warranted by the data, for example, the Bayesian information criterion (Schwarz, 1978), the Akaike information criterion (Akaike, 1974) or F-tests (Aster *et al.*, 2005). To date, however, 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 work presented here, we propose to invert simultaneously for the number of Voronoi cells, for their geometries and for the seismic velocities. That is, the parameterisation 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, as within an optimisation 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 study, however, this is not the case within a Bayesian sampling framework. It turns out that high dimensional (many cell) models are naturally discouraged. This results from a substantial property of Bayesian inference referred to as 'natural parsimony', i.e. preference for the least complex explanation for an observation. In general, mathematical models with the smallest number of parameters are preferred as each parameter introduced into the model adds some uncertainty to it. Excessively complex models suffer from overfitting and have poor predictive power. Therefore, given a choice between a simple and complex model that provide similar fits to data, the simpler one will be favoured (See MacKay, 2003, for a discussion).

In a Bayesian formulation all information is represented in probabilistic terms (i.e. degrees of belief). Each individual model is given a probability of existence given the data. This probability distribution is taken as the complete solution of the inverse problem and allows assessment of trade-offs, constraints and resolution. Bayesian inference combines prior information on the model with the observed data to produce the posterior probability density function (standard references are Box and Tiao, 1973; Smith, 1991; Gelman *et al.*, 1995). Geophysical applications of Bayesian inference are described in Tarantola and Valette (1982), Duijndam (1988a,b) and Mosegaard and Tarantola (1995).

In the approach we propose here, the number and positions of the Voronoi nuclei are unknown in the problem and treated as model parameters. This makes the function g in Equation (1.2) highly non-linear, i.e. the posterior distribution is far from being a simple multidimensional Gaussian and may have a large number of maxima. Therefore there is no simple analytical formulation for the solution and the only practical way to determine the posterior probability distribution is to evaluate

it at different positions in the model space by mean of Monte Carlo algorithms.

Monte Carlo sampling and integration of multidimensional distributions is an active area of research in computational statistics. For summaries see Flournay and Tsutakawa (1989) and Smith and Roberts (1993). Over the past 15 years, geophysicists have begun to use Markov chain Monte Carlo (McMC) methods, which directly simulate the posterior distribution, that is, draw random samples distributed according to the posterior distribution (see Koren *et al.*, 1991; Gallagher *et al.*, 1997; Gouveia and Scales, 1998) and see Gallagher *et al.* (2009) for a recent review. McMC algorithms produce a chain of random models (Monte Carlo) where each new model only depends on the model previously generated (Markov chain). It can be shown that the distribution of samples produced by an McMC random walk will converge towards the posterior probability distribution when the number of iterations goes to infinity (Tierney, 1994). An application of McMC to a seismic inverse problem is given by Malinverno (2002). Note here that McMC is not the only way to sample the posterior distribution and algorithms such as particle filters can be used alternatively. See van Leeuwen (2009) for a recent review on particle filtering within a geophysical context.

Bayesian statisticians have considered the problem of a variable number of unknowns, for more than 10 years. As a consequence 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 recent review, see Sisson, 2005). At present, the reversible jump algorithm (rj-McMC) of Green (1995)(see also Green and Mira, 2001; Sambridge *et al.*, 2006; Gallagher *et al.*, 2009) 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 and Leaney (2000) to the inversion of zero-offset vertical seismic profiles. (For a more complete treatment see Malinverno and Leaney (2005)). Subsequent work was by Malinverno (2002) who applied it to electrical resistivity sounding, and Piana Agostinetti and Malinverno (2010) recently inverted teleseismic receiver functions to infer a 1D Earth model with variable number of layers.

In this study, we develop an entirely new approach to the tomography problem based on the reversible jump algorithm. The scheme we propose here is closely re-

lated to a process known as partition modelling (e.g. Denison *et al.*, 2002) which is a statistical analysis tool for non-linear classification and regression. Partition modelling has been applied successfully in disease mapping (e.g. Denison and 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, 2009).

1.6 Importance of data noise

As mentioned before, there is a trade-off between model simplicity and data fit. If too many unknown parameters are used in an inverse problem, the distance between estimated and observed data may become smaller than the actual data noise. In this case, the measurements are overfitted and the solution model may show spurious features due to the noise in the data. We shall show that in a transdimensional Bayesian formulation, the model dimension is directly adjusted in order to fit the data to the degree required by the estimated noise. Hence, the solution model depends on the data but also on the assumed level of noise. While this property can be seen as an advantage, the posterior solution given by the reversible jump scheme strongly depends on the estimated data uncertainty, and hence this can be a problem if the user knows little about the measurements errors.

Fortunately, an expanded Bayesian formulation called Hierarchical Bayes (Gelman *et al.*, 1995) can take into account the lack of knowledge we have about data errors. Instead of being fixed, the variance of the measurement errors can have a broad prior uncertainty and posterior inference can be performed. In geophysics, Malinverno and Briggs (2004) and Malinverno and Parker (2006) were the first to use a Hierarchical Bayes formulation and simultaneously invert for the data noise and model complexity. They demonstrated the practical application of this approach to a simple linear inverse problem: using seismic travel times measured by a receiver in a well to infer compressional wave slowness in a 1D Earth model. In their work, the posterior distribution was Gaussian, and its mean and covariance could be easily computed analytically. In the work presented here, we propose to apply Hierarchical Bayes to the fully non-linear tomographic problem, where the posterior is numerically estimated with the reversible jump McMC algorithm.

1.7 Organisation of the thesis

This thesis is structured such that the reader follows chronologically the work I have carried out during my PhD program. Chapter 2 presents the general inversion methodology and each new chapter proposes an improvement and/or an application of the same class of algorithms.

Chapter 2: A self-adaptive parameterising approach to tomographic inverse problems

After a detailed introduction to Bayesian methods and McMC sampling, our novel inversion methodology is presented in the case of a fixed number of Voronoi cells and in the general context of straight ray (i.e. linear) tomography. 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 parameterisations. We use data contaminated with random noise in order to test the ability of the method recover a synthetic model and to infer model uncertainty.

Chapter 3: Seismic tomography with the reversible jump algorithm

Here the method is extended to transdimensionality (i.e. variable number of Voronoi cells) by using the reversible jump algorithm. We also show how the method can be used iteratively to perform non-linear tomography including ray bending. Computational cost issues are treated and it is shown how the algorithm can be parallelised. After having illustrated the improved version on synthetic experiments, the method is used to invert ambient noise data to infer a tomographic image of Rayleigh wave group velocity for the Australian continent.

Chapter 4: Accounting for data Noise Uncertainty – Theory and Application to Palaeoclimate data

In this chapter we treat the issue of data noise and its relation to model complexity. The Bayesian formulation of the problem is extended, to treat as an unknown and estimate the uncertainty on the level of data noise. The purpose of this chapter is to introduce the Hierarchical Bayes methodology, and hence all ideas are simply illustrated on 1D regression problems that are either linear or nonlinear. Hence, the

reader can visually appreciate on a single figure the data vector, the data noise, the true and estimated model, the data fit, etc. An application to palaeoclimatology is presented where the goal is to infer the position and number of abrupt changes in noisy geochemical records.

Chapter 5: Multiscale seismic tomography with an Hierarchical Bayes formulation

In this chapter the Hierarchical Bayes methodology is applied to seismic tomography in the case of a multiscale problem. Three ambient noise datasets that span the Australian continent at different scales are simultaneously inverted. We show that the extended Bayesian formulation turns out to be particularly useful when dealing with multiple data types that have different unknown levels of noise. With scant prior knowledge on data errors, the algorithm is able to naturally adjust the fit to different datasets and to provide a velocity map with a spatial resolution adapted to the quantity of information present in the data.

Chapter 6: Transdimensional inversion of receiver functions with the Hierarchical Bayes algorithm

The purpose of this chapter is to show that the class of algorithm presented in this thesis is not restricted to seismic tomography but is rather a general approach to inverse problems. Here we propose to apply our methodology to invert receiver function waveforms (RF), which is a well known highly non linear problem. The particularity of receiver functions is that they are time series, and hence the data noise is correlated. We show how to 'parameterise' the data covariance matrix and invert for both the magnitude and correlation of noise. The algorithm is first tested on synthetic data contaminated by correlated synthetic noise, and then applied to data collected by a broadband station located on the Haiman island in China.

Chapter 7: Conclusions and future work

The results of each chapter are listed briefly. We summarise the main results, and outline directions for further study.

1.8 Publication Schedule

Some of the chapters in this thesis have already been published or submitted for publication. A summary of manuscripts is included here to ensure that adequate acknowledgment is given to co-authors. Note however, that I took the lead on developing theory as well as performing all numerical calculations and data analysis, excepted for the last paper submitted where the research carried out in chapter 4 was combined to other calculations.

- **Chapter 2** Bodin, T., Sambridge, M., and Gallagher, K. (2009), A self-parametrizing partition model approach to tomographic inverse problems, Inverse Problems, 25, 055009.
- **Chapter 3** Bodin, T and Sambridge,M., (2009), Seismic tomography with the reversible jump algorithm, Geophys. J. Int, 178, 1411-1436.
- **Chapter 4** Gallagher, K., Bodin, T., Sambridge, M., Weiss, D., Kylander, M., Large, D. (2010), Inference of abrupt changes in noisy geochemical records using Bayesian Transdimensional changepoint models. Submitted to Geochimica Cosmochimica Acta.

Chapter 2

A Self-parameterising Approach to Tomographic Inverse Problems

The aim of this first chapter is to introduce the partition modelling formulation of the tomographic problem, and hence we place ourselves in a common and simple situation, where the problem is to recover a 2D velocity field from a dataset made of travel times of seismic waves propagating from sources to receivers as straight rays. Here the number of Voronoi cells defining the velocity field is fixed, although in chapter 3 we treat the number of cells as a parameter itself in the inversion. This chapter is divided into 2 main sections. We first give a complete description of the Bayesian (seismic) tomography algorithm based on partition modelling. Then, we present results of synthetic experiments in a case 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.

2.1 Method

2.1.1 The model parameterisation

The seismic velocity field is discretised by a set of Voronoi polygons as shown in Figure 2.1. Given a set $\mathbf{c}$ of n nuclei in the 2D plane ($\mathbf{c} = \{\mathbf{c}_1, .., \mathbf{c}_n\}$ where $\mathbf{c}_i \; \epsilon \; \Re^2$), the Voronoi tessellation defines n non-overlapping regions $R_1, R_2, \; ... \; , R_n$ 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 Figure 2.1. Note that the cell nuclei are not

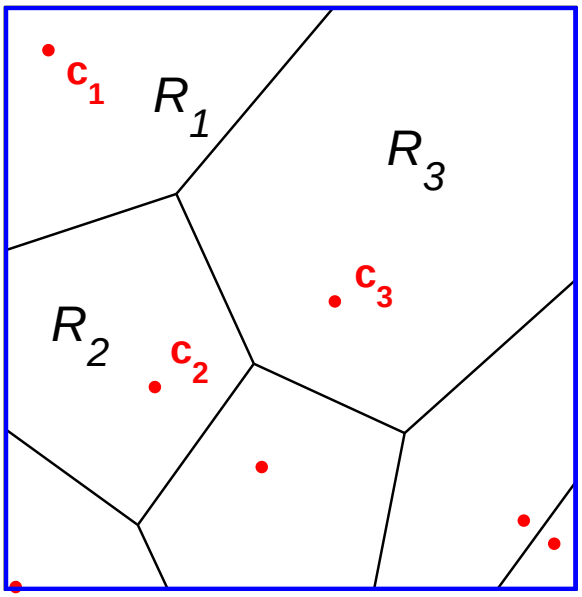


Figure 2.1: Examples of Voronoi diagram (black) which form a set of irregular cells that partition the plane. Each cell contains the part of the plane R_i closest to its nucleus $\mathbf{c}_i$ (red dot), and so the shape of the parameterisation is entirely defined by the location of the nuclei.

necessarily in the geometric centre, rather the cell boundaries are perpendicular bisectors of each pair of neighbouring nuclei. Each cell is therefore characterised by the two coordinates of its nucleus $\mathbf{c}_i$ and for the seismic tomography problem, by a constant 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 ($\mathbf{v} = (v_1, .., v_n)$ where $v_i \; \epsilon \; \Re$). Here we use only the simplest possible representation of velocity within each partition, i.e. a constant. Higher order polynomials are possible, e.g. a linear gradient or quadratic, which would require additional unknowns for each cell.

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. 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 parameterisation will adapt to the spatial variability in the resolving power of the data and so maximise the information extracted.

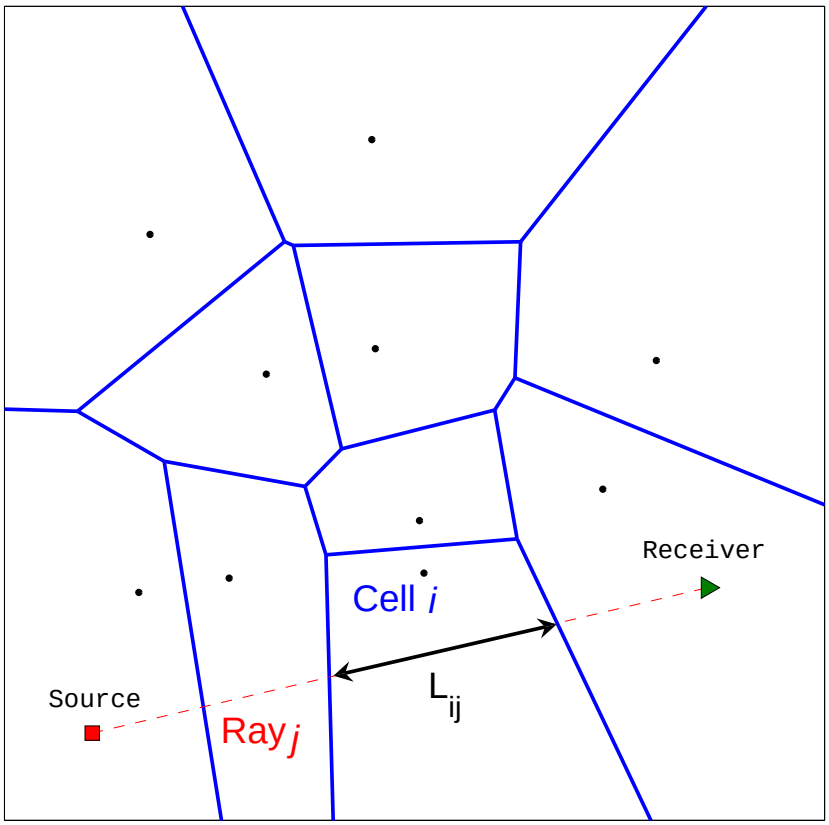


Figure 2.2: A seismic ray (dashed red line) joins a source to a receiver with a straight line. The travel time t_j is simply computed by integrating the inverse of the cell velocity v_i along the ray path. That is, using the length $\mathbf{L}_{ij}$ of the ray across each cell

2.1.2 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 j^{th} ray is then simply given by:

$$t_j = \sum_{i=1}^{n} \frac{\mathbf{L}_{ij}}{v_i} \tag{2.1}$$

where $\mathbf{L}_{ij}$ is the length of ray j across cell i (see Figure 2.2) and v_i is the velocity value assigned to cell i . If the ray j does not pass through cell i, then $\mathbf{L}_{ij} = 0$.

Equation (2.1) will be recognised by the reader as a standard linear tomographic system of equations in slownesses (i.e. $1/v_i$).

In the examples we subsequently consider, the rays remain straight and do not depend on the velocity field. This assumption is reasonable for many seismic problems and is also relevant in some practical cases as in x-ray tomography, ultrasonic computed tomography (Natterer, 2001) or in teleseismic tomography (e.g. Aki *et al.*, 1977; Graeber *et al.*, 2002).

2.1.3 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 consists 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, like 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 subsequent sections, we first consider noise free synthetic data and later add noise to see how it propagates into model uncertainty.

2.1.4 Bayesian formulation

Having parameterised 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) and Gelman *et al.* (1995). Summaries within a geophysical context are given by Tarantola and Valette (1982), Duijndam (1988a), and

Mosegaard and Tarantola (1995). The aim is to recover the posterior distribution: that is, the model probability density 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 multidimensional 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:

$$p(\mathbf{m} \mid \mathbf{d}_{obs}) = \frac{p(\mathbf{d}_{obs} \mid \mathbf{m})p(\mathbf{m})}{p(\mathbf{d}_{obs})} \tag{2.2}$$

where $x|y$ means x given, or conditional on, y, i.e. the probability of having x when for a given value of y. $\mathbf{d}_{obs}$ is a vector defined by the set of observed data and $\mathbf{m}$ is the vector of the model parameters. $p(\mathbf{d}_{obs} \mid \mathbf{m})$ is the likelihood of observing the 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 $\mathbf{d}_{obs}$. The term, $p(\mathbf{d}_{obs})$ 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 $\mathbf{m}$. Note however that in the next chapters the number of Voronoi cells n wil be concidered as an unknown, and the conditional evidence $p(\mathbf{d}_{obs}|n)$ will play a central role. It will be estimated indirectly with the number of samples generated by the algorithm in each state space. The total evidence $p(\mathbf{d}_{obs})$ term will be disregarded throughout this study, and we therefore write (2.2) as a proportionality relationship:

$$p(\mathbf{m} \mid \mathbf{d}_{obs}) \propto p(\mathbf{d}_{obs} \mid \mathbf{m})p(\mathbf{m}) \tag{2.3}$$

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.*, 1995).

2.1.5 The likelihood

The likelihood $p(\mathbf{d}_{obs} \mid \mathbf{m})$ 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

$$\Phi(\mathbf{m}) = \left\| \frac{g(\mathbf{m}) - \mathbf{d}_{obs}}{\sigma_d^2} \right\|^2 \tag{2.4}$$

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

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

$$p(\mathbf{d}_{obs} \mid \mathbf{m}) \propto \exp\left(\frac{-\Phi(\mathbf{m})}{2}\right) \tag{2.5}$$

2.1.6 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. A weakness of the Bayesian formulation is that only information that can be expressed as a probability distribution can be made use of. 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 practice many authors simply choose a convenient probability density function, somewhat contrary to strict Bayesian principles.

In this work, we assume minimal prior knowledge and use a 'nearly' uninformative uniform priors. Note we used the phrase 'nearly' to acknowledge that all priors contain information and it is not possible to have a completely uniformative prior.

Although uniform distributions are very informative about their bounds, in this study we use "wide" prior distributions, i.e. the likelihood distribution is virtually null at bounds defining the prior, and hence the values of bounds do not affect the posterior distribution. In this sense, we pretend to use uninformative priors.

In this thesis we only consider priors that are separable and hence can be written as a product of independent 1-D priors on each variable. In some cases one might want to introduce joint priors on a subset of the variables, for example by making the prior variance of velocity in each Voronoi cell dependent on the size of the cell. In principle this could be done with additional calculations to determine each prior (e.g. to calculate Voronoi areas), but the algorithm would then be more difficult to implement and more computationally expensive. Hence, the prior probability distributions for the $3n$ parameters, 2D Voronoi centres and velocities in each cell, are independent from each other, and so can written in separate form:

$$p(\mathbf{m}) = p(\mathbf{c})p(\mathbf{v}). \tag{2.6}$$

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

$$p(v_i) = \begin{cases} 1/(\Delta V) & \text{if} \quad V_{min} < v_i < V_{max} \\ 0 & \text{otherwise} \end{cases} \tag{2.7}$$

and since the velocity in each cell is independent,

$$p(\mathbf{v}) = \prod_{i=1}^{n} p(v_i). \tag{2.8}$$

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 2.1 and borders the area to map, i.e. the zone covered by seismic rays. Hence we have

$$p(\mathbf{c}_i) = \begin{cases} 1/(\Delta x \Delta y) & \text{if} \quad \mathbf{c}_i \in A \\ 0 & \text{otherwise} \end{cases} \tag{2.9}$$

where Δx and Δy are the dimensions of the spatial domain A. Since the position

of each nucleus is independent, we also have

$$p(\mathbf{c}) = \prod_{i=1}^{n} p(\mathbf{c}_i). \tag{2.10}$$

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

2.1.7 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) and 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.

In this work we use the well-known Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970), which 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 $q(\mathbf{m}' \mid \mathbf{m})$ conditional only on the current model $\mathbf{m}$. 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.

A simple example of a proposal distribution would be a Gaussian distribution

with zero mean and diagonal covariance matrix $C = \text{diag}[\sigma_1^2, \sigma_2^2, \dots]$

$$q(\mathbf{m}' \mid \mathbf{m}) \propto \exp\left(-\frac{1}{2}(\mathbf{m} - \mathbf{m}')^T C^{-1}(\mathbf{m} - \mathbf{m}')\right). \tag{2.11}$$

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

$$\mathbf{m}' = \mathbf{m} + u\sigma_i \mathbf{e}_i \tag{2.12}$$

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 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, i.e. probability of generating a perturbed model at $\mathbf{m}'$ when starting from $\mathbf{m}$ is the same as the reverse proposal distribution, i.e. probability of starting from $\mathbf{m}'$ and generating a sample at $\mathbf{m}$. Hence we have $q(\mathbf{m}' \mid \mathbf{m}) = q(\mathbf{m} \mid \mathbf{m}')$.

Once a proposed model has been drawn from the distribution $q(\mathbf{m}' \mid \mathbf{m})$, the new model is then accepted with a probability $\alpha(\mathbf{m}' \mid \mathbf{m})$, i.e. a uniform random deviate, U, is generated between 0 and 1. If $U \leq \alpha$, the move is accepted, the current model $\mathbf{m}$ is replaced with $\mathbf{m}'$ and the chain moves to the next step. If $U > \alpha$, 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, $\alpha(\mathbf{m}' \mid \mathbf{m})$, is the key to ensuring that the samples will be generated according to the target density. Gilks *et al.* (1996) showed that the chain will converge to the posterior distribution, $p(\mathbf{m} \mid \mathbf{d}_{obs})$, if

$$\alpha(\mathbf{m}' \mid \mathbf{m}) = min\left(1, \frac{p(\mathbf{m}' \mid \mathbf{d}_{obs})}{p(\mathbf{m} \mid \mathbf{d}_{obs})} . \frac{q(\mathbf{m} \mid \mathbf{m}')}{q(\mathbf{m}' \mid \mathbf{m})}\right). \tag{2.13}$$

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) and Neal *et al.* (1993). Another way to produce independent samples is to parallelise 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 chapter 3).

2.1.8 Implementation of the algorithm

Our implementation of the algorithm is as follow. Having randomly initialised 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 updates its position $\mathbf{c}_i$ or its seismic velocity value v_i. Each step is divided in three stages:

1. 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 on the current position $\mathbf{c}_i$ (See Figure 2.3). The covariance matrix for the 2D Gaussian function is proportional to the identity matrix, with the constant of proportionality a parameter to be chosen.

2. Solve the forward problem : Sort the rays passing through the cells, whose geometry or velocity has been modified, and with (2.1) update the estimated travel times for only the rays affected. The new estimated travel times are compared with the observations to build the misfit (2.4), the likelihood (2.5) and the posterior value of the proposed model $p(\mathbf{m}' \mid \mathbf{d}_{obs})$.

3. Randomly decide whether or not to replace and update the current model according to the acceptance probability distribution $\alpha(\mathbf{m}' \mid \mathbf{m})$ given in (2.13). Since proposal distributions are symmetric (i.e. $q(\mathbf{m}' \mid \mathbf{m}) = q(\mathbf{m} \mid \mathbf{m}')$), we have

$$\alpha(\mathbf{m}' \mid \mathbf{m}) = min\left(1, \frac{p(\mathbf{m}' \mid \mathbf{d}_{obs})}{p(\mathbf{m} \mid \mathbf{d}_{obs})}\right). \tag{2.14}$$

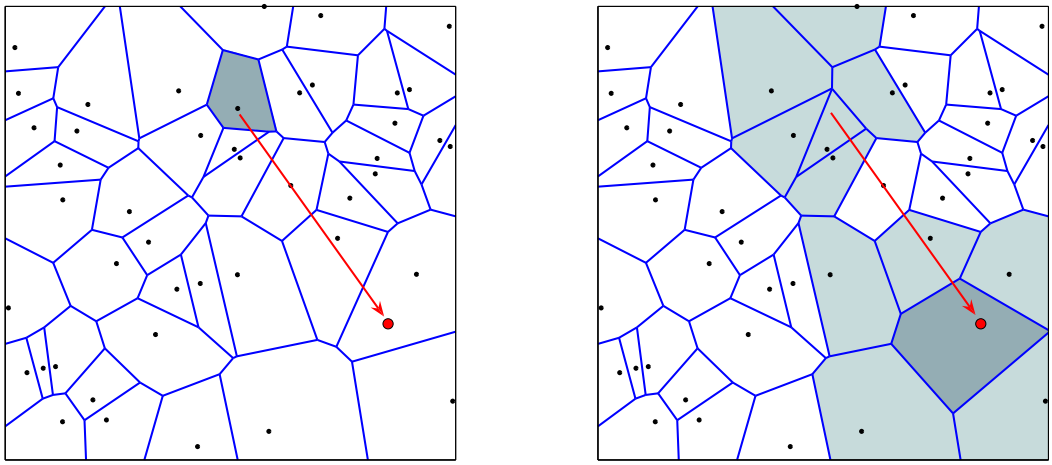


Figure 2.3: Move of a Voronoi nucleus. The two panels represent the Voronoi tessellation before and after the move. The new position (red circle) is drawn from a 2D Gaussian probability distribution centred on the old position of the nucleus. Note that only the cells in grey are affected by the move.

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. If the priors are always the same, the acceptance condition is based on the ratio of the likelihoods. When the proposed model is rejected, the current model is retained for the next iteration and also added again to the samples collected. Note that the proposed model can fall outside the range defined by the uniform prior distribution. In this case, the prior (and hence the posterior) of the proposed model is null, and the model is rejected. In this way the proposal distributions are seen as true Gaussians rather than truncated versions.

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.* (1995).

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), and Brooks and Roberts (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) and 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 normalising constant in (2.3) 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 achieve 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 is performing 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 may not be properly sampled and the chain may not converge at all. 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.

2.1.9 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 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 properties of an ensemble instead of a single model. Importance sampling provides the possibility to get 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 :

$$E[f(\mathbf{m})] = \int f(\mathbf{m}) p(\mathbf{m} \mid \mathbf{d}_{obs}) d\mathbf{m} \tag{2.15}$$

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):

$$E[f(\mathbf{m})] = \frac{1}{S} \sum_{s=1}^{S} f(\mathbf{m}_s) \tag{2.16}$$

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). All the models sampled have particular parameterisation defined by their cell geometry. When a large number of models with different parameterisations are stacked, the Voronoi polyhedra overlap so the 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. We take the average velocity field as a representative 'solution' to our inverse problem, although having quantified the posterior distribution we clearly have more information on the model space.

In the case that $f(\mathbf{m}) = \big(\mathbf{m} - E[\mathbf{m}]\big)^2$, where $E(\mathbf{m})$ is the average velocity field (or posterior mean), the expectation $E[f(\mathbf{m})]$ is then the variance of the posterior. It is directly obtained by taking the variance of all the models sampled. Here we

take the continuous map of the posterior standard deviation $\sqrt{E[f(\mathbf{m})]}$ as an 'error' map about the expected mean solution.

It is important to collect enough independent samples so that the solution maps are stationary and represent well the posterior mean and variance. Integrating the information coming from different models with different parameterisation 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-regularising inversion algorithm.

2.2 Synthetic data examples

2.2.1 Experimental setup

To illustrate the new algorithm in the case of an uneven 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 2.4. The upper-left part of the model is covered by many ray paths, whereas the lower-right part is left with relatively few.

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 2.5 shows the true velocity field that we try to recover. The areas in red have a velocity of 5 km/s and the velocity in the blue areas are of 4 km/s. The velocity field contains high contrast discontinuities. It will be shown that the dynamic parameterisation 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

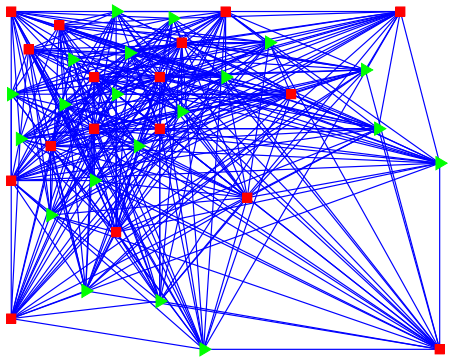


Figure 2.4: Geometry of rays. 340 rays join 17 sources (red squares) to 20 receivers (green triangles).

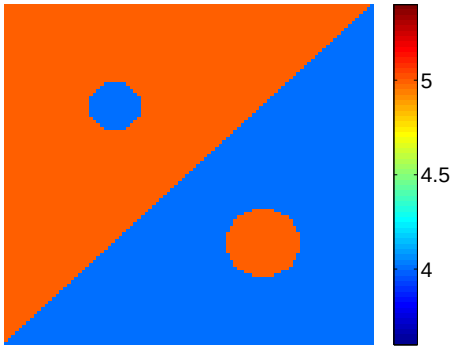


Figure 2.5: True velocity field (km/s).

velocity structure. While this is only a first order approximation for the seismic problem, it serves well to demonstrate the approach we are advocating.

We compare our Monte Carlo sampler to some widely used regularisation 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.

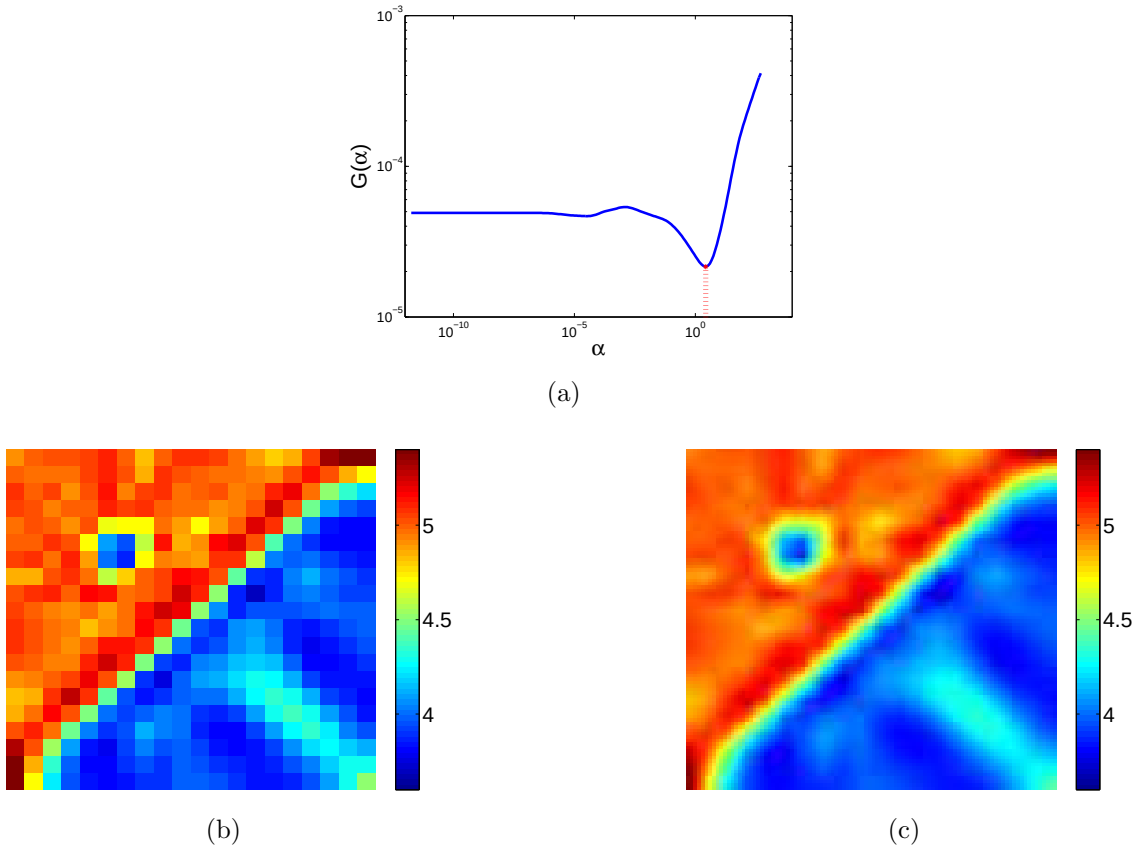
2.2.2 Noise free experiment: Tikhonov regularisation vs partition modelling

2.2.2.1 The regularisation process in linear inversion

Most of the methods using a predefined fixed parameterisation 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. Regularisation procedures must be used to choose a solution among all the acceptable possibilities. This resulting solution will have properties reflecting the particular choice of regularisation. A common choice of regularisation is to use the norm of the first or second derivative of the model (Aster *et al.*, 2005). The inversion scheme consists then in minimising a linear combination of two criteria, and takes the form :

$$min\left[\|\mathbf{Gm}-\mathbf{d}\|^2+\alpha^2\|\mathbf{Dm}\|^2\right] \tag{2.17}$$

where the first term is the data misfit and **Dm** is a finite difference approximation

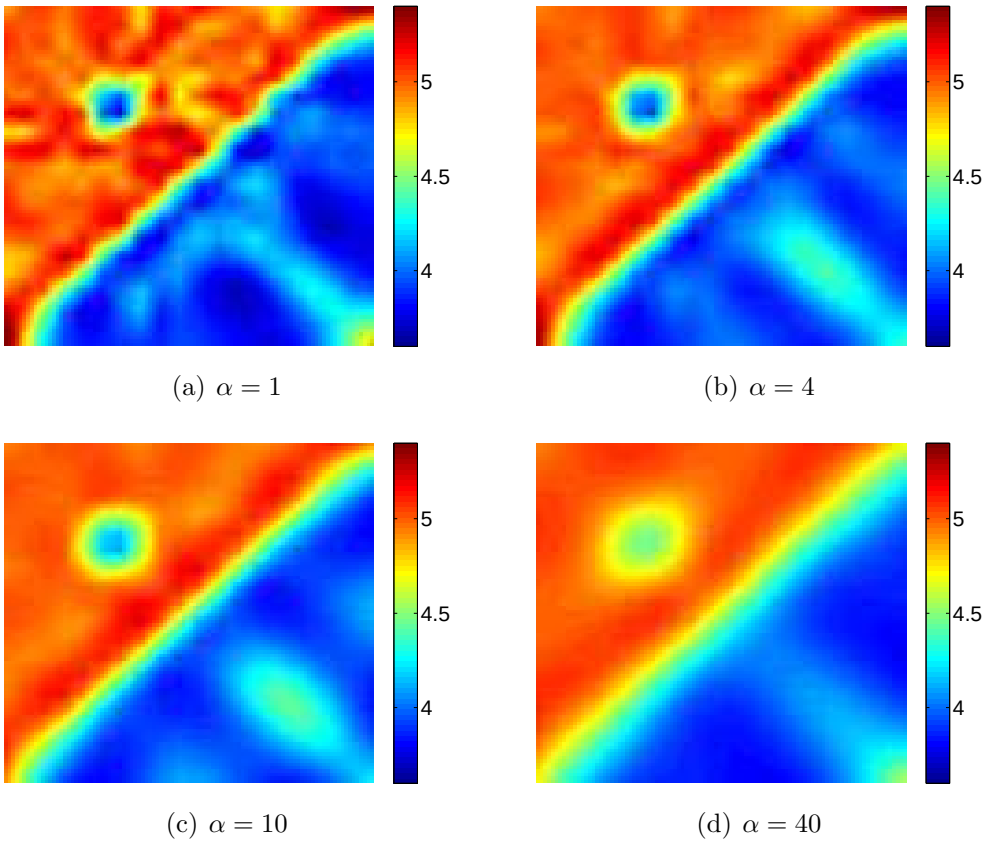


(a)

(b) (c)

Figure 2.6: Noise free experiment. Linear inversion after linear interploation over a regular grid : second order Tikhonov regularisation 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)). (c) Solution Model after Linear Interpolation (km/s). The colour scale are the same as for the true model.

that is proportional to the first or second derivative of $\mathbf{m}$. By minimising the semi-norm $\|\mathbf{Dm}\|^2$, the regularisation techniques will favour models that are relatively flat (first order regularisation) or smooth (second order regularisation). The parameter α gives relative weight to the two terms and is called the regularisation factor (note here α is different from the acceptance probability used in a Bayesian inversion and defined in (2.13)). Determining a solution through solving the problem (2.17) is known as a Tikhonov regularisation. For any given value of α, the solution to (2.17) is unique and can be easily found with least square optimisation 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 regularisation factor. Here



(a) $\alpha = 1$ (b) $\alpha = 4$

(c) $\alpha = 10$ (d) $\alpha = 40$

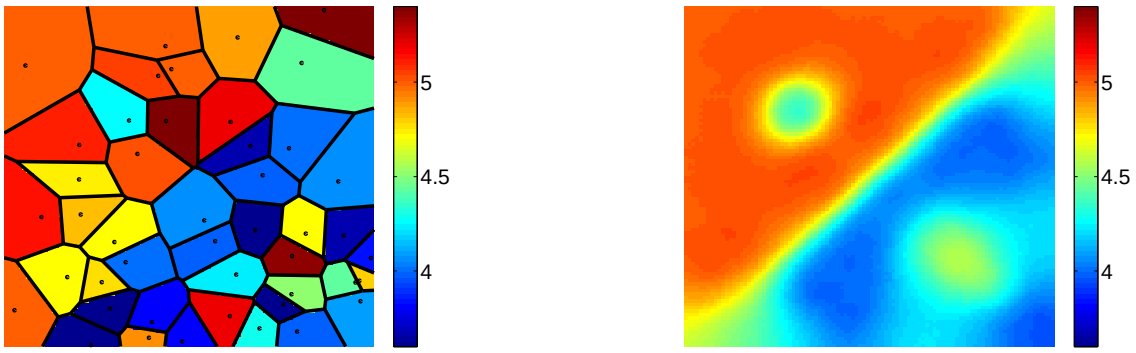
Figure 2.7: Linear Inversion with a regular grid (20 × 20 cells). Noise-free experiment. Results for different values of the regularisation parameter (km/s).

we adopt the General Cross Validation (GCV) method (See Aster *et al.* (2005) for details).

The GCV function obtained with a second order Tikhonov regularisation for a uniform grid of 400 cells is plotted in Figure 2.6(a). The red dashed line shows the minimum of the GCV function which correspond to an α of 2.6. The corresponding solution model after solving (2.17) with this value of α is shown in Figure 2.6(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 2.6(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 regulari-



(a) (b)

Figure 2.8: Partition modelling with 45 mobile Voronoi cells. Results for the noise-free experiment. (a) Best model sampled: posterior maximum (km/s). (b) Average solution model: posterior mean (km/s).

sation parameter α. We observed the classic trade-off between smooth models with poor spatial resolution and instability (Menke, 1989) (see Figure 2.7). 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 2.6 and 2.7 represent the best results obtained from a range of experiments within a regularisation framework.

2.2.2.2 Tomography with partition modelling

As mentioned above, our Monte Carlo sampler does not need explicit regularisation to produce a single smooth velocity field. We select a representative 'solution' as the average of an ensemble of models. Note that each individual model consists of a different configuration of a finite number of Voronoi cells (see Figure 2.8(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 2.8(a). The average of the post burn-in samples collected (i.e. the posterior mean) is shown in Figure 2.8(b). The scales are the same as 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 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 regularisation framework (Figure 2.7(c)) is obtained with $\alpha = 10$, and this represents a compromise across the entire 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 parameterisation but constructively reinforces the well constrained ones. That is, real discontinuities in the velocity field are preserved because the dynamic parameterisation has been able to adapt to the structural features of the unknown model (i.e. many models approximate the significant features well).

2.2.3 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, 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.

2.2.3.1 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 regularisation scheme. As in the noise-free case, we tried to use the first and second order Tikhonov regularisation 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 2.9 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. The panel 2.9(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 regularisation 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 regularisation stabilises 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 2.9(d). The errors on the solution model are clearly underestimated.

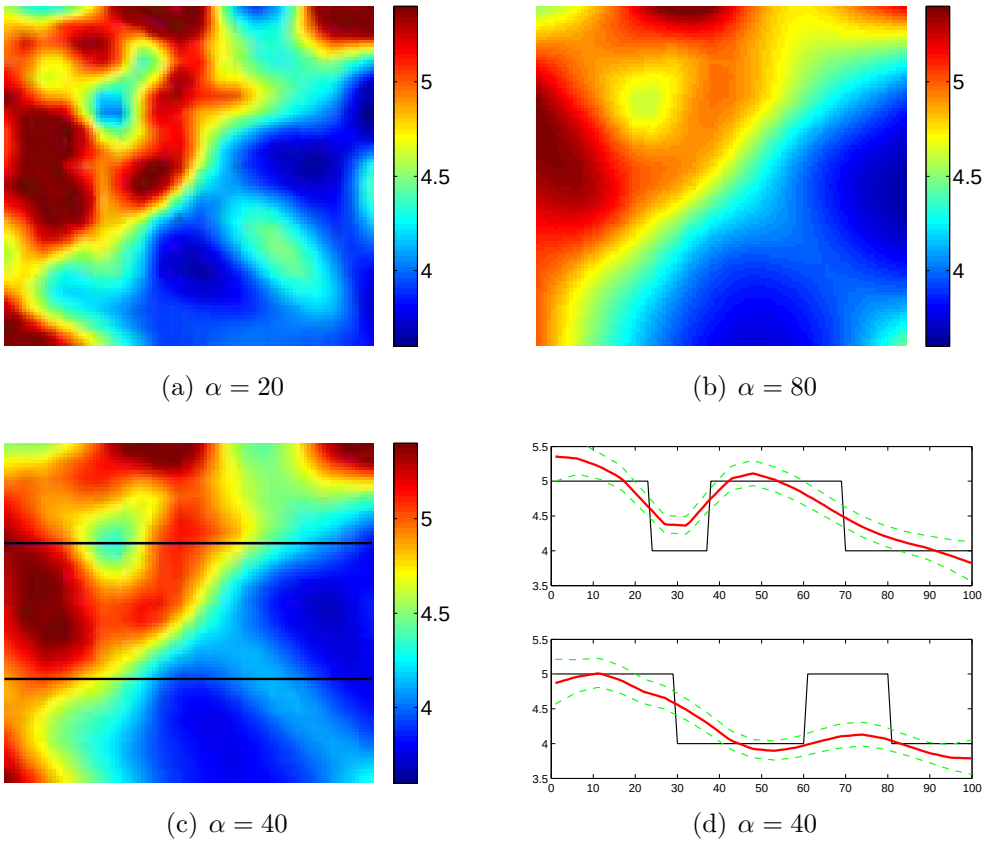
2.2.3.2 Partition modelling with noise

In contrast to regularisation 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 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 2.10. The map 2.10(a) shows the posterior mean 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 2.10(c) shows the map of the posterior standard deviation, i.e. the estimated error. The map 2.10(d) shows the true error for the average solution (i.e. the absolute value of the difference between the true model (Fig. 2.5) and the average solution (2.10(a))). The panel 2.10(b) shows the cross sections of the map as in Figure 2.9.

The model uncertainty map obtained in this way (2.10(c)) appears to be strikingly similar to the actual error (2.10(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 on panel 2.10(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.

In these experiments, the Monte Carlo sampler provides a reliable estimation of



(a) $\alpha = 20$ (b) $\alpha = 80$

(c) $\alpha = 40$ (d) $\alpha = 40$

Figure 2.9: Results with 12 % random Gaussian noise. Linear inversion (after interpolation) with a regular grid of 400 cells (km/s). (d) cross section for y=30 km and y=70 km (km/s). Black: true model, red: solution model. The dashed green lines are the average model plus and minus one value of the interpolated standard deviation.

the model uncertainty both in terms of amplitudes and lateral variations. This is a considerable advantage over the linear methods using a fixed grid.

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

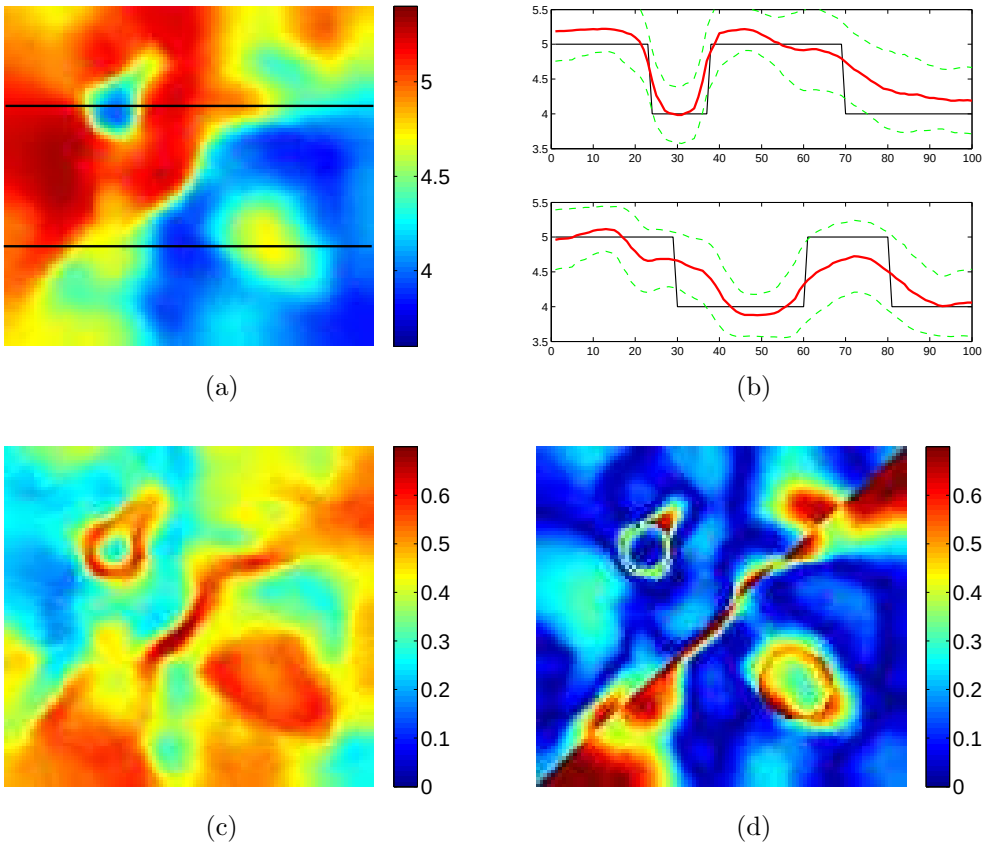


Figure 2.10: Results with 12 % random Gaussian Noise. Partition modelling with 45 mobile Voronoi cells. (a) Average model: posterior mean (km/s). (b) cross section for y=30 km and y=70 km (km/s). The black line is the amplitude of the true model, the red line is the estimated average model.The dashed green lines are the average model plus and minus one value of the local standard deviation. (c) Estimated error: posterior standard deviation (km/s). (d) True error. Absolute difference between the true and the estimated model (km/s).

data. This can easily be seen on the map of the estimated error on Figure 2.10(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 were the problem is ill-posed. In a situation where a lot of models almost equally satisfy the data, regularisation 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 optimisation methods.

2.3 Conclusion

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 utilised partition modelling which allows efficient exploration of the model space by sampling models of varying parameterisation. Although the numerical experiments reported in 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 regularisation of the model parameters is not required, thus avoiding global damping procedures and the often subjective process of finding an optimal regularisation value.

Second, the McMC approach lets us consider the issue of model parameterisation (e.g. the way of discretising the velocity field) as part of the modelling process. The velocity field has been parameterised 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 parameterisation 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 parameterisation involves fewer parameters to achieve better resolution than a fixed grid.

Another advantage is that the 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 accurate estimation of the velocity uncertainty as a novel result from this study.

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 parameterisation used here is highly novel in geophysical applications and would appear to have considerable potential for similar problems elsewhere. Moreover, the optimisation of the McMC sampler has proved to be efficient in terms of computational costs and our preliminary results are encouraging for applications to larger datasets. Here the method has been tested for straight ray linear tomography, but we show in

next chapter that, if used iteratively, it can be extended to non-linear tomography with bent rays at increased computational cost.

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.

Chapter 3

Seismic Tomography With the Reversible Jump Algorithm

In this chapter we improve and refine the partition modelling tomography algorithm presented previously. A first improvement involves the forward calculation. Instead of assuming straight rays, here ray theory is used to compute ray geometries from a reference velocity field. We show how the algorithm can be used iteratively to perform a non-linear tomography including ray bending. A second improvement is the extension to transdimensionality, that is the number of model parameters becomes an unknown in the problem. In this way, the number of Voronoi nuclei is variable during the inversion so the level of detail in the solution is directly dictated by the data (see Figure 3.1). We introduce the reversible jump algorithm (rj-McMC) (Green, 1995), and show how it can be used to add or remove cells in the partition models. Therefore the parameterisation becomes fully adaptive, since the user does not need to specify model complexity before hand. The number of unknowns is not fixed in the problem, and hence the posterior probability distribution is defined in a number of spaces with different dimensions. If the model is defined by $3n$ unknowns (2D coordinates of each nuclei + 1 velocity value for each cell, where n is the number of Voronoi nuclei), the posterior will be a function of $3n$ variables. This transdimensional probability distribution is taken as the complete solution of the inverse problem. We show how the rj-McMC algorithm is implemented to generate an ensemble of vectors $\mathbf{m}[= (\mathbf{c}, \mathbf{v})]$ with variable length, whose density reflects that of the posterior distribution. Finally, the algorithm has been optimised by direct parallelisation, and also by implementing the delayed rejection scheme which allows for adaptive proposal distributions.

This chapter is constructed in a similar manner as chapter 2. In a first section,

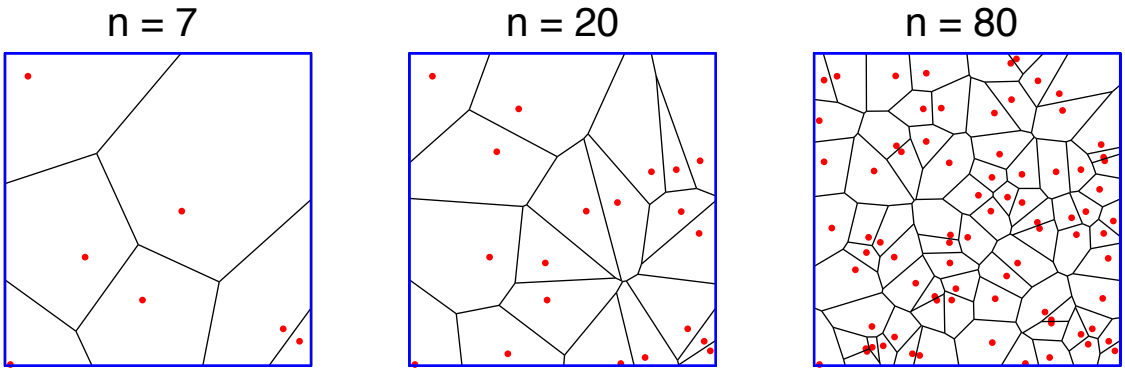


Figure 3.1: Examples of Voronoi diagrams (black) which form a set of irregular cells that partition the plane. The three panels show 7, 20 and 80 randomly distributed nuclei respectively. As the number and position of the nuclei change the Voronoi diagram corresponds to a multi-scale parameterisation of the velocity field.

we give a detailed presentation of the reversible jump tomography. Then, we show how the algorithm has been optimised, give some measure of computational cost, and show how the sampling efficiency is largely improved when multiple Markov chains simultaneously sample the model independently of each other . In section 3.3, we present results of synthetic experiments in a similar situation as chapter 2 , i.e. where the ray coverage is far from ideal, and compare our approach to standard methods that use regular (ideal) parameterisations. We use data contaminated with random noise in order to test the ability of the method to infer model uncertainty. In section 3.4, the method has been used with ambient noise data to infer a tomographic image of Rayleigh wave group velocity for the Australian continent.

3.1 Method

3.1.1 An iterative linearised approach

Given a complete description of a velocity field, the theory of seismic wave propagation allows us to predict travel times 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 and Brown, 2003). We employ the Fast Marching Method (FMM) (Sethian and Popovici, 1999; Rawlinson and Sambridge, 2004) to calculate travel times and raypaths in a laterally heterogeneous 2D medium.

Since the ray paths are dependent on the velocity model, the tomographic prob-

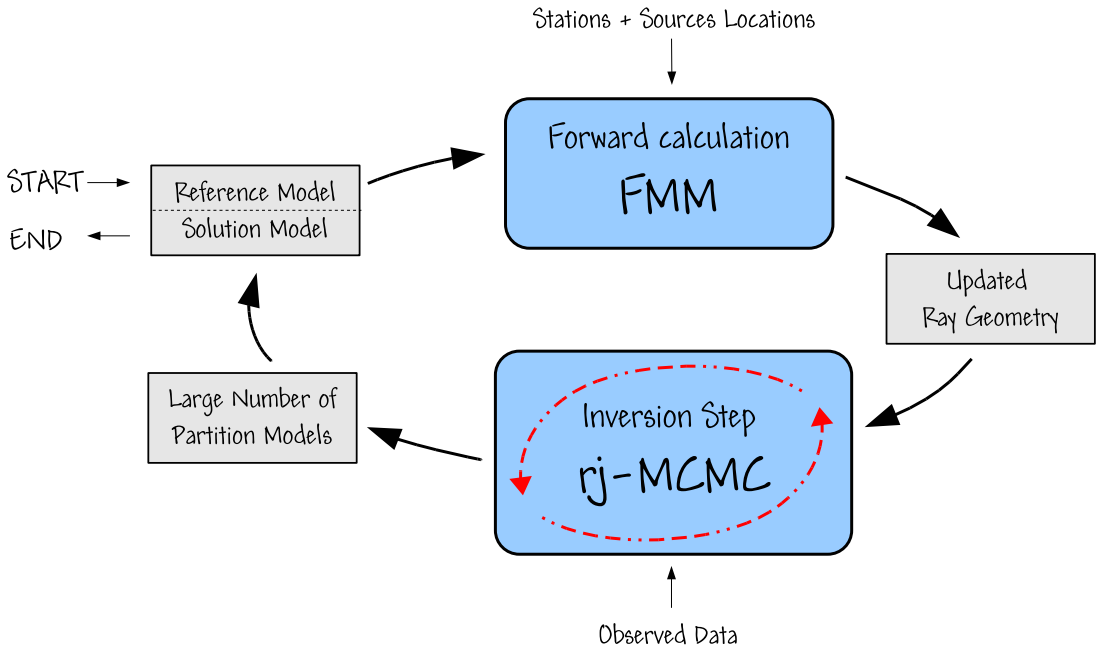


Figure 3.2: The transdimensional McMC sampler is used in conjunction with the Fast Marching eikonal solver to build an iterative linearised solution method. The inputs are an initial continuous reference velocity model, the stations and sources locations and the observed travel times. At each iteration, the Markov chain in the 'inner loop' (dashed red arrows) produces an ensemble of potential solutions which are spatially averaged to produce a reference model for the next iteration of the outer loop (black arrows). Only in the outer loop are raypaths updated.

lem is nonlinear. The development proposed here takes into account non-linearity by iteratively updating raypaths (e.g. Gorbatov *et al.*, 2001). Our scheme can be described using two loops as depicted in Figure 3.2. The outer loop (solid black arrows in Figure 3.2) is similar to many previous tomographic schemes. At each iteration ray paths and travel times 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 Figure 3.1) 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 linearised 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 Markov chain Monte Carlo 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, i.e. 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 sections 2.1.9 and 3.1.7)

The Markov chain algorithm within the inner loop (dashed red arrows in Figure 3.2) requires calculation of travel times for a large number of partitioned velocity models of the type in Figure 3.1. 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

$$t = \int_{R_0} \frac{1}{v(\mathbf{x})} dr \tag{3.1}$$

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 explicitely linearise the travel time expression in (3.1) 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 raypaths are kept fixed (3.1) is equivalent to the linearisation in slowness commonly used in tomographic algorithms, i.e. travel times given by (3.1) are the same as those obtained by evaluating

$$t = \int_{R_0} s_o(\mathbf{x}) dr + \int_{R_0} \delta s(\mathbf{x}) dr, \tag{3.2}$$

where $v_o(\mathbf{x})$ is the reference velocity field and $s_o(\mathbf{x}) = 1/v_o(\mathbf{x})$. Hence travel times given by (3.1), are equivalent to a first order accurate slowness perturbation. For fixed velocities in Voronoi cells, the travel time of the jth ray is then given by

$$t_j = \sum_{i=1}^{n} \frac{L_{ij}}{v_i} \tag{3.3}$$

where L_{ij} is the length of ray j across cell i (see Figure 3.3) 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

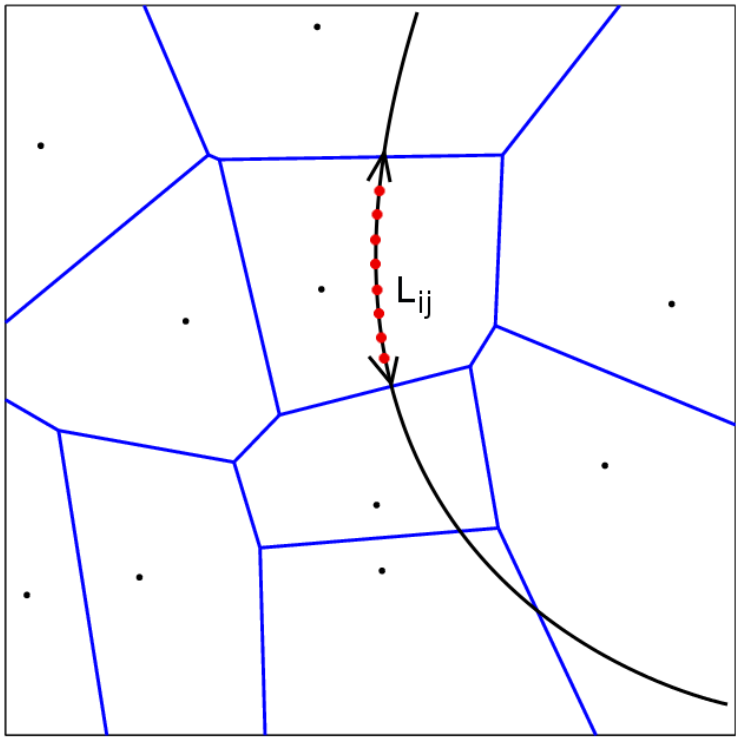


Figure 3.3: A seismic ray (thick black line) is bent according to a reference velocity model. For a given partition model, the estimated travel time t_j for the jth ray is simply computed by integrating the inverse of the cell velocity v_i along the ray path. That is, using the length L_{ij} of the ray across each cell.

lengths in the cells are calculated by sampling along rays at a predetermined step length (red points in Figure 3.3) and by finding the cell containing the midpoint of the current ray segment. Point location within 2D Voronoi cells is efficiently implemented with the scheme described in Sambridge and Gudmundsson (1998). Note, the curvature of the ray in Figure 3.3 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 parameterisation 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.

3.1.2 The prior

As in chapter 2, 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,

$$p(\mathbf{m}, n) = p(\mathbf{m} \mid n)p(n). \tag{3.4}$$

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

$$p(n) = \begin{cases} 1/(\Delta n) & \text{if} \quad n \,\epsilon\, I \\ 0 & \text{otherwise} \end{cases} \tag{3.5}$$

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

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

$$p(\mathbf{m} \mid n) = p(\mathbf{c} \mid n)p(\mathbf{v} \mid n). \tag{3.6}$$

Even though in the prior the parameterisation variables $\mathbf{c}$ are independent of the velocity variables $\mathbf{v}$, 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 $J = \{v_i \,\epsilon\, \Re \mid V_{min} < v_i < V_{max}\}$. Hence we have

$$p(v_i \mid n) = \begin{cases} 1/(\Delta v) & \text{if} \quad v_i \,\epsilon\, J \\ 0 & \text{otherwise} \end{cases} \tag{3.7}$$

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

$$p(\mathbf{v} \mid n) = \prod_{i=1}^{n} p(v_i \mid n). \tag{3.8}$$

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 $\left(\frac{N!}{n!(N-n)!}\right)$ 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

$$p(\mathbf{c} \mid n) = \left[\frac{N!}{n!(N-n)!} \right]^{-1}. \tag{3.9}$$

Combining together (3.5), (3.7), (3.8), and (3.9), the full prior probability density function can be written as

$$p(\mathbf{m}) = \begin{cases} \dfrac{n!(N-n)!}{N!(\Delta v)^n \Delta n} & \text{if} \quad (n \,\epsilon\, I \quad \text{and} \quad \forall i \,\epsilon\, [1,n],\ v_i \,\epsilon\, J) \\ 0 & \text{otherwise} \end{cases} \tag{3.10}$$

By multiplying (3.10) by the likelihood, the posterior distribution can be evaluated for any given model $\mathbf{m}$. The task is then to generate samples whose density follows the posterior distribution.

3.1.3 Principle of the reversible jump Markov chain Monte Carlo

In the transdimensional case, the dimension of the model space is itself a variable ($3n$) and the posterior becomes a transdimensional function. This can be sampled with a generalisation of McMC (see chapter 2) called reversible jump (Green, 1995) which allows inference on both model parameters and model dimensionality. rj-McMC is an extension of the afore-presented Metropolis-Hasting algorithm (Metropolis et al., 1953; Hastings, 1970), and similarly consists of a two stage process of proposing a model probabilistically and then accepting or rejecting it. As in the fixed dimension case, the proposal is made by drawing the new model, $\mathbf{m}'$, as a random deviate from a probability distribution $q(\mathbf{m}' \mid \mathbf{m})$ conditional only on the current model $\mathbf{m}$. However, here the proposed model, $\mathbf{m}'$, may be a vector of different length than the current model $\mathbf{m}$, corresponding to partition models with differing number of Voronoi cells.

Once a proposed model has been drawn from the distribution $q(\mathbf{m}' \mid \mathbf{m})$, the new model is then accepted with a probability $\alpha(\mathbf{m}' \mid \mathbf{m})$. It can be shown (Green, 1995, 2003) that the chain of sampled models will converge to the transdimensional posterior distribution, $p(\mathbf{m} \mid \mathbf{d}_{obs})$, if

$$\alpha(\mathbf{m}' \mid \mathbf{m}) = \min\Big[1,\ \textbf{prior ratio} \ \times\ \textbf{likelihood ratio} \ \times\ \textbf{proposal ratio} \ \times |\mathbf{J}|\Big] \tag{3.11}$$

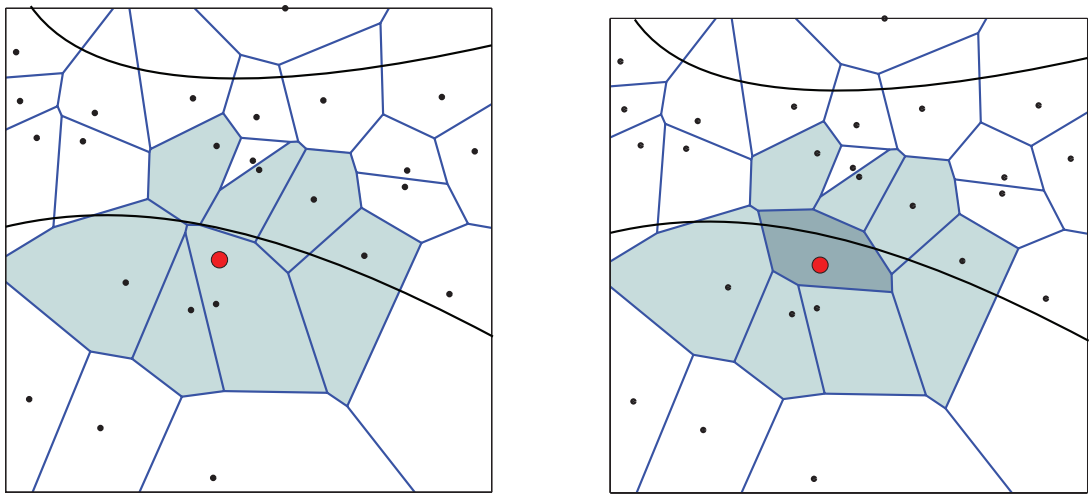
$$= \min\left[1, \frac{p(\mathbf{m}')}{p(\mathbf{m})} \times \frac{p(\mathbf{d}_{obs} \mid \mathbf{m}')}{p(\mathbf{d}_{obs} \mid \mathbf{m})} \times \frac{q(\mathbf{m} \mid \mathbf{m}')}{q(\mathbf{m}' \mid \mathbf{m})} \times |\mathbf{J}|\right] \tag{3.12}$$

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 'normalises' the difference in volume between two spaces of different dimension. In section 3.1.5, 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.* (2002) and Green (2003). The likelihood function $p(\mathbf{d}_{obs}|\mathbf{m})$ is not changed from the fixed dimension algorithm and is given by (2.5). The difference is that here, travel times are computed by integration along bend rays calculated in the reference model.

The expression for $\alpha(\mathbf{m}' \mid \mathbf{m})$ 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, $q(\mathbf{m} \mid \mathbf{m}')$, to the forward step, $q(\mathbf{m}' \mid \mathbf{m})$. 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 (3.12). 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, $p(\mathbf{m} \mid \mathbf{d})$. 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. Of course, the sampling distributions need to be chosen sensibly. If we always choose the proposed model to have velocity equal to the previous model then we could never converge to the posterior solution. In the next section we describe the proposal distributions used for each type of variable in the tomographic problem.

3.1.4 Proposal distributions

Here we use four different types of perturbation to the model $\mathbf{m}$. One is a 'birth' step which adds a Voronoi cell to the existing parameterisation. Another is a 'death'



(a) (b)

Figure 3.4: 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 neighboring 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, i.e. (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 travel times of rays passing through these cells need updating.

step which removes one of the Voronoi cells. The third is a 'move' step which 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 parameterisation and one changes the velocity values. Together they form a randomised perturbation which is able to generate a wide range of velocity models from few to many degrees of freedom with multiple spatial scalelengths (See Figure 3.4).

3.1.4.1 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:

- At every even step of the chain : randomly pick one velocity parameter, say v_i

and perturb its value using a Gaussian probability density $q_{v1}(v_i' \mid v_i)$

$$q_{v1}(v_i' \mid v_i) = \frac{1}{\sigma_1\sqrt{2\pi}} \exp\left\{-\frac{(v_i' - v_i)^2}{2\sigma_1^2}\right\}. \tag{3.13}$$

Hence we have

$$v_i' = v_i + u \times \sigma_1 \tag{3.14}$$

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.

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

  1. 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 parameterisation (See Figure 3.4). A velocity value v_{n+1}' needs to be assigned to the new cell. This is drawn from Gaussian proposal probability density $q_{v2}(v_{n+1}' \mid v_i)$ with the same form as (3.13), 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.
  2. 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 Figure 3.4).
  3. MOVE : randomly pick a cell and perturb the position of its nucleus $\mathbf{c}_i$ according to a 2D Gaussian proposal probability density $q_c(\mathbf{c}_i' \mid \mathbf{c}_i)$ centred on the current position $\mathbf{c}_i$.

     $$q_c(\mathbf{c}_i' \mid \mathbf{c}_i) = \frac{1}{2\pi\sigma_c^2} \exp\left(-\frac{1}{2\sigma_c^2}(\mathbf{c}_i' - \mathbf{c}_i)^T(\mathbf{c}_i' - \mathbf{c}_i)\right), \tag{3.15}$$

     The covariance matrix for the 2D 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.* (1995).

3.1.4.2 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 (3.12) 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

$$\begin{aligned} q_{v1}(v_i' \mid v_i) &= q_{v1}(v_i \mid v_i') \\ q_c(\mathbf{c}_i' \mid \mathbf{c}_i) &= q_c(\mathbf{c}_i \mid \mathbf{c}_i') \end{aligned} \tag{3.16}$$

and so in both cases the proposal ratio is one

$$\frac{q(\mathbf{m} \mid \mathbf{m}')}{q(\mathbf{m}' \mid \mathbf{m})} = 1, \tag{3.17}$$

which simplifies the acceptance probability expression (3.12). 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 (3.12). 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

$$\frac{q(\mathbf{m} \mid \mathbf{m}')}{q(\mathbf{m}' \mid \mathbf{m})} = \frac{q(\mathbf{c} \mid \ \mathbf{m}')}{q(\mathbf{c}' \mid \mathbf{m})} . \frac{q(\mathbf{v} \mid \mathbf{m}')}{q(\mathbf{v}' \mid \mathbf{m})}. \tag{3.18}$$

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

$$q(\mathbf{c}' \mid \mathbf{m}) = 1/(N-n), \tag{3.19}$$

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

$$q(\mathbf{v}' \mid \mathbf{m}) = \frac{1}{\sigma_2\sqrt{2\pi}} \exp\left\{-\frac{(v'_{n+1} - v_i)^2}{2\sigma_2^2}\right\}, \tag{3.20}$$

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

$$q(\mathbf{c} \mid \mathbf{m}') = 1/(n+1) \tag{3.21}$$

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

$$q(\mathbf{v} \mid \mathbf{m}') = 1. \tag{3.22}$$

Substituting these expressions in (3.18) we obtain

$$\left(\frac{q(\mathbf{m} \mid \mathbf{m}')}{q(\mathbf{m}' \mid \mathbf{m})}\right)_{birth} = \frac{\sqrt{2\pi}(N-n)}{(n+1)}\sigma_2 \exp\left\{\frac{(v'_{n+1} - v_i)^2}{2\sigma_2^2}\right\}. \tag{3.23}$$

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 parameterisation. This exponentially increasing probability density is ultimately restrained when combined with the likelihood ratio term in (3.12) which would penalise 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 (Figure 3.4). 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

$$\left(\frac{q(\mathbf{m} \mid \mathbf{m}')}{q(\mathbf{m}' \mid \mathbf{m})}\right)_{death} = \frac{n}{\sigma_2\sqrt{2\pi}(N-n+1)} \exp\left\{-\frac{(v'_j - v_i)^2}{2\sigma_2^2}\right\} \tag{3.24}$$

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 1) generate new samples easily and 2) determine the proposal ratio for insertion into the acceptance probability expression (3.12). To complete the evaluation of the acceptance probability $\alpha(\mathbf{m}' \mid \mathbf{m})$ we need the Likelihood ratio (2.5) and prior ratio (3.10).The likelihood evaluation involves calculation of travel times in the proposed model $\mathbf{m}'$. In the

linearised formulation used here, we simply integrate along rays calculated in the current reference model although in a fully nonlinear 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 $(\sigma_1, \sigma_2, \sigma_c)$. 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 $\alpha(\mathbf{m}' \mid \mathbf{m}) = 1$ 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; Green, 2003; Brooks *et al.*, 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 and 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 one cell 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 and 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 toward the posterior distribution is preserved. Details are given in section 3.2.1. In this work, Delayed Rejection was only used for moves that do not jump between dimensions although Green and 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.

3.1.5 The Jacobian

By definition the Jacobian $|\mathbf{J}|$ in (3.12) only needs to be calculated when there is a jump between two models of different dimensions, i.e. 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

$$(\mathbf{c}, \mathbf{v}, \mathbf{u}_c, \mathbf{u}_v) \longleftrightarrow (\mathbf{c}, \mathbf{v}, \mathbf{c}'_{n+1}, v'_{n+1}) = \mathbf{m}'. \tag{3.25}$$

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 $[0, 1, ..., N-n]$. 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

$$v'_{n+1} = v_i + \mathbf{u}_v \tag{3.26}$$

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.*, 2002) showed that the Jacobian term is always unity for discrete transformations. Therefore, the Jacobian term only accounts for the change in variables from

$$(\mathbf{v}, \mathbf{u}_v) \longleftrightarrow (\mathbf{v}, v'_{n+1}) = \mathbf{v}'. \tag{3.27}$$

Hence, we have

$$|\mathbf{J}|_{birth} = \left|\frac{\delta(\mathbf{v}')}{\delta(\mathbf{v}, \mathbf{u}_v)}\right| = \left|\frac{\delta(\mathbf{v}, v'_{n+1})}{\delta(\mathbf{v}, \mathbf{u}_v)}\right| = \left|\frac{\delta(v_i, v'_{n+1})}{\delta(v_i, \mathbf{u}_v)}\right| = \begin{vmatrix} 1 & 0 \\ 1 & 1 \end{vmatrix} = 1 \tag{3.28}$$

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

3.1.6 The acceptance probability

To complete our description of the algorithm, we now substitute expressions for each proposal ratio into (3.12) 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

$$\alpha(\mathbf{m}' \mid \mathbf{m}) = \min\left[1, \frac{p(\mathbf{m}')}{p(\mathbf{m})}.\frac{p(\mathbf{d}_{obs} \mid \mathbf{m}')}{p(\mathbf{d}_{obs} \mid \mathbf{m})}\right]. \tag{3.29}$$

Since the dimension of the model does not change, according to (3.10), the prior ratio is either null or unity and we have

$$\alpha(\mathbf{m}', \mathbf{m}) = \begin{cases} \min\left[1, \frac{p(\mathbf{d}_{obs}|\mathbf{m}')}{p(\mathbf{d}_{obs}|\mathbf{m})}\right] & \text{if} \quad \forall i \in [1, n],\ v_i \in J) \\ 0 & \text{otherwise} \end{cases} \tag{3.30}$$

For both the move and velocity update steps, this only requires the ratio of the Likelihoods and hence calculation of travel times 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 $\alpha_2(\mathbf{m}'' \mid \mathbf{m})$ is more complicated to determine, an expression is given in section 3.2.1. Note that by reducing the variance of the second proposal, we attempt a less ambitious move which 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 (3.10), the prior ratio takes the form

$$\left(\frac{p(\mathbf{m}')}{p(\mathbf{m})}\right)_{birth} = \begin{cases} \left[\frac{(n+1)!(N-n-1)!}{N!(\Delta v)^{n+1}\Delta n}\right]\left[\frac{n!(N-n)!}{N!(\Delta v)^{n}\Delta n}\right]^{-1} & \text{if } \left((n+1)\ \epsilon\ I \text{ and } v'_{n+1}\ \epsilon\ J\right) \\ 0 & \text{otherwise} \end{cases} \tag{3.31}$$

$$\left(\frac{p(\mathbf{m}')}{p(\mathbf{m})}\right)_{birth} = \begin{cases} \frac{n+1}{(N-n)\Delta v} & \text{if} \quad \left((n+1)\ \epsilon\ I \quad \text{and} \quad v'_{n+1}\ \epsilon\ J\right) \\ 0 & \text{otherwise.} \end{cases} \tag{3.32}$$

After substituting (2.5), (3.23), and (3.32) into (3.12), the acceptance term reduces to

$$\alpha(\mathbf{m}', \mathbf{m}) = \begin{cases} \min\left[1, \frac{\sigma_2\sqrt{2\pi}}{\Delta v}.\exp\left\{\frac{(v'_{n+1}-v_i)^2}{2\sigma_2^2} - \frac{\phi(\mathbf{m}')-\phi(\mathbf{m})}{2}\right\}\right] & \text{if } \left((n+1)\ \epsilon\ I \text{ and } v'_{n+1}\ \epsilon\ J\right) \\ 0 & \text{otherwise} \end{cases} \tag{3.33}$$

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 which penalises velocities if they change so much that they degrade fit to data.

For the death step, the prior ratio in (3.32) must be inverted. After substituting this with (2.5) and (3.24) into (3.12), and after simplification we get the acceptance probability

$$\alpha(\mathbf{m}', \mathbf{m}) = \begin{cases} \min\left[1, \frac{\Delta v}{\sigma_2\sqrt{2\pi}}.\exp\left\{-\frac{(v'_j-v_i)^2}{2\sigma_2^2} - \frac{(\phi(\mathbf{m}')-\phi(\mathbf{m}))}{2}\right\}\right] & \text{if} \quad (n-1)\ \epsilon\ I \\ 0 & \text{otherwise} \end{cases} \tag{3.34}$$

where i indicates the cell that we remove from the current tessellation $\mathbf{c}$ and j indicates the cell in the proposed tessellation $\mathbf{c}'$ 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, i.e. 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 which 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 model 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 (3.5). (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.)

3.1.7 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 linearised scheme described by Figure 3.2, 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 parameterisation (an example is seen in section 3.3).

Instead, and as done in the fixed dimension case, we look at the spatially averaged model defined by taking the mean of the distribution of velocity values at each point across the 2D 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 visualising the reference model. At each iteration of the outer loop in Figure (3.2) 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, i.e. by calculating the standard deviation of velocities as a function of position. In this way a large number of models with different parameterisations 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.

3.1.8 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, i.e. 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 and 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 and Carlin, 1996; Gelman and 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 ten 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 work we assess convergence (or non convergence) without using parameters (**c,v**) defining the partitioned models. Instead, we look for convergence in terms of numbers like the velocity value at a given point in the 2D field or the model dimension. In the case of a parallelised algorithm, we shall show that the average dimension of the model space over the ensemble of independent chains plotted as a function of iterations is a good tool to diagnostic convergence (see Figure 3.6).

3.2 Optimising the algorithm

Markov chains are conceptually simple, easy to implement and often appear to be ideally suited for Bayesian analysis. However, some major issues and drawbacks can make them impractical. A first weakness of McMC schemes is that they rapidly can become computationally expensive. For example, in our tomography problem, as a large amount of data are used, the number of cells needed to describe the velocity field with adequate resolution increases. And if the model is defined by too many parameters, 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. This is a reason why Monte Carlo methods have not been habitually used in tomographic imaging.

Another issue is that the samples produced by a Markov chain are correlated. Each model generated in the chain is 'close' to the previous model. Hence, a sequence of samples produced by a Markov chain, if too small, will only describe the posterior distribution in a given region of the model space. The size of cells becomes smaller as we increase the dimension. Changing the velocity value in one cell in a 3-cell model represents changing the velocity value of approximatively one third of the velocity field. This is a much larger step compared to changing the wavespeed in one cell belonging to a 100-cell models. Therefore as we increase the model dimension the sizes of 'steps' in the random walk become smaller and sampled models become more correlated. When a large number of data are used, the random walk advances very slowly around the model space, a situation known as 'poor mixing'. When this is the case, the sequence of samples needed to properly describe the posterior distribution becomes very long, and the algorithm runs for a longer time.

The reversible jump tomography algorithm encounters the two problems aforementioned, and hence is much more computationally expensive than standard tomography optimisation schemes. 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 travel times are recomputed in the evaluation of (3.3). For example, Figure 3.4 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 travel times 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. Two other features used to improve the sampling efficiency are the use of a delayed rejection scheme (Tierney and Mira, 1999) as well as parallelising the algorithm. These are detailed in the next subsections.

3.2.1 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 and Mira, 2001). Tierney and 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 1D target distribution $\pi(x)$ shown in Figure 3.5. 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 and Mira, 2001).

When at x, we propose a new state y_1 with density $q_1(y_1 \mid x)$. As in Hastings

(1970), this is accepted with probability

$$\alpha_1(y_1 \mid x) = \min\left[1, \frac{\pi(y_1)}{\pi(x)}\frac{q_1(x \mid y_1)}{q_1(y_1 \mid x)}\right]. \tag{3.35}$$

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

$$\alpha_2(y_2 \mid x, y_1) = \min\left[1, \frac{\pi(y_2)}{\pi(x)}\frac{q_1(y_1 \mid y_2)}{q_1(y_1 \mid x)}\frac{q_2(x \mid y_2, y_1)}{q_2(y_2 \mid x, y_1)}\frac{\{1-\alpha_1(y_1 \mid y_2)\}}{\{1-\alpha_1(y_1 \mid x)\}}\right] \tag{3.36}$$

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 i-th stage, the acceptance probability at that stage is , following Mira (2001),

$$\alpha_i(y_i \mid x, y_1, \ldots, y_{i-1}) = \min\left[1, \frac{\pi(y_i)}{\pi(x)}\frac{q_1(y_{i-1} \mid y_i)q_2(y_{i-2} \mid y_i, y_{i-1})\ldots q_i(x \mid y_i, y_{i-1}, \ldots y_1)}{q_1(y_1 \mid x)q_2(y_2 \mid x, y_1)\ldots q_i(y_i \mid x, y_1, \ldots, y_i)}\right.$$
$$\left.\frac{\{1-\alpha_1(y_{i-1} \mid y_i)\}\{1-\alpha_2(y_{i-2} \mid y_i, y_{i-1})\}\ldots\{1-\alpha_{i-1}(y_1 \mid y_i, y_{i-1}, \ldots, y_2)\}}{\{1-\alpha_1(y_1 \mid x)\}\{1-\alpha_2(y_2 \mid x, y_1)\}\ldots\{1-\alpha_{i-1}(y_{i-1} \mid x, y_1, \ldots, y_{i-2})\}}\right]. \tag{3.37}$$

The implementation of a general delayed rejection algorithm using (3.37) 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 $q_1(\mathbf{m}' \mid \mathbf{m})$ with large variance. The proposed model is accepted with probability

$$\alpha_1(\mathbf{m}' \mid \mathbf{m}) = \min\left[1, \frac{p(\mathbf{m}' \mid \mathbf{d}_{obs})}{p(\mathbf{m} \mid \mathbf{d}_{obs})}\right]. \tag{3.38}$$

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 $q_2(\mathbf{m}'' \mid \mathbf{m})$ 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

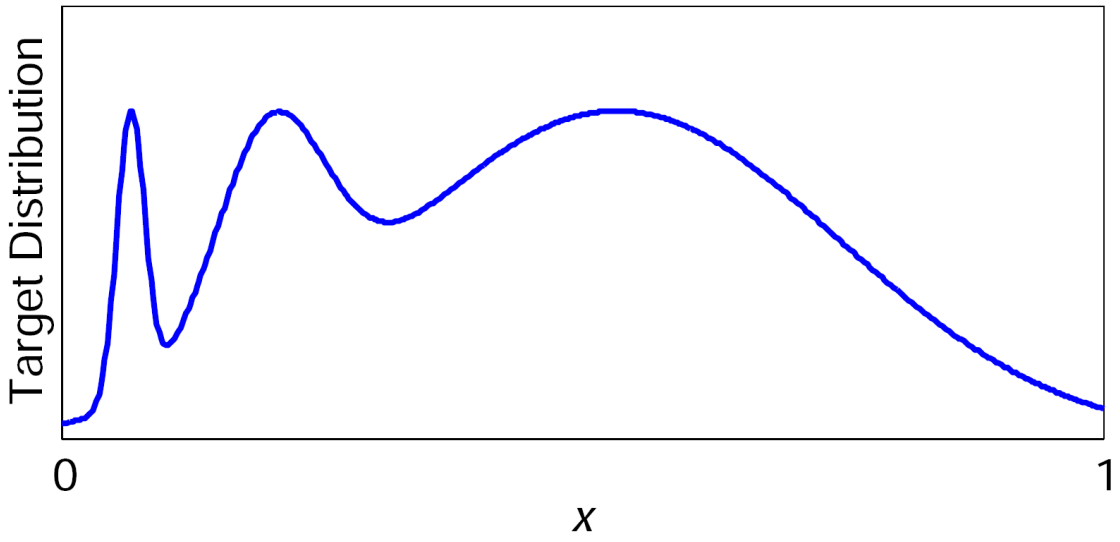


Figure 3.5: An example of a 1D probability density function with variable 'length scale'.

second try in (3.36) is simplified

$$\alpha_2(\mathbf{m}'' \mid \mathbf{m}) = \min\left[1, \frac{p(\mathbf{m}'' \mid \mathbf{d}_{obs})}{p(\mathbf{m} \mid \mathbf{d}_{obs})} \frac{q_1(\mathbf{m}' \mid \mathbf{m}'')}{q_1(\mathbf{m}' \mid \mathbf{m})} \frac{\{1 - \alpha_1(\mathbf{m}' \mid \mathbf{m}'')\}}{\{1 - \alpha_1(\mathbf{m}' \mid \mathbf{m})\}}\right]. \tag{3.39}$$

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.

3.2.2 Parallelisation of the algorithm

Another computational consideration is that the algorithm is straightforward to parallelise 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). Figure 3.6 shows an example of 5 chains independently sampling the same transdimensional distribution. The number of Voronoi cells for each chain is plotted against iterations. Each chain starts from a different state (with a number of cells drawn randomly from the uniform prior between 2 and 100), and after 10^5 iterations, one

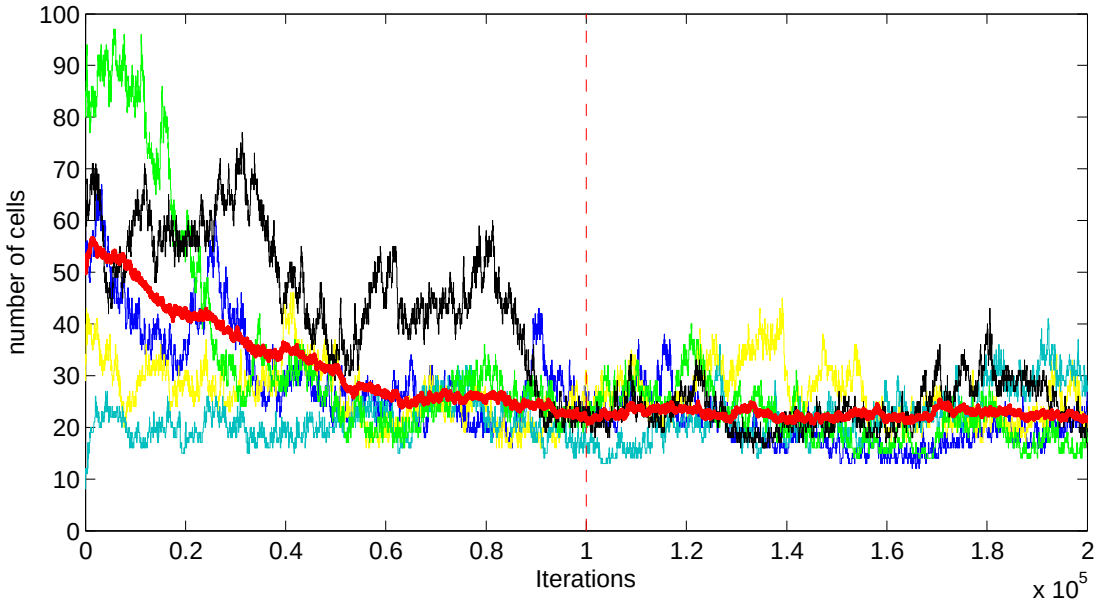


Figure 3.6: Convergence of chains. The number of cells is plotted against iterations for 5 chains running in parallel. For each chain, the initial number of cells is drawn randomly form the prior distribution $n = [2, 100]$. The red line is the average over all the parallel chains, and it is used as a measure of convergence to determine the length of the burn-in period (red dashed line).

can see that they have all converged towards the target distribution and only sample models having around 22 cells. The curve in red shows the average over 100 independent chains, and is a good measure of convergence that can be used to determine the length of the burn-in period (red dashed line).

Our Monte Carlo algorithm proceeds by averaging large numbers of computed values. It is therefore straightforward to have different processors compute different values, and then use a simple average of these values to produce a final answer (see Bradford and Thomas, 1996, for an example of parallel McMC). At each point of the velocity field, we wish to compute the unknown seismic wave velocity value V. We have a computer program which is capable of producing some large number n of samples, $v_1, ..., v_n$, and then to estimate V by the estimator

$$E = \frac{1}{n} \sum_{i=1}^{n} v_i. \tag{3.40}$$

This estimate has a variance proportional to $1/n$ (Press *et al.*, 1992). Now let us suppose we have N computers available to us. We wish to use all of these computers to better estimate V. The simplest idea is to run our same program on each of the

N computers (with the seed of the random number generator different on each computer). Each independent chain j then produces n samples $v_1^j, ..., v_n^j$, computes their average E^j, and reports the result back to a master program. The master program then averages these N results to obtain a master result

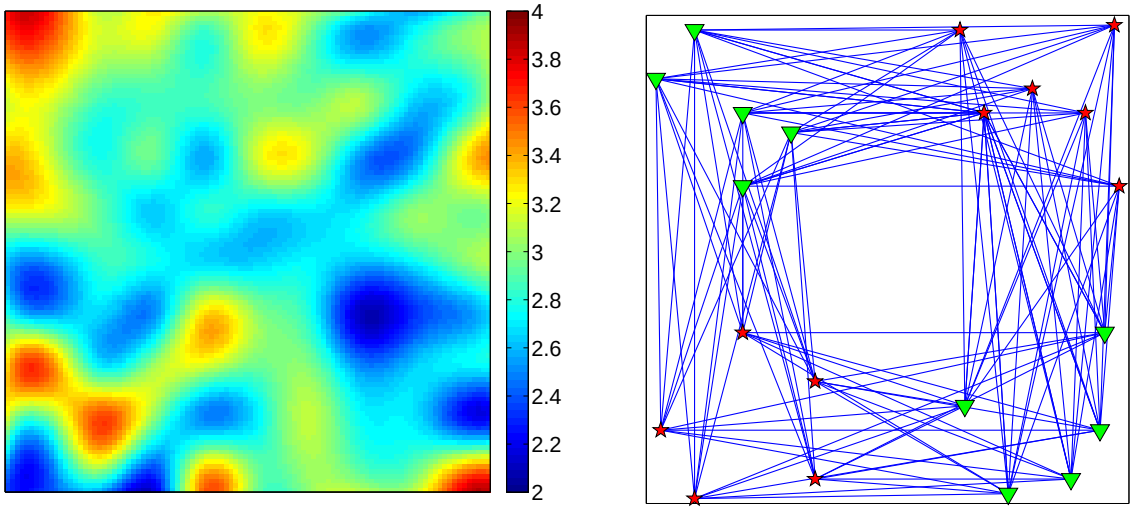
$$\bar{E} = \frac{1}{N} \sum_{j=1}^{N} E^j. \tag{3.41}$$

Therefore, the variance of the estimate is reduced by a factor of N. In this way, the communication between processors is minimised, so that parallel processing is easily facilitated. This kind of Monte Carlo algorithm is so ideally suited to parallel computation that it would be labeled 'embarrassingly parallelisable' by computer scientists (Rosenthal, 2000). Clearly, once a burn-in period has been walked, the algorithm will go 10 times faster if it is run on 10 processors that if it is run in a single processor.

It would appear that, in this way, we have obtained linear speed-up, that is the program runs N time faster as on a single computer. Unfortunately, this is not rigorously the case. Within each independent chain, an initial given number of steps b needs to be discarded before samples start to be collected. It is only after this burn-in period that the McMC procedure asymptotically converges to an ensemble of models whose density is proportional to the posterior distribution. Hence, if n now represents the number of post burn-in samples needed to describe the posterior, each chain of the parallel program need to be run for $b + n/N$ steps. The 'speed-up', i.e. the ratio of time taken by a sequential algorithm to the time taken by a parallelised algorithm to collect n post burn-in samples, is therefore given by

$$\text{Speedup}(N) = \frac{b + n}{b + n/N} \tag{3.42}$$

To illustrate the benefits of parallelisation, we compare the sequential and parallel algorithms on a simple experiment (here the rays are considered straight). Figure 3.7 shows the true velocity model and the geometry of rays that have been used. The sources and receivers are clustered in regions that could represent seismogenic areas and arrays of seismic stations. As a result, the ray coverage is irregular and rays often lay in the same direction. Those are features often encountered in seismic tomography.



(a) (b)

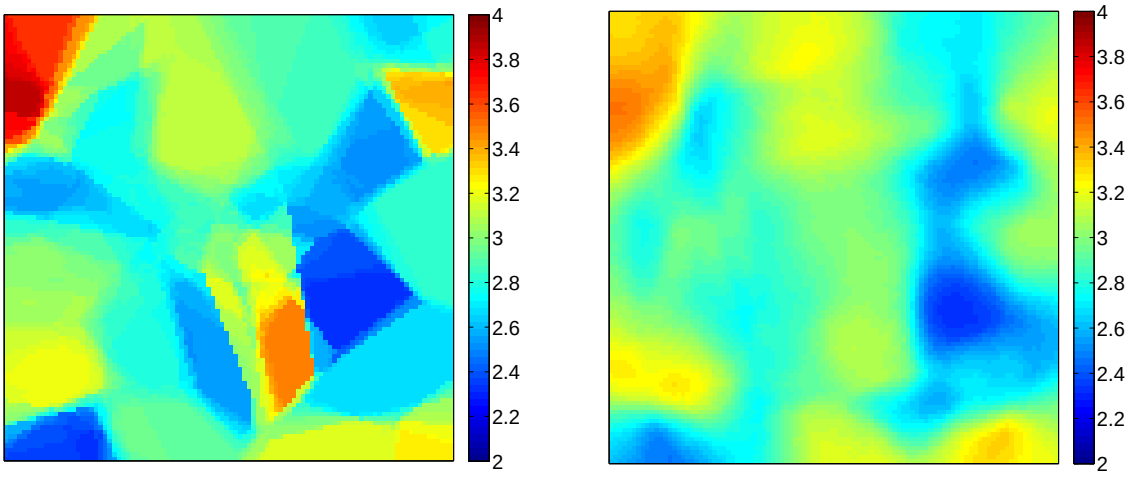
Figure 3.7: Left: True velocity field (km/s). Right: Geometry of rays. 110 rays join 11 sources (red stars) to 10 receivers (green triangles).

3.2.2.1 Advantage in computational time

Let us suppose that we only have 1000 s (approximatively 15 minutes) of computational time available to solve the 2D problem described above. Figure 3.8 compares the results obtained with a single computer against a cluster of 256 computers. In both cases, a burn-in period was set to 50000 samples, after which the chains produced 5000 samples used for Bayesian inference. In 15 minutes, a single computer is only able to collect 5000 samples to estimate the velocity at each point (Figure 3.8(a)) whereas the parallel algorithm can generate $256 \times 5000 = 1.28.10^6$ post burn-in samples (note that if we had wanted to produce the same number of samples with a single computer, that would have taken around 6.5 hours). As can be seen in Figure 3.8(a), the sequential algorithm has not produced enough samples, and hence the average solution map still contains the Voronoi discontinuities present in individual models. Clearly, the velocity estimates given by the parallel tomography in Figure 3.8(b) are closer to the true velocities in Figure 3.7(a).

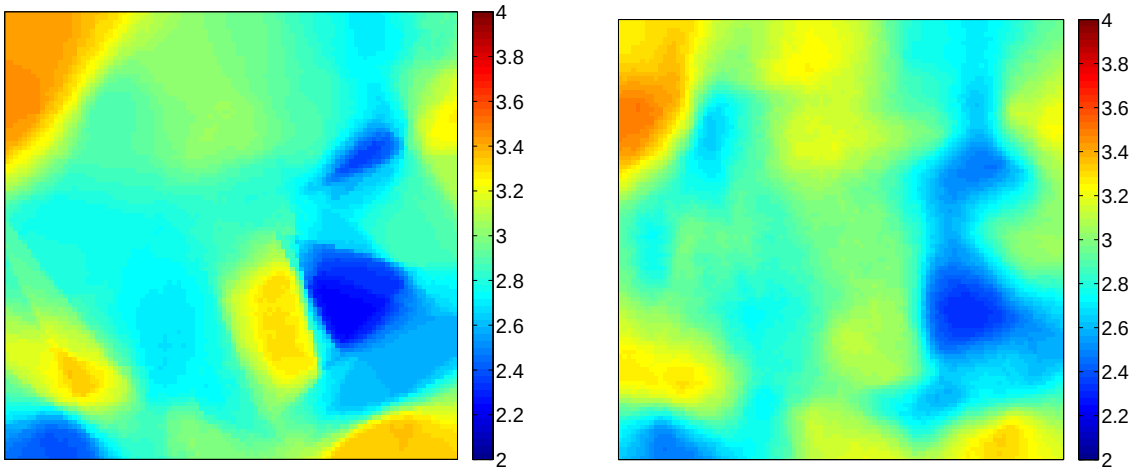
3.2.2.2 Advantage in performance

Another benefit of parallel Markov chains is that they naturally produce independent samples. Figure 3.9(a) shows the image obtained when running the parallel algorithm for $N = 376$ and $b = n = 50000$. The result is not equivalent to the image



(a) (b)

Figure 3.8: Average solution map obtained with a sequential (a) and parallel (b) algorithm for the same amount of computational time.



(a) (b)

Figure 3.9: Average solution map obtained with a sequential (a) and parallel (b) algorithm for the same number of post burn-in samples.

3.9(b) we obtain upon running our program on a single computer and collecting exactly the same number n of post burn-in samples. The samples produced by a Markov chain are correlated, i.e. they follow a path in the model space. Therefore, information on the target distribution is restricted to that path and regions of interest can be left out. This is why estimation of the posterior is improved if a large number of chains independently explore the model space. In simple words,

the map produced by an explorer allowed to walk for 100 km will be less accurate than the map produced by the common work of 10 men, each walking for 10 km, and starting at different locations. From a visual inspection, it is evident that the parallel algorithm gives more accurate and unbiased velocities 3.9(b) than the single processor scheme in 3.9(a)

Here, the number of post burn-in samples n needed to estimate the posterior is equal to the length b of the burn-in period. The estimation of 3.9(a) took about 15 minutes against 30 minutes for 3.9(b). Thus, the parallelisation upon 376 processors is not efficient in terms of computational time (according to (3.42), the speed-up is indeed bounded by 2). However, Figure 3.9 clearly shows that parallelisation is worthwhile in terms of performance of the algorithm.

3.2.3 Computational time

It is difficult to give a general idea of computational time for the reversible jump tomography. As showed in Figure 3.10, it basically depends on two factors (green boxes): the required length of the Markov chain and the time taken to solve the forward problem each time a model is tested. These two factors depend on 5 elements that are specific to each problem (blue boxes): the number of data, the level of data noise, the complexity of the true model, and the form of prior and proposal probability distributions. Note that the level of data noise given by the user prior to the inversion plays an important role in determining the number of model parameters. This is a major feature of the algorithm and will be studied in detail in next chapters. The prior distribution on the model defines the volume of the model space and hence the time needed by the chain to converge (i.e. the length of the burn-in period). For example, if the prior on the number of cells in Figure 3.6 would have been a uniform distribution over the range 15-25, the convergence would have taken less iterations, since the number of cells of initial models would have been closer to their expected posterior value.

To give some measure of computational cost, the synthetic examples presented in Section 3.3 can all be performed without parallelisation on a standard desktop workstation (Intel core 2 duo with CPU running at 2.1 GHz) using about 150 minutes CPU time. However, these are simple examples and as we invert for more travel times, the required number of cells increases and parallelisation becomes necessary. The ambient noise tomography performed in section 3.4 was carried out on 98 parallel processors and took around 6 hours.

The relation between the number of cells in a model and the computational cost

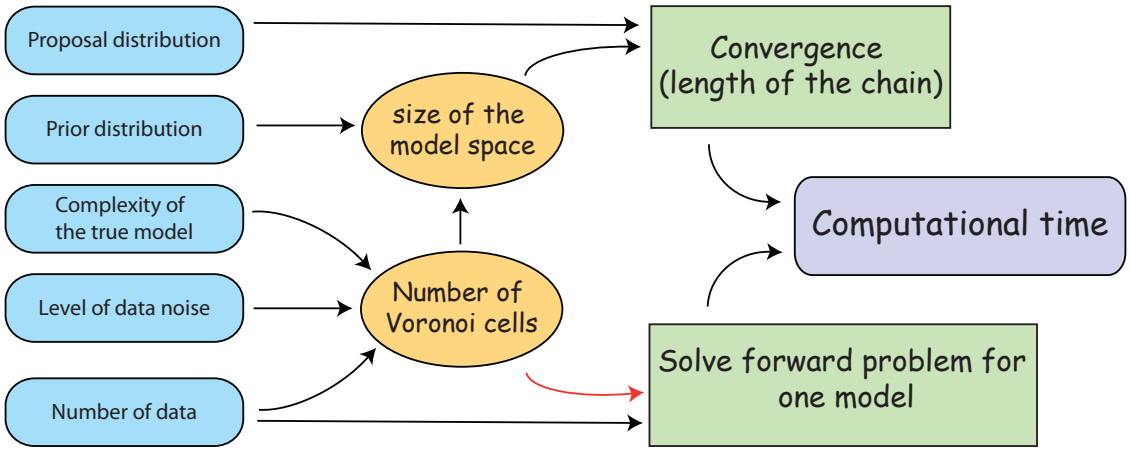


Figure 3.10: Scheme showing the elements determining the computational cost of the algorithm. The relation represented by the red arrow is examined in Figure 3.11.

of the forward calculation needed at each iteration of the inner-loop is shown in Figure 3.11. This corresponds to the arrow in red in Figure 3.10. For each rj-McMC inversion that has been carried out in this thesis, we estimate the average time taken to solve the forward model (the time to compute travel times for rays passing through the perturbed cells), and plot it against the average number of cells in the ensemble. One can clearly see that the computational cost increase as we add more unknowns in the problem.

3.3 Synthetic data examples

3.3.1 Experimental setup

A synthetic data set is constructed by using the Fast Marching Method (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 set up is similar to the one used in chapter 2 to illustrate the fixed dimension algorithm. It 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 Figure 3.12. The areas in red have a velocity of 5 km/s and the blue areas are of 4 km/s. The velocity field presents high contrast discontinuities. The blue heterogeneity represents a velocity anomaly of -20% of the red background and the red heterogeneity is +25% of the blue background. Hence this 'simple' problem is reasonably nonlinear and serves to illustrate

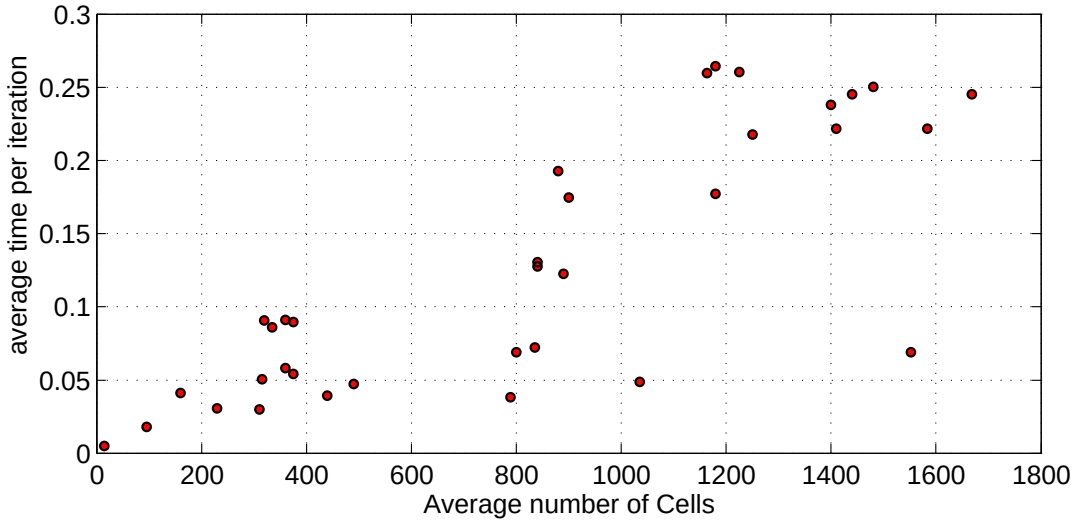


Figure 3.11: The speed of the forward calculation (i.e. overall time taken by the algorithm divided by the number of samples in each chain) is plotted against the average number of Voronoi cells in the ensemble. This corresponds to the red arrow in Figure 3.10. It is clear that the forward calculation takes longer as we add unknowns in the problem.

the algorithm.

The ray geometry associated with the synthetic velocity field is shown in Figure 3.13. All the calculations are performed in 2D 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

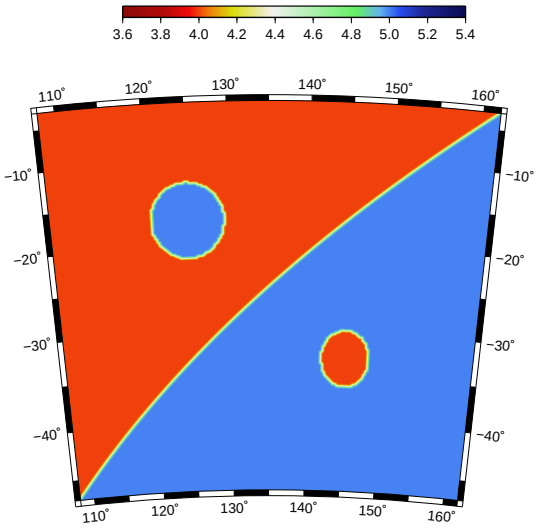


Figure 3.12: True velocity field (km/s).

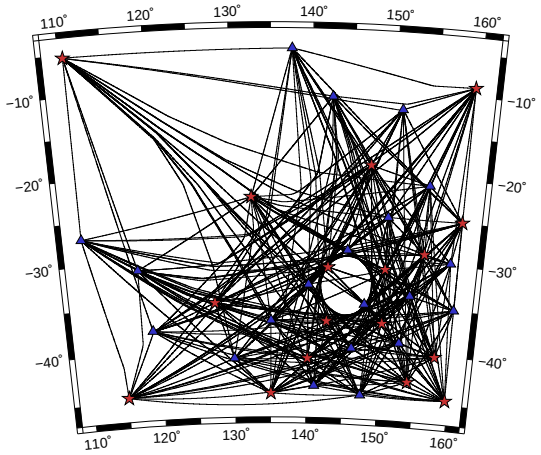


Figure 3.13: True ray paths. 340 rays join 17 sources (red stars) to 20 receivers (blue triangles). Due to the principle of least time, the rays seem to avoid the low velocity anomaly and to be attracted towards the high velocity anomaly.

regular parameterisation. Both approaches are linearised methods and use the same forward modelling to update the geometry of rays. The Subspace scheme uses a matrix inversion approach including implicit regularisation to solve the linearised 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 travel time is 473 s. For a homogeneous initial model with velocities equal to 4.5 km/s, without noise in the observed data, the average difference between observed and estimated travel times is about 35 s. Some random Gaussian noise has been added to the observed travel times with a standard deviation of 9.5 s (i.e. 2% of the average observed travel time).

3.3.2 Fixed parameterisation tomography with the Subspace method

3.3.2.1 The regularisation process

Most of the methods using a predefined fixed parameterisation 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. Regularisation procedures must be used to choose a solution among all the acceptable possibilities. This resulting solution will have properties reflecting the particular choice of regularisation.

In order to discard the models that are unrealistic and make the solution unique, criteria other than the misfit can be minimised 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 in minimising an objective function which is a linear combination of different criteria

$$\Phi(\mathbf{m}) = \left\| \frac{\mathbf{Gm} - \mathbf{d}}{\sigma_d} \right\|^2 + \varepsilon \|\mathbf{m} - \mathbf{m}_0\|^2 + \eta \|\mathbf{Dm}\|^2 \tag{3.43}$$

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 minimising the semi-norm $\|\mathbf{Dm}\|^2$, the regularisation techniques favour models that are relatively flat (first order regularisation) or smooth (second order regularisation). 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.

3.3.2.2 Fixed parameterisation and B-spline interpolation

Here, the velocity field is defined by a uniform grid of nodes with bi-cubic B-spline interpolation (Virieux and 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, i.e. 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 parameterising the velocity field is common in surface wave tomographic studies (e.g. Yoshizawa, K. and Kennett, B. L. N, 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 2D interpolation.

3.3.2.3 The Subspace method

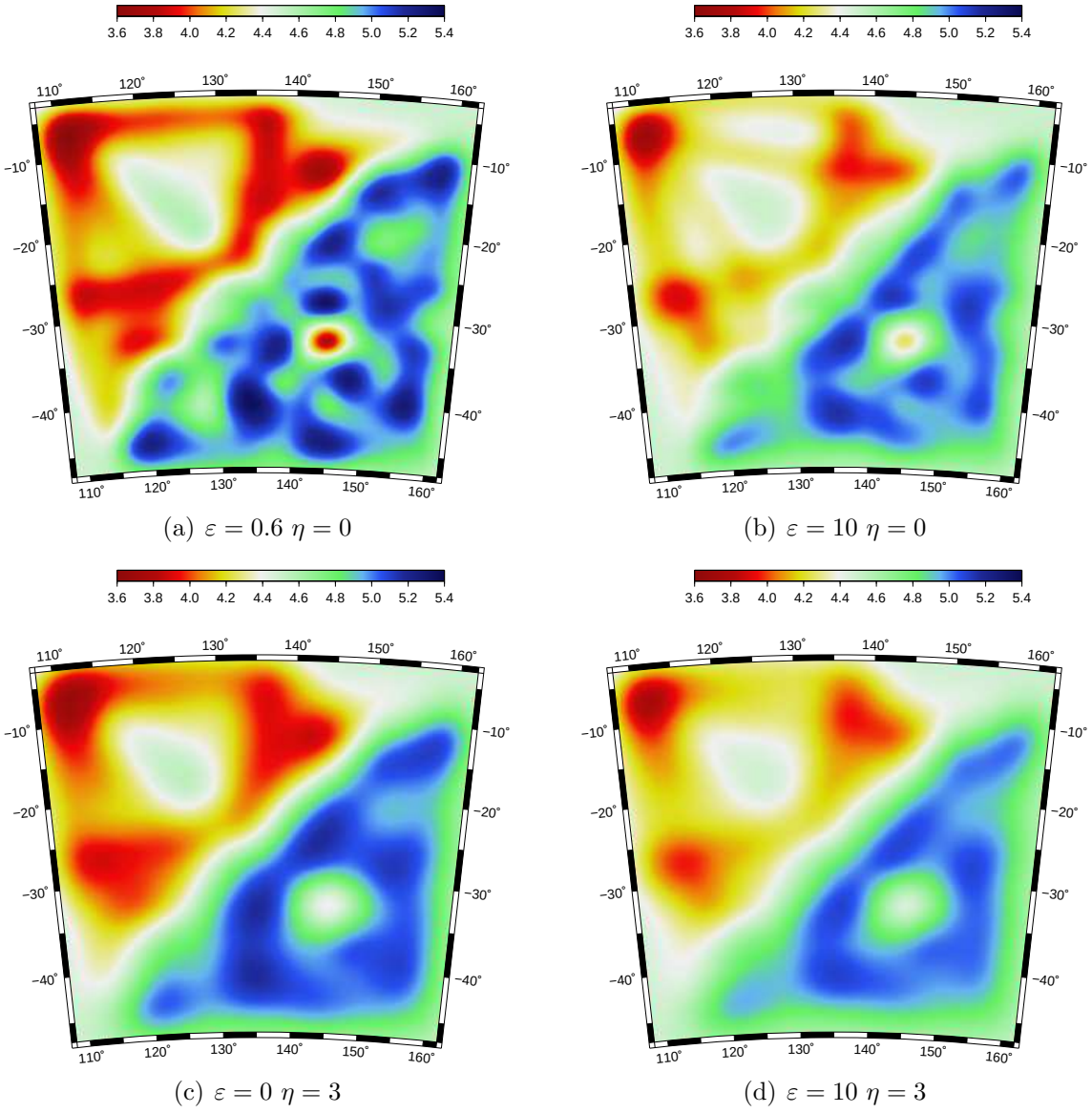
Here, the unknown which is sought for during the inversion step is a perturbation of the reference field (which causes only a perturbation to the geometry of rays). The problem is then locally linearised around the reference model. The perturbed solution becomes the reference model for the next iteration. The process is stopped when, for example, the data are satisfied, that is when the normalised χ^2 misfit measure (corresponding to the first term in (3.43) 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 linearised 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 and Sambridge (2003) and Rawlinson *et al.* (2006). The advantage of this approach is that the optimisation of (3.43) can proceed with only the inversion of an $m \times m$ matrix at each iteration. The set of vectors which 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 orthogonalise subspace vectors, and remove those directions which 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.

3.3.2.4 Results

Figure 3.14 shows the results obtained after 6 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 regularisation parameters, we observed the classic trade-off between smooth models with poor spatial resolution (3.14(d)) and instability (3.14(a)) (Menke, 1989). The solution shown in 3.14(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 3.14(b) is damped and 3.14(c) is a smoothed solution. The solution model in 3.14(d) has been produced with relatively large values for both regularisation 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



(a) $\varepsilon = 0.6$ $\eta = 0$

(b) $\varepsilon = 10$ $\eta = 0$

(c) $\varepsilon = 0$ $\eta = 3$

(d) $\varepsilon = 10$ $\eta = 3$

Figure 3.14: Subspace Inversion with a regular grid (20 × 20 nodes) and B-spline interpolation. Results after 6 iterations for different values of damping and smoothing (km/s). A random gaussian noise has been added to the data with a standard deviation equal to 2% of the average observed travel time.The colour scales are the same as for the true model.

to map the heterogeneities. Figure 3.14 represents the best results we have been able to obtain with the regularisation framework. Of course, automated procedures methods also exist for selecting optimal values of the regularisation parameters like 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 regularisation 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 3.14(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.

3.3.3 Reversible jump tomography

3.3.3.1 The average model: a naturally smooth solution

Posterior inference was made using an ensemble of 5000 models. We ran the rj-McMC algorithm for 560000 steps in total. The first 60000 steps were discarded as burn-in steps, only after which the sampling algorithm was judged to have converged. Then, every 100^{th} model visited in the last 500000 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 (Figure 3.2) 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 Figure 3.15(a). It would appear to be a rather poor recovery of the true model in Figure 3.12. However, the spatial average of the post burn-in samples collected shown in Figure 3.15(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 Figure 3.15(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, i.e. without the need to impose an explicit smoothing function, choose regularisation parameters or interpolation procedure. Average velocity maps like Figure 3.15(b) are in a sense self-regularised solutions.

The average solution is clearly quite different from the models obtained with a fixed grid. The artifacts of the later are not present and the discontinuities have been recovered with better accuracy. Furthermore, the ficticious gradients which are evident in Figure 3.14 are removed in Figure 3.15(b) giving a more faithful recovery of the true model in Figure 3.12. The normalised χ^2 misfit measure for the average solution model is 0.86 which is of the same order as for the solution in 3.14(c) (0.92).

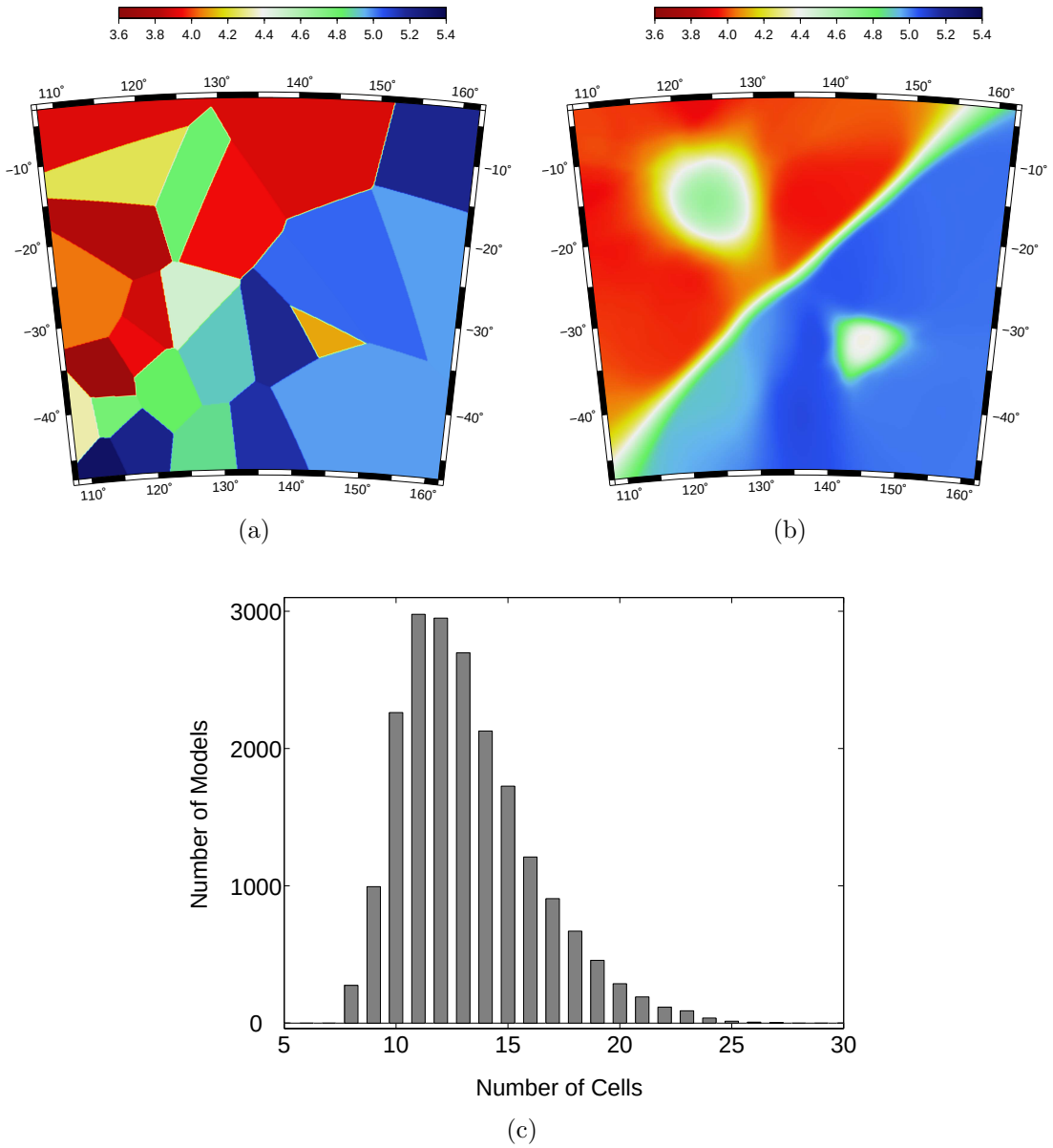


Figure 3.15: Reversible jump tomography. Results after 4 iterations (km/s). Left: best model sampled (i.e posterior maximum). Middle: Average solution map. The scales are the same as for previous figures. Right : Information on the number of cells in the Voronoi tessellation. Posterior probability density for the parameter n, $p(n, \mathbf{d}_{obs})$

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. Figure 3.15(c) shows $p(n, \mathbf{d}_{obs})$, i.e 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 suggest 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 wavespeed, 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 $p(n, \mathbf{d}_{obs})$ 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).

Figure 3.15(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 which results, as expected, from averaging many overlapping Voronoi cells in different configurations. The solution in the regularisation 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 parameterisation looks to have adapted to the structural features of the underlying model.

3.3.3.2 The variance map: an estimate of model uncertainty

A well known problem with regularised inversion algorithms which construct a single optimal model is that it is impossible to assess the distance between the estimated and true model. Regularisation or damping helps to stabilise 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. 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

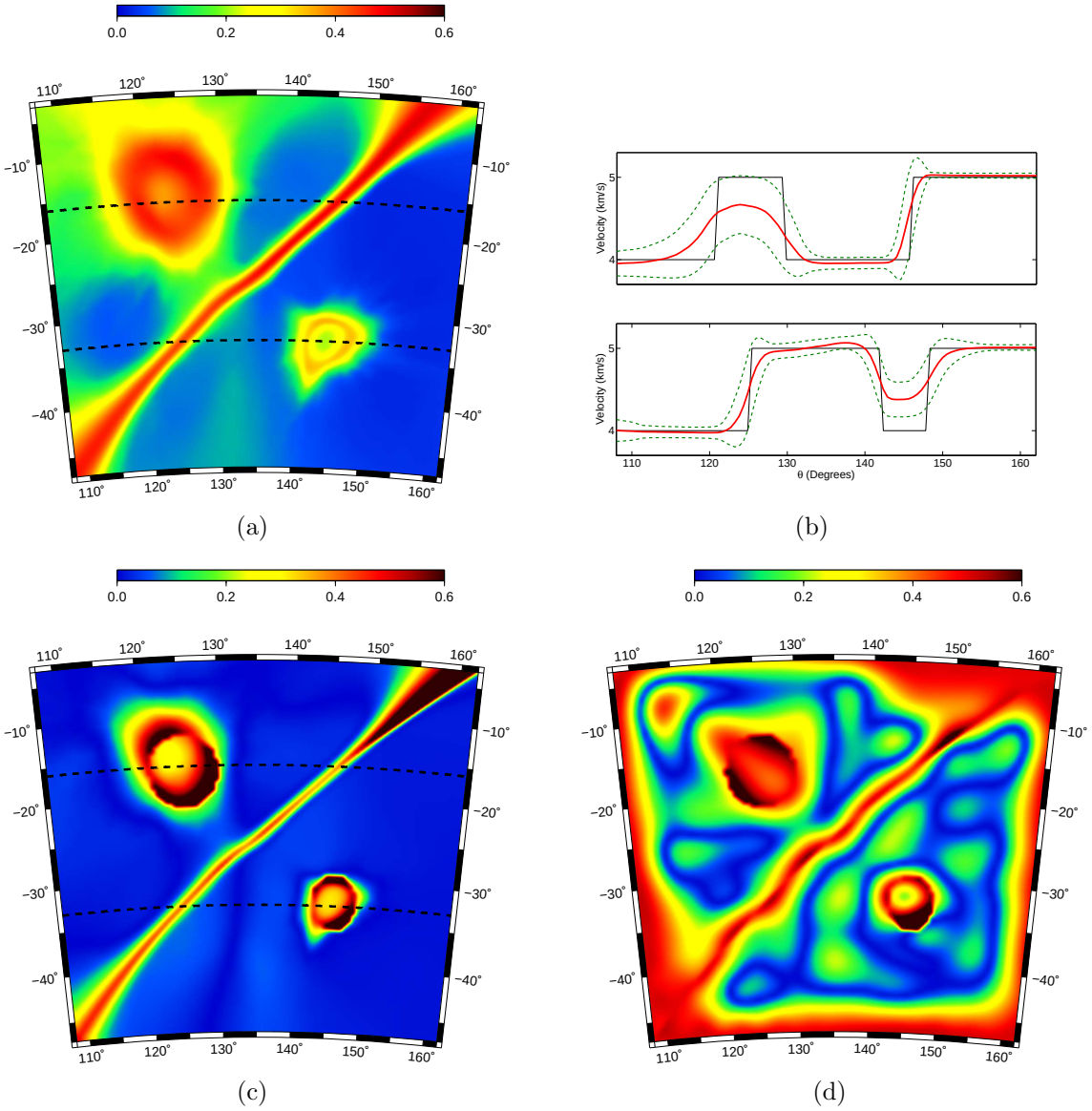


Figure 3.16: Top left: estimated error map (km/s). The two dashed lines show the cross sections presented on 3.16(b). Top right: cross sections showing the true model (black), the solution model (red) and +/- one standard deviation (dashed green). Bottom left: actual error for the reversible jump tomography. Absolute difference between the true model in 3.12 and the estimated model in 3.15(b) (km/s). Bottom right: actual error for best Subspace solution. Absolute difference between the true model in 3.12 and the estimated model in Figure 3.14(c) (km/s).

problem have been suggested (Rawlinson *et al.*, 2008).

In contrast to regularisation 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. Figure 3.16 shows the results. The model uncertainty map obtained in this way (Figure 3.16(a)) appears to be similar to the actual error in 3.16(c), but with higher amplitude. Indeed, as can be seen in 3.16(b), the one standard deviation estimated error (green dashed lines) includes the true model (black solid line) for more than 90% of the profiles. In these experiments, the Monte Carlo 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 3.16(c) can be also compared to the true error map for the Subspace solution in 3.16(d). Values are clearly smaller in 3.16(c), and hence the average solution obtained with the RJ-McMC approach is closer to the true model than the Subspace solution in 3.14(c). This can be quantified with the 'norm' for 3.16(c) being almost half of the 'norm' for 3.16(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.

3.3.4 Example with a Gaussian random model

3.3.4.1 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 on Figure 3.17 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 (Figure 3.18) are the same as in the previous example. Some random Gaussian noise has been added to the observed travel times with a standard deviation of 5 s (i.e. about 1% of the average observed travel time).

3.3.4.2 Comparing regularised 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 regularisation parameters. Here we show solutions obtained for different grid sizes (Figure

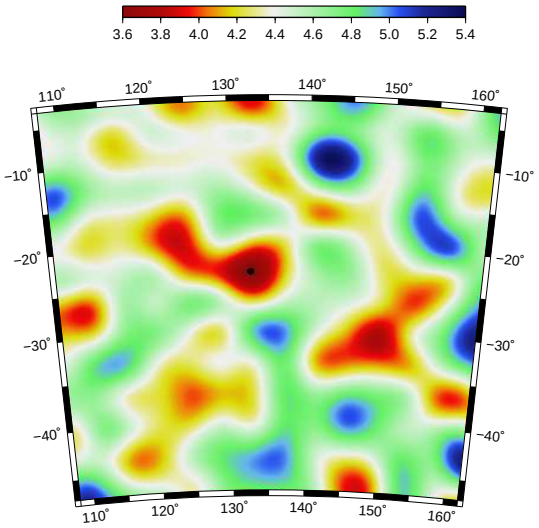


Figure 3.17: True velocity field (km/s). Grid of 14 × 14 nodes which is B-spline interpolated

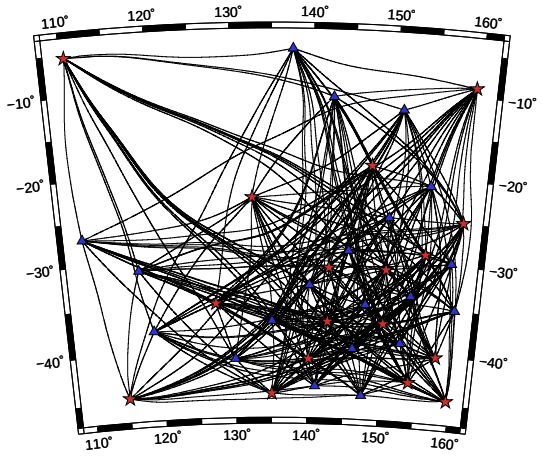


Figure 3.18: Geometry of rays. The sources (red stars) and receivers (blue squares) are located as in previous example. However, rays take different path due to the different velocity heterogeneities.

3.19) and for each case, the regularisation parameters are chosen with an "L-curve" technique (Aster *et al.*, 2005).

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 Figure 3.19(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 which is maximised at the corner of the 'L-curve'. The corner obtained at $\eta = 5$ offers a compromise between minimising 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 regularised solution for different gird sizes and results are shown in Figure (3.19).

Results clearly shows that solutions are acutely dependent on the grid size. Some features are missed if the grid is too coarse as in 3.19(a) whereas gradients not present in the true model may appear with a too fine grid as in 3.19(c). Note that the solution in 3.19(b) has been produced with the 'perfect' grid size, as it is the same node spacing used to construct the true model in 3.17

In general the average solution obtained with the Reversible jump tomography

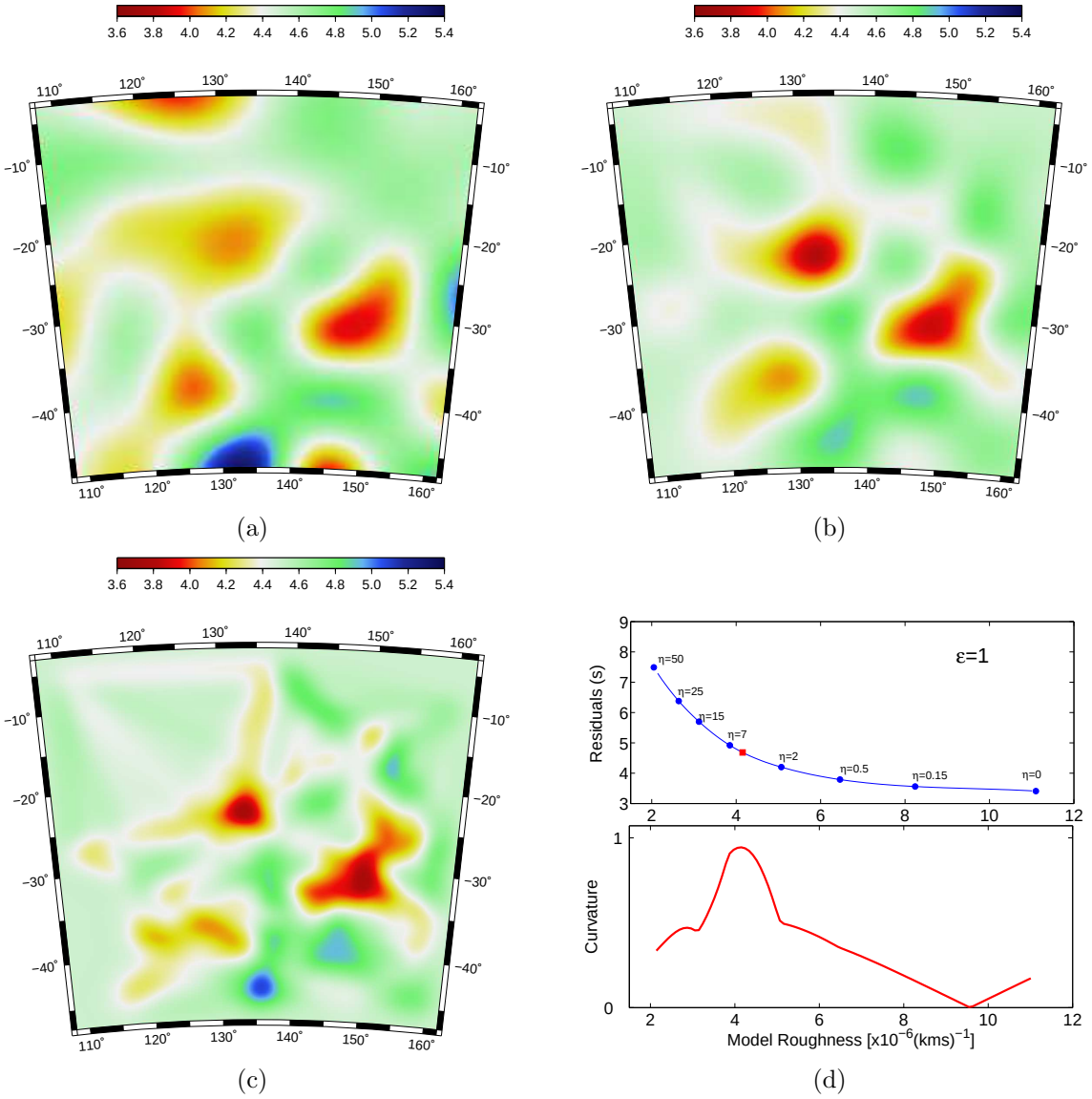
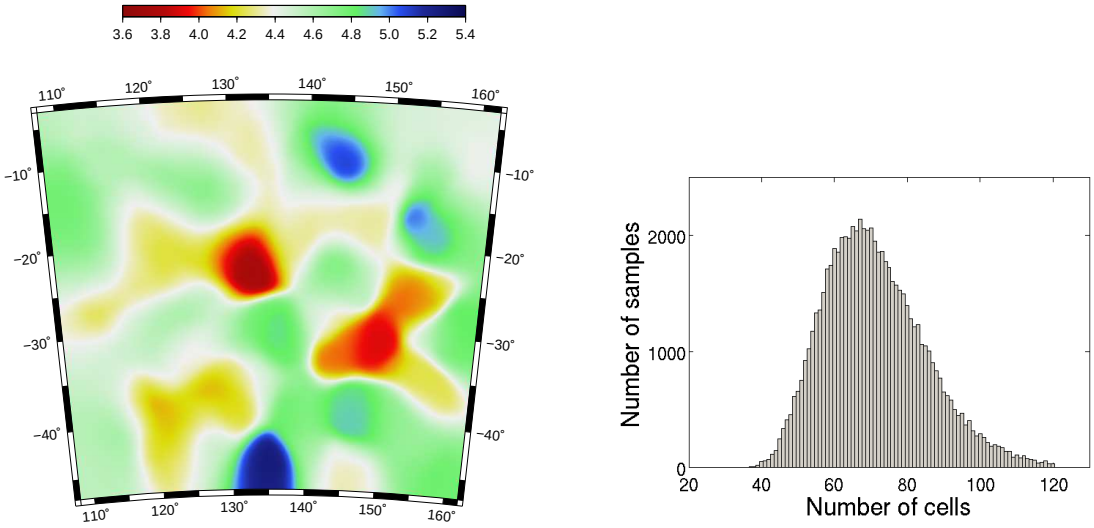


Figure 3.19: Results with the Subspace inversion for different grid sizes. Each solution is obtained after finding the regularisation parameters with an L-curve technique. (a) Grid size = 7×7 nodes. (b) Grid size = 14×14 nodes. Grid size used to produce the synthetic model. (c) Grid size = 35×35 nodes. (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.

in Figure 3.20(a) seems to recover the velocity anomalies with a better amplitude than any of the solutions obtained with the regularised method in Figure 3.19. Interestingly the recovered velocity map seems to be an improvement over the case where the regularised scheme uses the actual parameterisation of the true model



(a) (b)

Figure 3.20: Reversible jump tomography. Results after 3 iterations (km/s). Left: Solution map obtained by averaging 10.000 post burn-in samples. The scales are the same as for previous figures. Right: Information on the number of cells in the Voronoi tessellation. Posterior probability density for the parameter n, $p(n, \mathbf{d}_{obs})$

(Figure 3.19(b)). 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 cell sizes. Figure 3.20(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.

3.4 Ambient noise data example

The use of ambient seismic noise to recover the travel times of surface waves between pairs of stations is rapidly becoming popular (Shapiro and Campillo, 2004). There are various causes of seismic noise. The most energetic component is the oceanic microseism which 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 which create continuous pressure on the sea bottom, with variable intensity. The disturbance of the sea bottom results in the emergence of

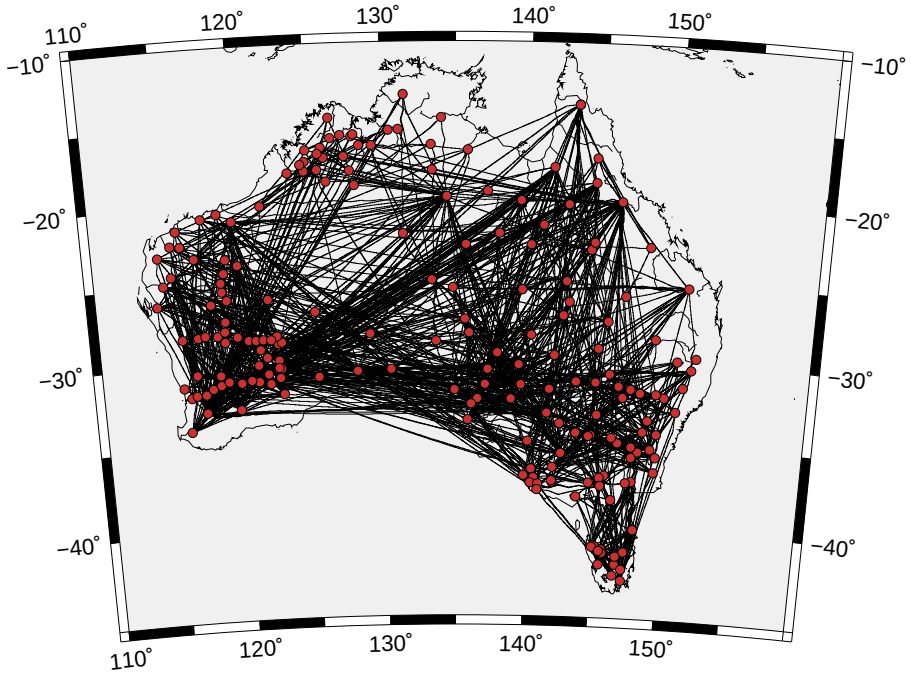


Figure 3.21: Map of connecting raypaths between couples of stations for 0.2 Hz. Red circles show the locations of the stations.

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 and Paul, 2003) showed that it is possible to perform the cross correlation between signals recorded at two stations and extract the Green's function (the signal that would be recorded at one station if an impulsive force was applied at another station). A simple demonstation of this property is based on a modal representation of a diffuse wavefield inside an elastic body (the Earth in our case) (Lobkis and Weaver, 2001). Shapiro and 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.

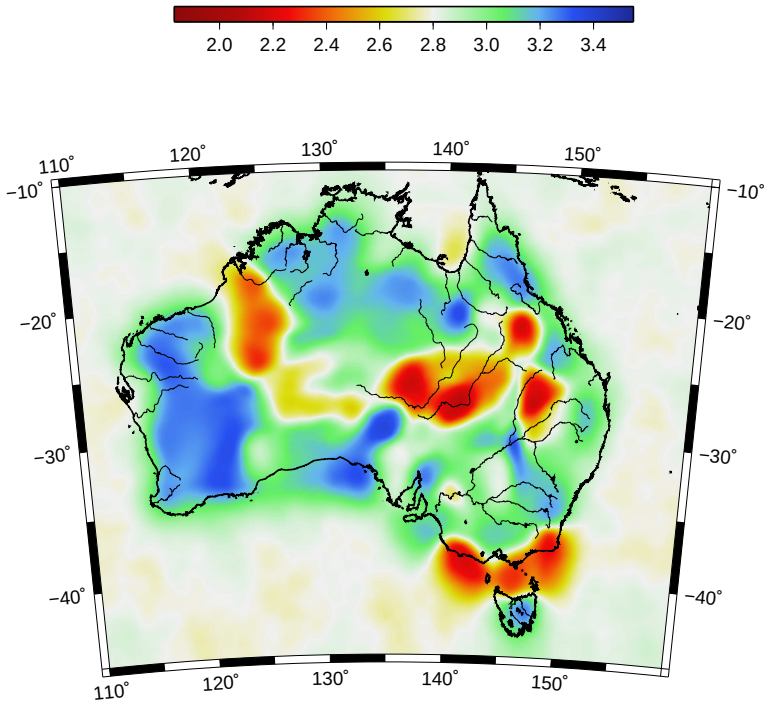
This new idea created an opportunity to use an important part of the recorded wavefield of the dense networks on the Earth that is normally neglected. New measurements can be obtained for paths that could not be sampled with the ballistic waves and therefore, can significantly improve the resolution of seismic images. Saygin (2007) compiled all the seismic broadband data from temporary and permanent

stations across the Australian continent from 1992 to 2006. The data was used to calculate the Green's functions between each possible station pairs which resulted in a coverage of the continent as in earthquake tomography studies with over 1000 individual ray paths. All of the available data was used for the calculation scheme from a lower duration limit of 15 days up to several years. Due to the inter-station distance and spectrum characteristics of the noise field, the extracted signal was mainly the Green's function of Rayleigh type surface wave for vertical components. For each frequency, the rayleigh wave arrival time could be picked on the envelope of the band-pass filtered seismogram. The extracted travel times were used to build a tomographic image of the group velocity for the Australian crust with frequency dependency.

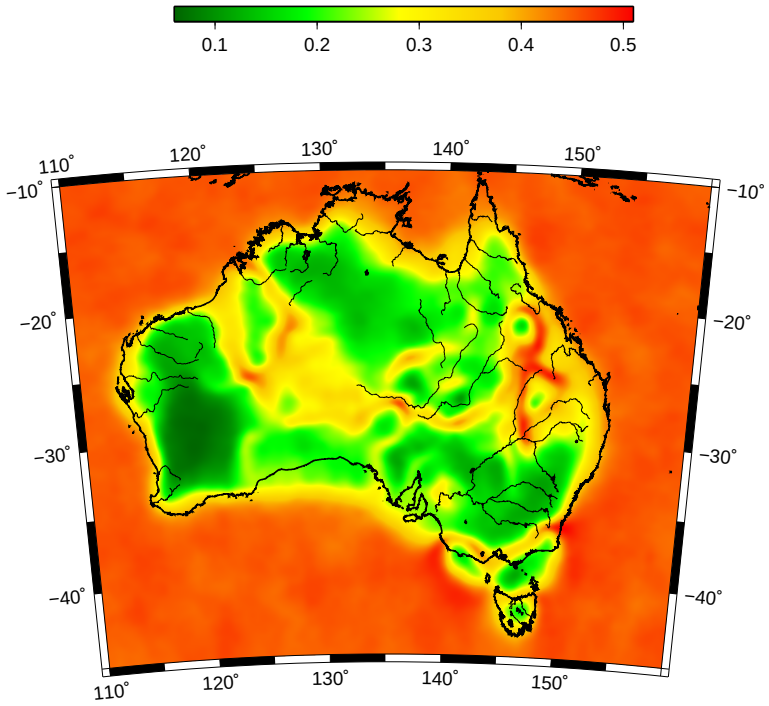
We propose here to use the same set of measured travel times for a period of 5 s (0.2 Hz) and test the reversible jump tomography. Apart from dealing with observational data, this example differs from the synthetic problem in the sense that all the receivers are considered as virtual sources. The 208 stations used in the experiment are represented by red circles in Figure 3.21. 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 travel time 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 Figure 3.21. 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.

3.4.1 Results

Figures 3.22(a) and 3.22(b) show the results obtained after 3 iterations with the new scheme. The prior on the number of cells $p(n)$ was set uniform with $n_{min} = 0$ and $n_{max} = 500$. Bounds for the uniform prior distribution for velocities were determined on the basis of the overall statistics of surface wave velocities in Australia, i.e. 1.9-3.5 km/s. 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 200000 steps were discarded as burn-in. Then, every 150^{th} model visited was taken in the ensemble. There is no data in the ocean areas and according to Bayesian principles, we recover the mean and the variance of the prior probability density function. Figure 3.23 shows the results of the posterior on the number of cells, $p(n, \mathbf{d}_{obs})$. Here there are more rays and hence more information than in the synthetic example. As a consequence, the algorithm has automatically chosen to parameterise the model



(a) Average solution map (km/s)



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

Figure 3.22: Reversible jump tomography. Results after 1 iterations of 4.10^5 Markov samples.

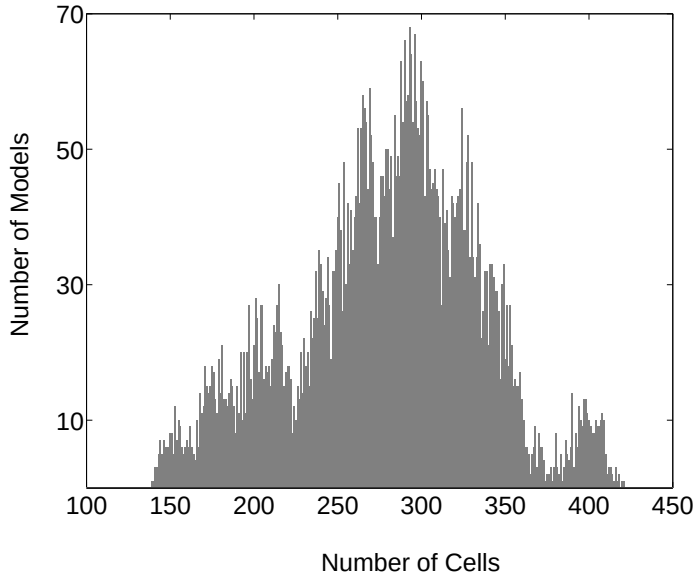


Figure 3.23: Posterior probability density for the parameter n

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 travel time used here are mostly sensitive to the structure in the first 3 km of the crust (Saygin, 2007). The Average model in Figure 3.22(a) reveals some features that can be correlated with the surface geology of the Australian continent in Figure 3.24. The zones of elevated wave speed in western Australia correspond with the Pilbara and Yilgarn cratons which 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 around 3.2 km/s (Betts *et al.*, 2002; Clitheroe *et al.*, 2000b). 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 (3.22(b)). The short period tomographic image shows multiple low velocity zones with velocities lower than 2.4 km/s, 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.* (2000a).

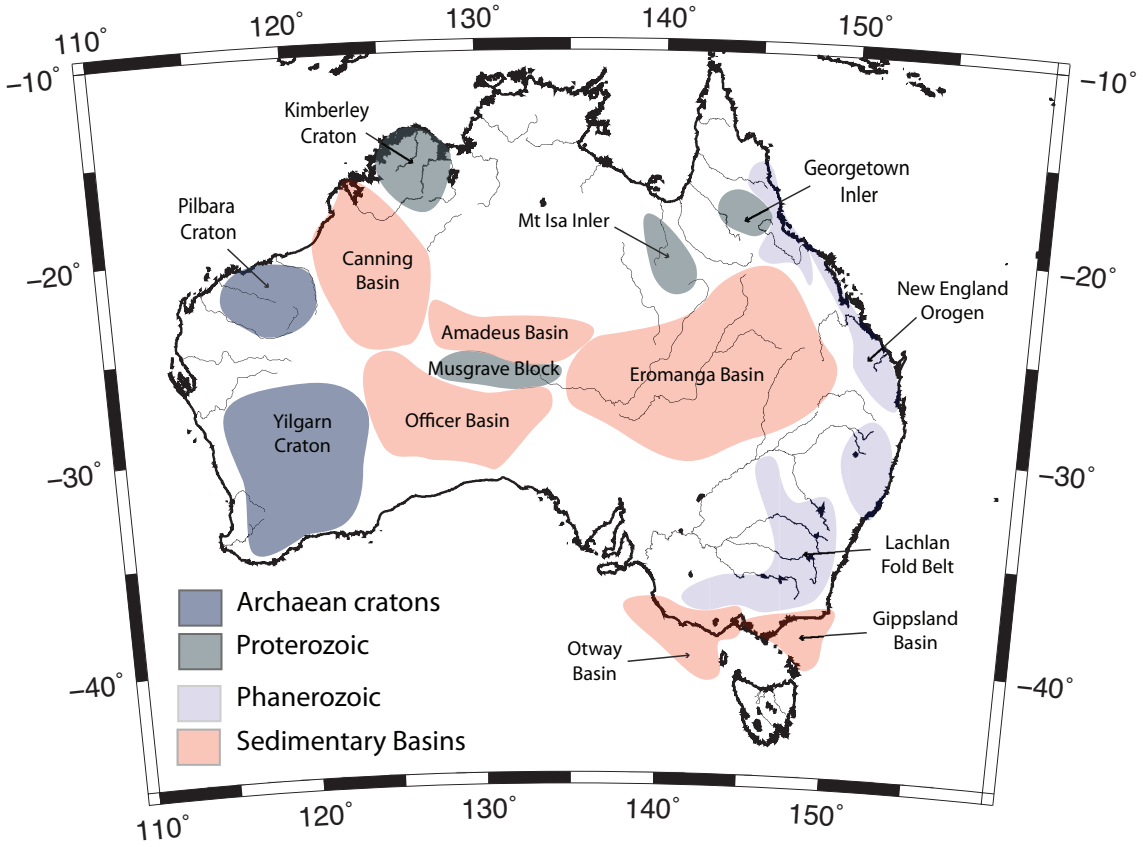


Figure 3.24: Surface geology of Australia

Although our analysis of the geological implications of Figure 3.22(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 parameterisation or any regularisation procedure. The standard deviation map in Figure 3.22(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.

3.5 Discussion

In this chapter we have improved the partition modelling tomography methodology presented in chapter 2. First, the Bayesian formulation has been extended to trans-dimensionality by means of the rj-McMC algorithm (Green, 1995, 2003). The model space is explored by sampling models of varying dimension. Second, the sampling efficiency has been improved by using the delayed rejection scheme (Tierney and Mira, 1999) which allows for proposal distributions that adapt to the shape of the posterior distribution. Third, The algorithm has been parallelised.

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 regularisation or fixed grid parameterisation. Moreover, the optimisation of the McMC sampler has proved to be efficient in terms of computational costs and these preliminary results are encouraging for applications to larger datasets, different problems and extension to 3D.

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 Markov chain Monte Carlo algorithms. (See Han and 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 and 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 work would be to extend this scheme to transdimensional moves as shown in Green and Mira (2001) or to use recent development of assisted or automated proposal generation for transdimensional sampling scheme (e.g. Tierney and 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 work 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 optimisation based approaches with fixed grids and globally constrained

regularisation schemes.

Chapter 4

Accounting for Data Noise Uncertainty – Theory and Application to Palaeoclimate Data

4.1 Motivation

Applying the reversible jump algorithm to real data from the seismic ambient noise problem highlighted new difficulties that would not have appeared with synthetic examples. These practical problems have eventually been beneficial as, by trying to solve them, our theoretical understanding had to be pushed further and the algorithm improved. In this chapter, we present one of these issues and show how it has been solved.

One characteristic of ambient noise cross-corelation travel times is that little is known on the measurements errors (Bensen *et al.*, 2007). These techniques are rapidly becoming popular and widely used but no clear method has been presented to date to quantify the uncertainty on the measured travel times (although Yao *et al.* (2006), Weaver *et al.* (2009) and Hubans *et al.* (2010) recently analysed phase and group velocity biases in ambient noise tomography). This lack of information on data errors does not represent an issue in linear (or linearised) tomography and has been disregarded. As shown below, this is because in an optimisation based inversion, the estimated model is independent of the scale of the data covariance matrix.

However, an important feature of transdimensional Bayesian inversion is that the level of data uncertainty estimated by the user prior to inversion (i.e. the data covariance matrix) directly determines the complexity of the solution (i.e. the

number of model parameters). We shall show that our transdimensional Bayesian procedure naturally adapts the complexity of the solution in order to fit the data up to the level of noise determined by the user. Certainly, a better fit can always be obtained by adding more unknowns, but the algorithm naturally prevents the data to be fitted more than the given level of data noise. Therefore it becomes clear that a major issue of transdimensional inversion is the quantification of data noise, which is often difficult.

In this chapter we propose to address the issue of noise estimation by extending the Bayesian formulation to hierarchical models which are able to take account of the lack of information the user has on the data errors. We use a Hierarchical Bayes formulation (Gelman *et al.*, 1995; Malinverno and Briggs, 2004; Malinverno and Parker, 2006) where the level of data noise is treated as an unknown in the inversion. In this way we let the data infer the appropriate level of data fit, and hence the approach fully takes into account the combination of effects contributing to the misfit.

The purpose of this chapter is to introduce the Hierarchical Bayes methodology. All ideas are illustrated on 1D regression problems that are either linear or nonlinear. Hence, the reader can visually appreciate on a single figure the data vector, the data noise, the true and estimated model, the data fit, etc... An advantage is that algorithms are simple to write and tests are quickly run. Furthermore, nonlinear regression methods have a wide range of potential applications in geosciences. However, the key idea presented here (i.e. data noise estimation as part of the inversion process) is not restricted to 1D regression model and a more general applicability will be presented in further chapters.

This chapter is divided into two sections. In the first part we present the issues related to data noise and its relation to model complexity both in the case of optimisation and Bayesian inversion. We introduce hierarchical models and show how they overcome the lack of direct information about data uncertainty. The second section shows a direct application of the method in palaeoclimatology.

4.2 Model dimension and data uncertainty: towards an expanded Bayesian formulation

In an inverse problem, it is well known that the data fit is improved as more unknowns are added into the problem. Therefore, if too many parameters are used, the distance between estimated and observed data may become smaller than the

actual data noise. In this case, the measurements are overfitted and the solution model may show spurious features due to the noise in the data.

4.2.1 Linear regression and chi-square statistical test

This effect can be easily observed on a simple linear regression problem. In Figure 4.1, we have fitted a synthetic dataset with 3 polynomials of different degrees. The true model (grey line) is a 3^{rd} order polynomial function and the data have been generated from the true model plus a random Gaussian noise. The left panels in Figure 4.1 clearly show that the data points are fit better as the order of the polynomial model is increased. The general least square solution used in linear regression is given by

$$\mathbf{m}_{L_2} = (G^T C_d^{-1} G)^{-1} G^T C_d^{-1} \mathbf{d}_{obs} \tag{4.1}$$

where C_d is the data covariance matrix. It is worth noting that $\mathbf{m}_{L_2}$ does not depend on the absolute value of the data noise. Indeed, $\mathbf{m}_{L_2}$ does not change when C_d is multiplied by a constant factor. In most cases, C_d is considered proportional to the identity matrix and is therefore removed from (4.1). In other words, the shape of the fitted polynomials in Figure 4.1 do not change as we multiply the error bar of each data point by a constant factor. In fact, only the estimated error on the model (i.e. the posterior model covariance matrix) depends on the data noise. Note that when the problem is not full rank (e.g. in seismic tomography) and requires regularization a common solution would look like

$$\mathbf{m}_{L_2} = (G^T C_d^{-1} G + \mu C_m^{-1})^{-1} (G^T C_d^{-1} \mathbf{d}_{obs} + \mu C_m^{-1} \mathbf{m}_0) \tag{4.2}$$

where $\mathbf{m}_0$ is a reference model, C_m is the a priori model covariance matrix and μ a regularization parameter. In this case, a scaling factor in C_d simply absorbs into the regularization parameter.

In optimisation methods, the aim is to find the model that minimises the data fit. Assuming that the data have random independent errors that are normally distributed with expected value zero, the chi-square χ^2_{obs} misfit measure is defined by

$$\chi^2_{obs} = \left\| \frac{G\mathbf{m}_{L_2} - \mathbf{d}_{obs}}{\sigma_{est}} \right\|^2 \tag{4.3}$$

where $\mathbf{m}_{L_2}$ is the least square solution and σ^2_{est} is the estimated variance of measurement errors in $\mathbf{d}_{obs}$. As can be seen on Figure 4.1, the χ^2_{obs} misfit measure decreases as the order of $\mathbf{m}_{L_2}$ increases.

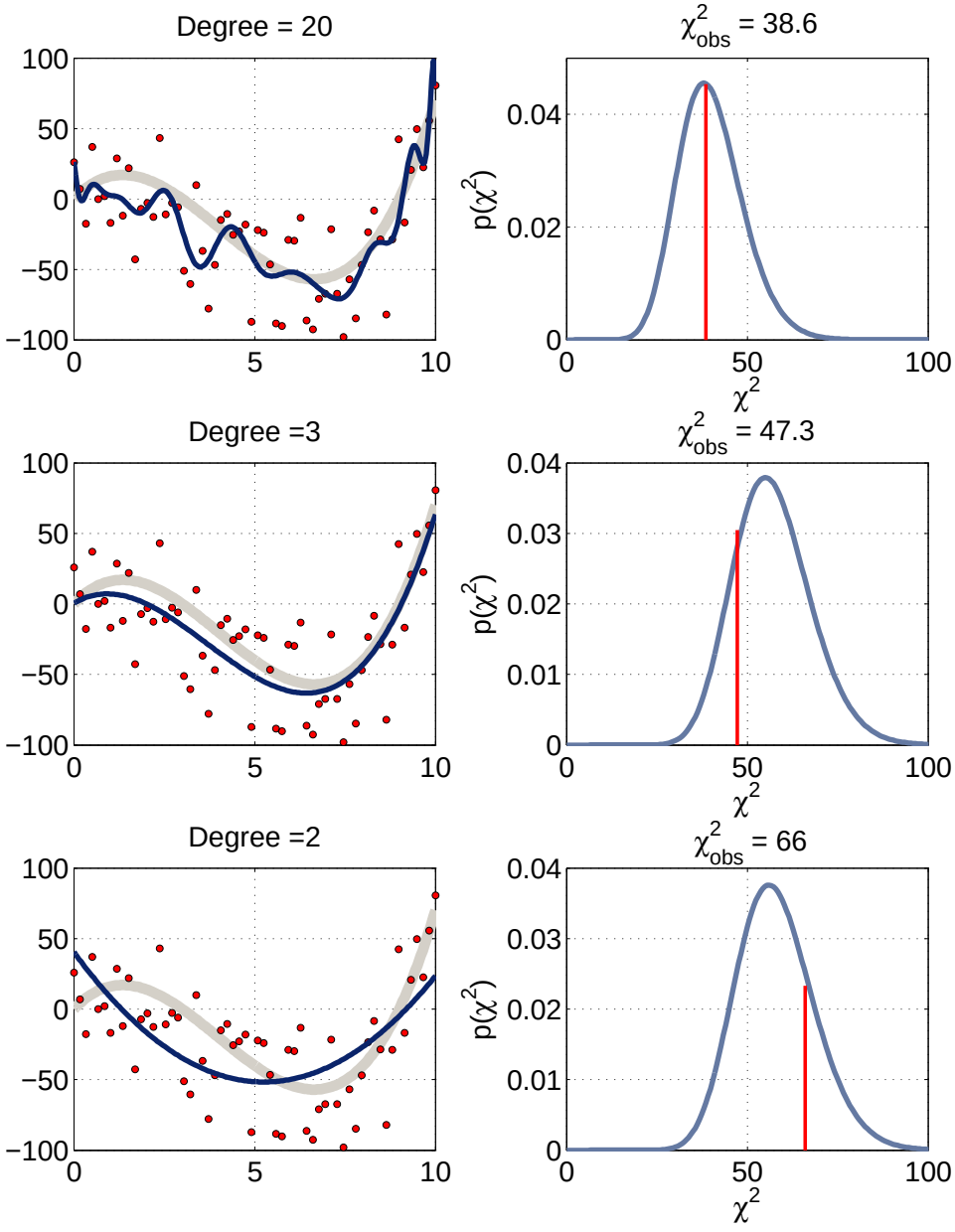


Figure 4.1: The synthetic true model (Thick light grey line) is a 3^{rd} order polynomial function. The synthetic data (60 red circles) are defined by the true model plus a random Gaussian noise with a standard deviation of 30. The data can be fit with polynomial functions by mean of a simple least squares linear regression scheme and left panels show the solutions (blue lines) for three polynomials models of different degrees. Right panels show χ^2 tests. For the three solutions, the χ^2_{obs} misfit measures (red line) have been computed with the correct data noise standard deviation ($\sigma_{est} = 30$). Therefore, the comparison between the χ^2_{obs} measures and their theoretical probability distributions are supposed to indicate the validity of the mathematical assumptions, that is the degree of the fitted polynomial. The solution obtained with the correct mathematical model (i.e. 3^{rd} order polynomial) is shown in the middle panels. However, the chi-square test does not give a clear preference to this model.

In order to get a solution model that fits the data to the required level given by the measurments errors, several statistical tests can be used to choose the number of model parameters. For instance, the chi-square statistical test (Aster *et al.*, 2005) can be carried out. Since the χ^2_{obs} misfit measure depends on σ_{est} which is the variance of a random variable, χ^2_{obs} is itself a random variable and has a χ^2 distribution with $\nu = N - n$ degrees of freedom (with N the number of data and n the number of model parameters). The probability density function for the χ^2 distribution is

$$f_{\chi^2}(x) = \frac{x^{\frac{1}{2}\nu-1}e^{-x/2}}{2^{\nu/2}\Gamma(\nu/2)}. \tag{4.4}$$

The expected value for χ^2_{obs} is ν.

In a chi-square statistical test, χ^2_{obs} is computed and compared to its estimated probability distribution. Hence, a χ^2_{obs} value far from the mode of the distribution means that some of the assumptions made are incorrect. For example, if χ^2_{obs} is much larger than its expected value, the data noise σ_{est} may have been underestimated. Another possibility is that the mathematical model is too simple (e.g. to smooth) and cannot produce a good data fit. In this case more unknowns should be incorporated to the problem. Conversely, a too small χ^2_{obs} value (relative to its expected value) may either indicate a too complex model, or that σ_{est} has been overestimated. For a discusion on chi-square statistical tests, see Aster *et al.* (2005).

In the example in Figure 4.1, we have used the correct value for σ_{est} when computing χ^2_{obs}, i.e. the value used to produce the synthetic data. Therefore, the chi-square statistical test provides information about the complexity of the model. The right panels of Figure 4.1 show the observed χ^2_{obs} values and their probability distribution $p(\chi^2)$ given by (4.4) for the 3 different fitted models. Although the 2^{nd} order polynomial function gives a relatively high $p(\chi^2_{obs})$ value, it is difficult to discard any of the three models according to this test. Note that to quantitatively compare the 3 solutions, we can scale the input error by dividing χ^2_{obs} by $\mathrm{E}(\chi^2)$. The numbers obtained are 0.96, 0.83 and 1.14 respectively which gives a preference to the most complex solution in the upper panel as it has the closest observed χ^2_{obs} value to its expected value. This shows that, even given the correct level of data noise σ_{est}, it is still difficult the recover the complexity of the true model.

Another test that can be invoked to estimate the complexity of the solution is the Bayesian Information Criterion (BIC) (Schwarz, 1978). Under the assumption that the model errors are normally distributed,

$$\mathrm{BIC} = \chi^2_{obs} + n\ln(N). \tag{4.5}$$

Schwarz (1978) used a Bayesian argument to show that, given any two estimated models, the model with the lower value of BIC is the one to be preferred. The BIC is an increasing function of χ^2_{obs} and an increasing function of n. Hence, lower BIC implies either fewer model parameters, better fit, or both. In our example, Bayesian Information criterion suggests that the second model (i.e. the 3^{rd} order polynomial) is to be prefered as the BIC values for the 3 proposed models are approximatively 120, 64, and 78 respectively.

In seismic tomography, some resolution tests (spike tests, checkboard tests) are also used to choose the number of unknowns, although regularisation procedures are used most of the time to smooth and damp over-dimensioned models. Hence, seismologists are often more interested in smoothing parameters rather than in model complexity. It is worth here emphasising the difference between the two approaches. The model complexity is the size of the vector used to approximate the true model whereas the smoothness is the first or second derivative of this vector. Thus the smoothness represents a measure of complexity in terms of the physical structure of the model. Tomographic models are often over-parameterised (i.e. they have too many model parameters which makes the problems under-determined) but appear simple due to the smoothing. Conversely, a regression model like a cosine with a small period can appear to be very complex because of its fast variations whereas it is described by only 3 model parameters (a period, a phase, and an amplitude). Hence different measures of complexity result in different behaviours.

Note that in optimization methods all measures of complexity and data misfit are 'point' measures in that they only depend on the best fit solution, i.e. a single solution. This can be contrasted with a transdimensional Bayesian approach where the solution is an ensemble of models with different complexities.

4.2.2 Non linear regression with the reversible jump algorithm

Here we apply the reversible jump scheme described in chapter 3 to the regression problem presented above. The fitted model is not a polynomial but a partition model which is simply the 1D equivalent to Voronoi cells. The parameterization is illustrated in Figure 4.2. As the position and number of nuclei defining the partition model are variable, the regression becomes non-linear (although linear in the coefficients for the model within a given partition).

The algorithm used here is virtually the same as in the reversible jump tomog-

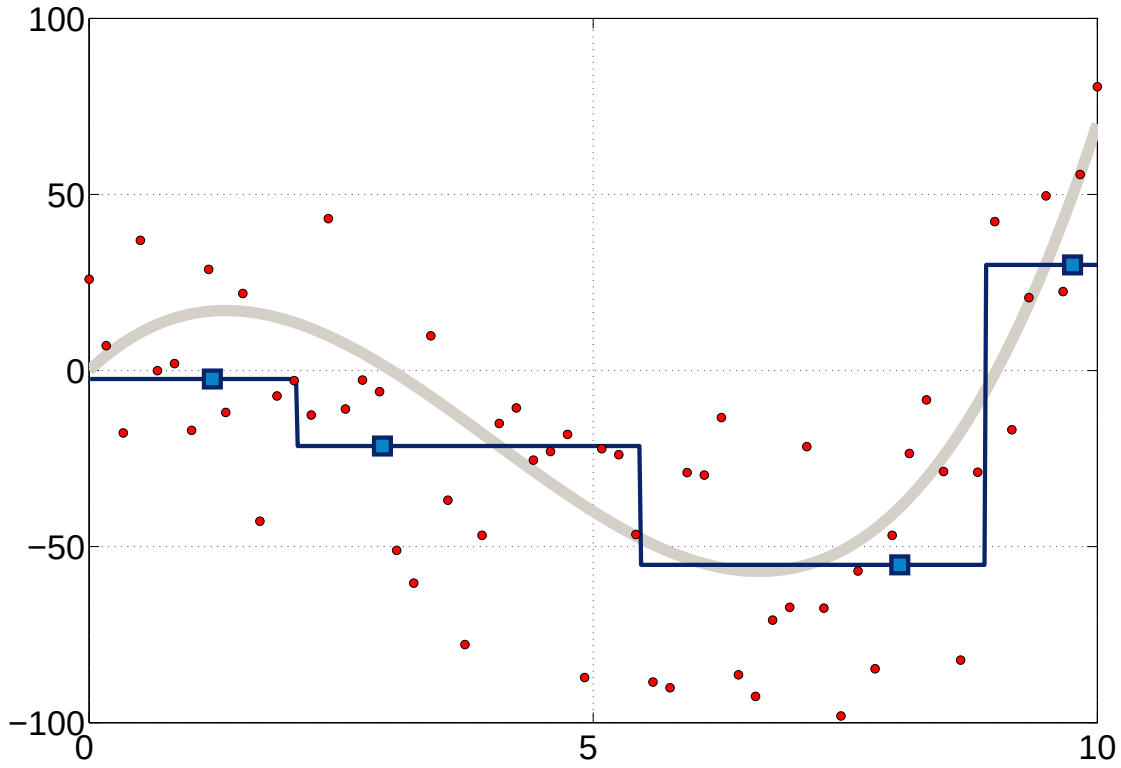


Figure 4.2: Figure establishing the parameterization of a 1D partition model. The 60-point data set (red dots) is the same as in Figure 4.1 which are simulated from the true model (thick light grey line) with added independent normal errors. Here, the regression model (dark blue line) is a 1D equivalent to the Voronoi cells partition model. It is defined by a variable number of nuclei (blue squares) each of which is associated a constant response value. The number, position, and response value of the nuclei are the model parameters to be inverted for.

raphy. The parameterisation is fully adaptive so the Voronoi nuclei (blue squares in Figure 4.2) can freely move horizontally and vertically during the procedure. A Markov chain is used to sample a large number of partition models that are distributed according to a target posterior probability distribution. This transdimensional distribution is defined by a Bayesian formulation as proportional to the likelihood distribution times a prior distribution. As in previous applications, the Markov chain repeatedly proposes and then either accepts or rejects the proposed models in the usual way. The final solution is given by the average over the ensemble of sampled models. All the model sampled have a particular parameterisation defined by the position of their nuclei. When a large number of models are stacked, their cells overlap so the spatial average model is effectively a continuous line. For

details on the algorithm, see chapter 3.

Here the prior on the number of cells is a uniform distribution over the range $[1, 50]$. These bounds have been set to large values so the prior is as least informative as possible. For each cell, the prior on the location of its nucleus is a uniform distribution over the range $[0, 10]$, and the prior on the constant response value associated with it (i.e. the vertical position of the nucleus) is a uniform distribution over the range $[-100, 100]$. We acknowledge that, even with a large width of prior on the response variable, it still affects the acceptance rate in the birth/death jumps and therefore it is impossible to use a completely non informative prior.

The likelihood is simply defined from a least square misfit function given by the distance between observed and estimated data.

$$p(\mathbf{d}_{obs} \mid \mathbf{m}) \propto \exp\left\{ \frac{-\left\| g(\mathbf{m}) - \mathbf{d}_{obs} \right\|^2}{2\sigma_{est}^2} \right\} \tag{4.6}$$

In contrast to optimization schemes, with the reversible jump approach the model dimension is directly adjusted in order to fit the data to the degree required by the estimated noise. Hence, the solution model depends on the data but also on the estimated data noise. As can be seen in the form of the acceptance term (see chapter 3), a proposed model is more likely to be accepted as the difference between the current and proposed model decreases. As we increase the value of the data noise, we decrease the difference between model misfits and the algorithm becomes more 'permissive' and accepts more easily 'worse' models. Conversely, when the estimated data noise is decreased, the required data fit is increased and the Markov chain naturally adds more parameters to provide 'better' models.

In order to experience the effect of the estimated data noise on the average solution, the algorithm was run three times with a different value for σ_{est} at each run. The three solution curves obtained are shown in the left panels of Figure 4.3. Dashed lines show the 95% credible intervals. Since the models in the ensemble solution have a varying number of cells, the complexity of the average model cannot be described with a single number n. However we plot on right panels the histogram of n across the ensemble solution that is directly proportional to the marginal posterior $p(n|\mathbf{d}_{obs})$. Note that the number of nuclei n scales linearly with total number of unknowns.

In the top panels, the noise has been underestimated ($\sigma_{est} = 15$), and hence data are over-fit. The middle panels show results obtained with the correct data noise estimation ($\sigma_{est} = 30$). Lower panels have been obtained with $\sigma_{est} = 80$, which

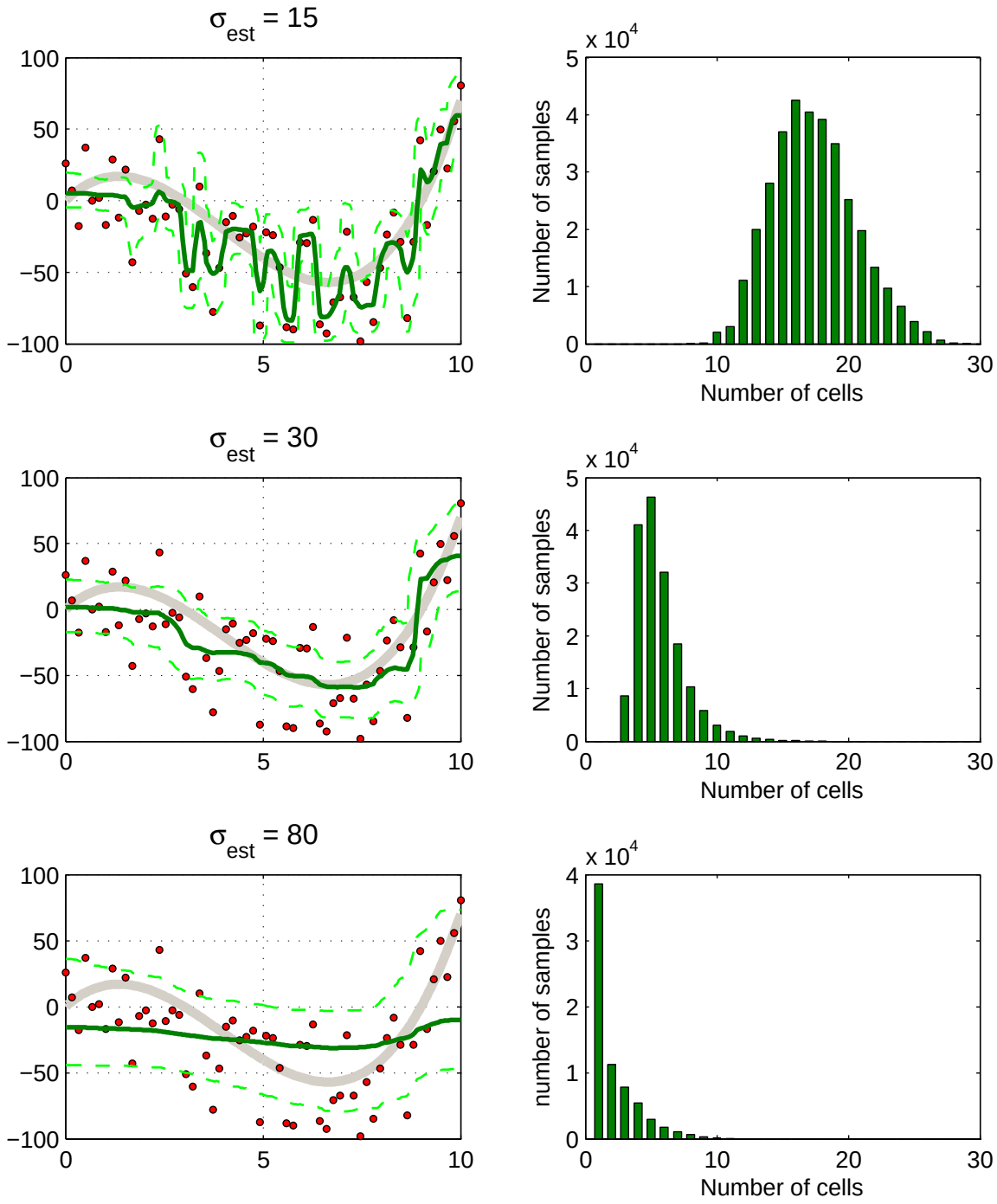


Figure 4.3: Non linear regression with the reversible jump algorithm. Left panels show average solution models (green lines) obtained with different σ_{est} values. The 95% credible intervals on the predictions are shown with dashed green lines. The right panels show the respective distributions for the number of cells. The two middle panels show results obtained with the correct level of data noise ($\sigma_{est} = 30$).

is clearly overestimated. The number of model parameter used, and therefore the complexity of the average solution, clearly depends on the estimated data noise. Intuitively, we can see that if we underestimate the data noise, then we will tend to fit many cells (i.e. the model becomes too complex in the top panel). On the other hand if we regard the data as more noisy than they really are, then we will tend to fit a model that has too few cells (i.e. the model is too simple in the lower panel). Hence the level of data noise effectively quantifies the usable information present in the data (a very noisy dataset does not contain much retrievable information), and thus it naturally controls the quantity of information that consequently should be present in the model (i.e. the number of model parameters).

Contrary to optimisation schemes, here there is no need for statistical tests or regularisation procedures to choose the adequate model complexity or smoothness coresponding to a given degree of data uncertainty. Instead, the reversible jump technique automatically adjusts the underlying parametrisation of the model to produce an average solution with just enough complexity to fit the data.

We acknowledge that the average solution curves are further away form the true model than the polynomial models obtained with linear fitting. This is because the mathematical models used here (i.e. partition models with constant value in each cell) are far from the true mathematical model (a polynomial function). Hence we have a parameterization error due to mismatch between the polynomial used to generate the data and the choice of model parameterisation used in the inversion. Better solutions (i.e. closer to true solution) would be obtained by allowing the partition models to follow a linear or quadratic function within each cell (see Denison *et al.*, 2002, for an example), or allowing the order of the polynomial to vary in each cell.

The value of σ_{est} imposed at the outset has a direct effect on the solution of the reversible jump algorithm and implicitely acts as a smoothing parameter. This can be seen as an advantage over linear inversions where the level of data noise is not accounted for and where the level of smoothing is chosen a priori. However, in some experiments as to be seen in next chapters, assessment of measurements errors can be difficult to achieve a priori. Without any reliable information about the data uncertainty, it is impossible to give a preference between two solutions obtained with different values of σ_{est}.

4.2.3 Uncertainty quantification: Hierarchical Bayes.

Fortunately, an expanded Bayesian formulation can take into account the lack of knowledge we have about data errors. Instead of being fixed, the variance of the measurement errors can have a broad prior uncertainty and posterior inference can be performed. Following current statistical terminology, σ_{est} becomes a 'hyperparameter', and the method used is known as Hierarchical Bayes (Gelman *et al.*, 1995). The value of the hyperparameter is determined by the data, and the final result is a posterior distribution for both the hyperparameter and the Earth-model parameters. Note that the number of Voronoi cells n is also a hyperparameter as it is not directly related to Earth properties. By letting the number of cells being a unknown, we were actually already using an 'Hierarchical Bayes' framework in chapter 3.

In geophysics, Malinverno and Briggs (2004) and Malinverno and Parker (2006) were the first to use a Hierarchical Bayes formulation and invert for the data noise. They applied this method to the linear Gaussian problem, where the relationship between Earth-model parameters and measurements is linear and where the prior distribution and the likelihood function are multivariate normal distributions. They demonstrated the practical application of this approach to a simple linear inverse problem: using seismic travel times measured by a receiver in a well to infer compressional wave slowness in a 1D Earth model. In their work, the posterior distribution was Gaussian, and its mean and variance could be easily computed analytically. In the work presented here, we propose to apply Hierarchical Bayes to a fully non-linear regression problem, where the posterior is numerically estimated with the reversible jump McMC algorithm.

The model to be inverted for is defined by the combined set $(\mathbf{m}, \sigma)$ where $\mathbf{m}$ is the vector containing the nuclei locations and values and σ is the unknown standard deviation of data errors. Therefore, for the 1D regression problem, the dimension of the model space is $2n + 1$ with n the number of nuclei (itself a parameter to be estimated). The Gaussian likelihood function is defined by

$$p(\mathbf{d}_{obs} \mid \mathbf{m}, \sigma) \propto \frac{1}{\sigma^N} \exp\left\{ \frac{-\left\| g(\mathbf{m}) - \mathbf{d}_{obs} \right\|^2}{2\sigma^2} \right\} \tag{4.7}$$

where N is the size of the data vector $\mathbf{d}_{obs}$. Note here that in contrast to common inversion treatment (e.g. Aster *et al.*, 2005), a new variable to be solved for, σ, appears in the front of the exponential term due to normalisation. In the usual case where σ is assumed known this is a simple factor which may be neglected. However here its presence contributes to the nonlinearity of the inverse problem.

The hierarchical algorithm is implemented in the very same manner as the conventional reversible jump, the only difference being that here we add an extra type of model perturbation, i.e. a change in the hyperparameter σ. As for other model parameters, each time σ is perturbed, a new value σ' is randomly proposed from a Gaussian proposal distribution $q(\sigma'|\sigma)$ centred at the current value σ. The likelihood of the proposed model $(\mathbf{m}, \sigma')$ is computed from (4.7), and the new value of data noise is either accepted or rejected according to the ratio of likelihoods of the proposed and current models (here the prior ratio and Jacobian are unity). As described previously, the variance of the proposal function is tuned by 'trial and error' using the rate of acceptance.

When treating the data noise as a variable, one would intuitively expect the algorithm to choose high values for σ as it would minimise the misfit (i.e. term between brackets in (4.7)). However, the Gaussian likelihood function is normalized by σ^N in (4.7) and a high σ also implies a low likelihood. Hence the value taken by σ has two competing effects on the likelihood. It is not surprising to see that, for a given partition model $\mathbf{m}_0$, the maximum of the likelihood function is obtained for

$$\sigma = \sqrt{\frac{\|g(\mathbf{m}_0) - \mathbf{d}_{obs}\|^2}{N}} \tag{4.8}$$

which is the mean value of the residuals, commonly called 'root mean square'.

Note that, instead of a single number, more parameters can be used to describe the data uncertainty. For example, as shown in the application to palaeoclimate data in section 4.3, if the measurements are collected with m different instruments, we look for a value for the noise of each instrument, $\sigma_{est} = [\sigma_1, ..., \sigma_m]$, and therefore the likelihood function takes a more complicated form. As will be seen in next chapter, the data noise in seismic tomography can be modelled as a linear function of ray length. In this case it is described by two parameters: the slope and intercept of the linear function. In chapter 6, we invert a seismic waveform, and hence the data noise is correlated. In this case we show how to invert for the noise correlation, that is the off-diagonal elements in the data covariance matrix. Data uncertainty could ultimately be described with $N \times N$ numbers, that is, the size of the data covariance matrix. Then, an apposite question would be : how many parameters do we need to describe the data noise, and can we use a transdimensional formulation to let the data answer that question as well? For the moment, we assume the data covariance matrix is proportional to the identity matrix and use a single value to describe the variance of the noise.

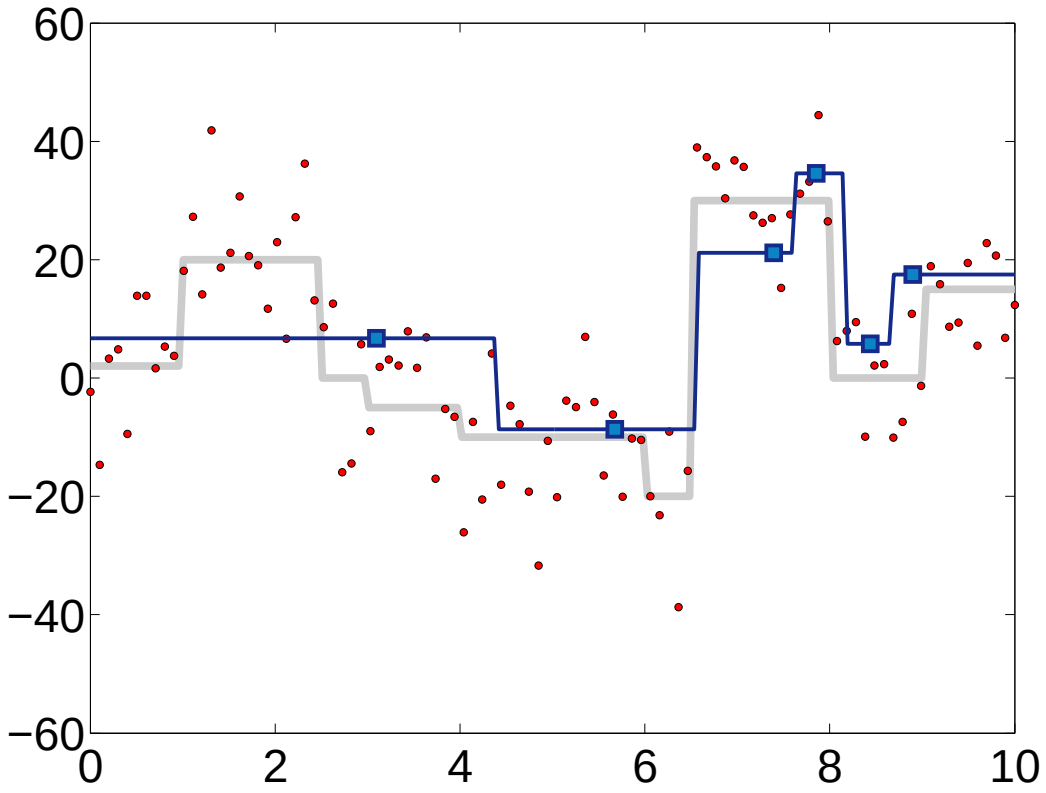


Figure 4.4: The synthetic true model (light grey line) is defined by 9 regions of constant value. The synthetic data (100 red circles) are defined by the true model plus a random Gaussian noise with a standard deviation of $\sigma_{true} = 10$. The blue line shows a sample generated by the reversible jump algorithm.

The Hierarchical Bayes algorithm is applied to the regression problem. Here, the synthetic true model is not a polynomial but a 1D partition model (Grey line in Figure 4.4) defined by 9 regions of constant value. Therefore, the samples generated along the Markov chain will be defined with the correct mathematical model. The data set consist of $N = 100$ points (red dots in Figure 4.4) that are simulated from the true model with added independent normal errors with standard deviation $\sigma_{true} = 10$ (this value is unknown during the inversion).

We first show in Figure 4.5 results obtained with the conventional Bayesian approach where a fixed value is given to the data noise. We use the word 'conventional' to refer to the reversible jump algorithm where the level of noise is fixed and given by the user at the outset. This experiment is in essence a repeat of that in Figure 4.3 only with the true model lying in the same partition model space. As previously, different choices made for σ_{est} result in markedly different posterior distributions.

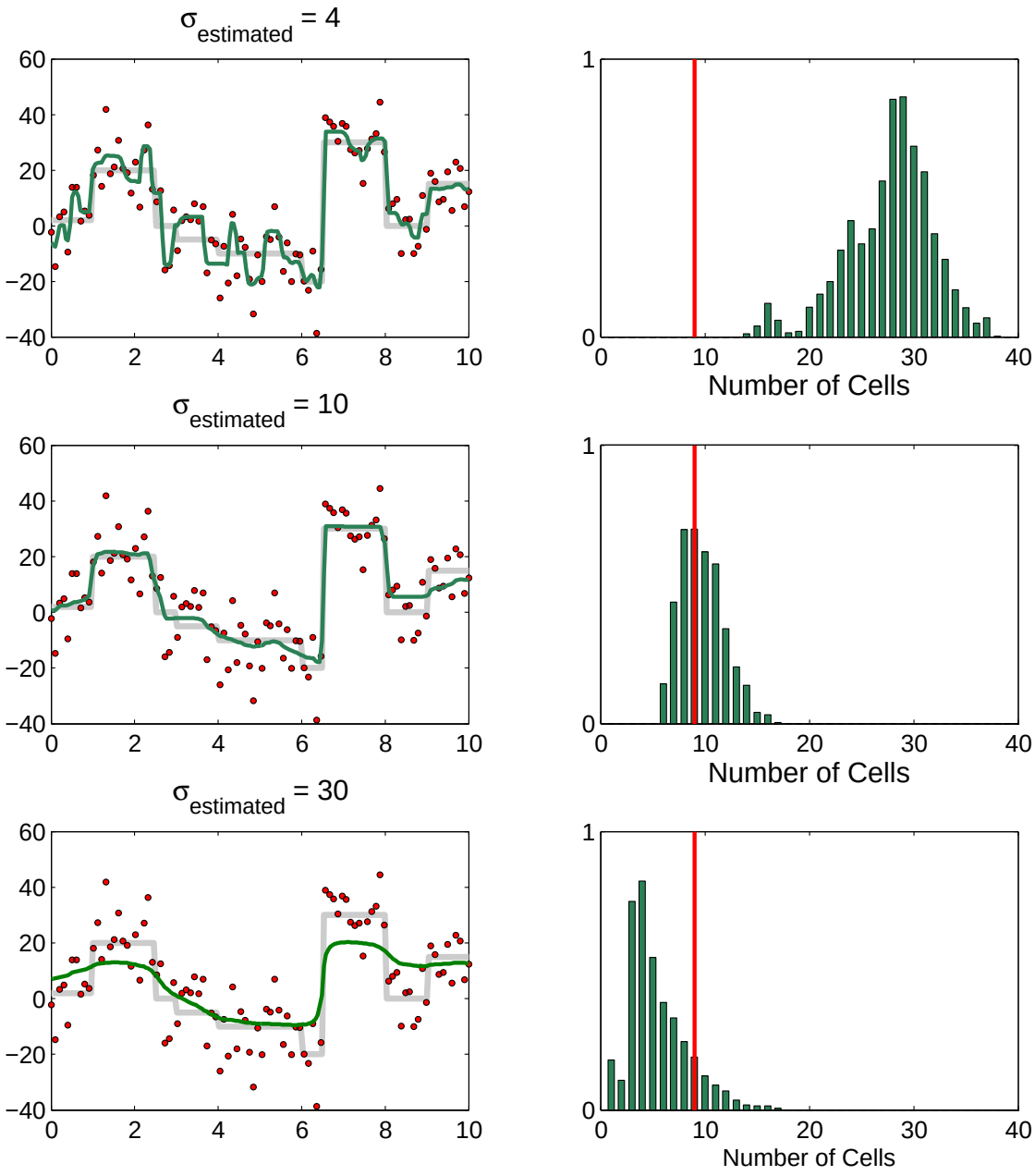


Figure 4.5: Non linear regression with the conventional reversible jump algorithm. Left panels show average solution models (green lines) obtained with different σ_{est} values. The right panels show the respective distributions for the number of cells. Red lines show the number of cells in the true model. The middle panels show results obtained with the correct data noise ($\sigma_{est} = 10$). Note that in this case, the posterior distribution on the number of cells is maximum at 9 which is the number of cells in the true model.

Hence, as in Figure 4.3, different solutions are obtained with different values of σ_{est}. Again, this demonstrates the difficulties associated with the conventional Bayesian approach when little is known about the data noise as many different results can be obtained for different values of σ_{est}.

The results obtained with the Hierarchical Bayes procedure are shown in Figure 4.6. As the hyperparameters n and σ are poorly known a priori, a wide (i.e loose) prior distribution has been assigned to them (uniform distribution over the range $[1, 40]$ for σ and over $[1, 50]$ for n). The Markov chain has sampled a large ensemble of models $(\mathbf{m}, \sigma)$ that represent the posterior distribution. The average solution is shown in the top panel of Figure 4.6. The discontinuities present in the data are well resolved. The posterior distribution on the number of cells is plotted in the middle panel in Figure 4.6. The red bar shows the number of cells in the true model. Similarly, in the third panel we construct the histogram on the value taken by the hyperparameter σ across the ensemble of models. Here the red bar shows the standard deviation σ_{true} of the random Gaussian noise present in the synthetic data.

With scant prior knowledge on both the data noise and the complexity of the true model, the algorithm is able to provide a solution model with the correct complexity, and that fits the data to the required level. The posterior distribution for the number of cells looks like a Gaussian distribution around 9, which is the number of cells in the true model. Furthermore, the expected posterior value for the hyperparameter σ is close to the true data noise which is 10.

By letting both n and σ being variable during the inversion, here it is clear that the hierarchical formulation of the Bayesian problem overcomes the problem of trade-off between data fit and model simplicity. The procedure effectively lets the data determine the hyperparameter values that are most sensible a posteriori. As opposed to optimization schemes where the aim is to find the best fitting model, here the goal is to find the best compromise between data fit and number of unknowns.

This extanded Bayesian formulation is a form of global inference: before the data are collected there is an initial state of uncertainty not only on the earth model parameters $\mathbf{m}$ but also on the hyperparameters n and σ. Once the information provided by the data is accounted for, we know more about earth properties, but also about hyperparameters. By allowing a range of possible prior hypotheses on the hyperparameters, this approach decreases the number of assumptions that need to be made before the inversion.

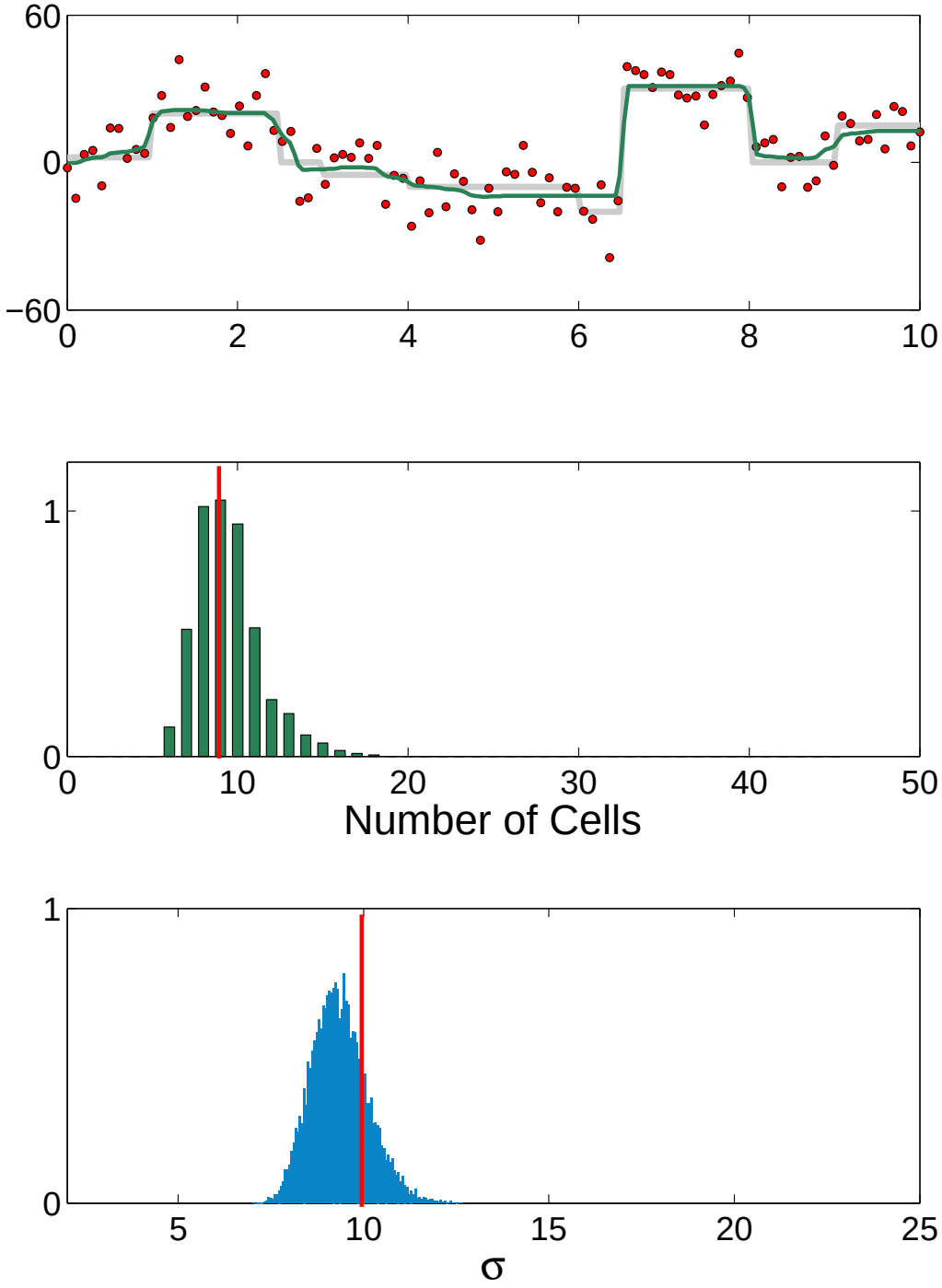


Figure 4.6: Results with Hierarchical Bayes. Top: Average solution model. Middle: Posterior distribution on the number of cell (the true value is 9). Bottom: Posterior distribution for hyperparameter σ (the true value of data error is 10)

4.2.4 Forward model uncertainty

It is important to note that the measurement errors are not necessarily the principal contributor to the data misfit. The error associated with the mathematical model or the forward theory can also contribute to the data misfit. That is, the function $g(\mathbf{m}) = \mathbf{d}$ can be an inaccurate approximation far from the reality, and may not be able to explain even perfect measurements adequately.

In the example shown in Figure 4.6, the partition models used by the Markov chain are of the same 'nature' as the true synthetic model. As a result, the data misfit is mainly due to the data random errors and the posterior $p(\sigma|\mathbf{d}_{obs})$ recovers the standard deviation of data noise. However, if the true model is not a partition model, but for instance a polynomial function as in the example in Figure 4.1 and 4.3, the samples proposed by the Markov chain will have more difficulty in fitting the data. Hence, the error associated with the forward model will contribute to the data misfit and this will be taken into account by the hyperparameter σ (see Figure 4.7).

In conventional Bayesian procedures, the number σ_{est} used to normalise the misfit in the likelihood function (4.6) is often taken as the standard deviation of measurements. This may be incorrect as errors due to assumptions made by the geophysicist may also contribute to the misfit and should also be accounted in the data covariance matrix (see Gouveia and Scales (1998) for a discussion). Scales and Snieder (1998) showed that in the context of Bayesian inference, the data noise should be defined as that part of the data that we do not wish/expect the model to explain.

For example, when performing seismic tomography assuming straight rays in a region with strong velocity anomalies (i.e. where rays are actually strongly curved), even if one has perfect measurements, it will be incorrect to use a small value for σ_{est} and try to fit the data too well. The problem is then to quantify the ability of the forward theory to explain the data. This number will have to be expressed in the same units as the data and added to the estimated data noise to get a correct value for σ_{est}. In the case of reflection seismic waveform inversion, Gouveia and Scales (1998) showed the necessity and the difficulty to account for all contribution to the misfit, i.e. ambient noise, near surface heterogeneities, scaling factor between field and synthetic data, and model discretisation errors.

In geophysical problems where assessment of data noise is problematic, a Hierarchical Bayes formulation enables to express this initial state of uncertainty on the different terms contributing to the data noise and to perform posterior inference on

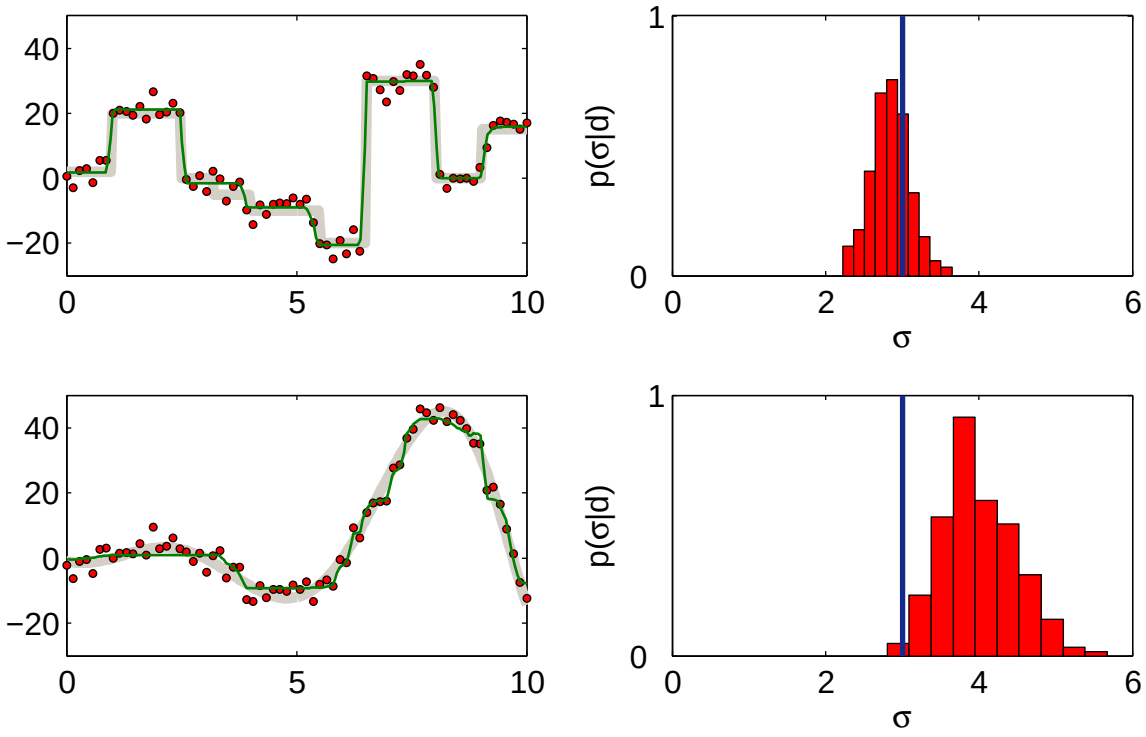


Figure 4.7: Same experiment as in Figure 4.6. The data points have been generated with random Gaussian errors with standard deviation of 3. The posterior distribution on the hyperparameter σ is shown on left. The blue line shows the true value of the level of data errors. Top: the true model is a partition model. Here, the misfit is mainly due to data errors and the algorithm recovers the true value of data errors. Bottom: the true model is a polynomial function. The samples proposed by the Markov chain have an inaccurate mathematical model and data points are harder to fit. In this case, the posterior distribution $p(\sigma|\mathbf{d}_{obs})$ shows both the data errors and the error in the mathematical model.

them. This is done with the use of hyperparameters which are determined directly from the data.

4.3 Application to change point modelling of palaeoclimate data

At present there are not yet physical theories that confidently predict when and how often the climate system experiences transitions or changes rapidly. Our information is largely empirical and based on geochemical proxy data (Ruggieri *et al.*, 2009). Examples include an effort to characterise global temperatures over the past

millennium from a variety of proxies (Briffa *et al.*, 1995; Esper *et al.*, 2002; Mann *et al.*, 1998; Jones and Mann, 2004), using borehole temperatures to inference past temperature histories (Hopcroft *et al.*, 2007, 2009), millennium long reconstructions of equatorial Pacific surface oceanography using stable isotopes from corrals (Cole *et al.*, 1993; Cobb *et al.*, 2003; Quinn *et al.*, 1998), high-resolution stable isotope and trace gas records from polar ice cores (Chappellaz *et al.*, 1993; Dansgaard *et al.*, 1993; Mayewski *et al.*, 1993; Petit *et al.*, 1999), and records of glacial-interglacial climate cycles derived from ocean sediment cores (Bloemendal and de Menocal, 1989; Imbrie *et al.*, 1989; Herbert and Mayer, 1991; Joyce *et al.*, 1990; Lisiecki and Raymo, 2005).

Although these studies use different tools to resolve various aspects of climate change at different time scales, they basically all solve a regression problem. That is, geochemical measurements are collected at different depths that are related to time (time series) and the problem is to find the 1D model (a function of depth or time) that predicts the data. Then, this model is used as a proxy to infer past climate events.

The current research effort towards understanding climate change, and the relative roles of natural and anthropogenic influences, often focuses on the inference of rapid or abrupt changes in the mean signal, over time. Thus one might idealise paleoclimatic data as a succession of periods with internally homogeneous statistical properties, bounded by abrupt shifts to subsequent or antecedent regimes (Kylander *et al.*, 2007). Therefore, the use of change point models turns out to be particularly relevant. This enables the regression model to automatically detect the position (and number) of climate shifts within data that are statistically significant.

Mudelsee (2000) proposed to quantify climate transitions by fitting a ramp to observed data. The approach consists of finding two changepoints (in this case, time points) between which the data are fit with a linear function. Outside these two points, the data are fit with a constant value. These parameters are estimated using standard least square criteria, with a brute search (global) approach to find the values of the changepoints. While the approach allows for the data noise to be a variable and indeed to vary between data points, the values are chosen subjectively (by eye).

Tomé and Miranda (2004) presented a method to find discrete changes in linear trends. Gradients between changepoints are fitted to a given time series, subject to constraints on the minimum distance between changepoints, and on the magnitude of the changes in each 'cell'. The problem is set up such that a user specifies a range

for number of changepoints, and finds the best fitting functions for each value of the number of change points in turn. Although the authors state that they can examine the data fit for each model (with different numbers of changepoints) to choose an appropriate value, they do not explicity do this in the paper. Rather they seem to favour visual inspection to select a preferred model, which also is required to have a constant distance between the changepoints.

Recently, Ruggieri *et al.* (2009) described various palaeoclimatic studies based on time series and developed a method to infer Milankovitch-type cycles from geochemical $\delta^{18}O$ data. They explicitly allow for changepoints between which the nature of the signal (defined by superimposed sine functions) can change abruptly. They apply their model to 2 sets of benthic $\delta^{18}O$ isotope data with time ranges going back to 2500 and 5000 kiloyears. As these authors state, an important limitation of their method is that they do not include the number of change points as a parameter to be inferred directly. Instead they examine the variation of the data fit as a function of the number of changepoints, and try to identify the upper limit such that adding more changepoints makes little difference to the data fit. One additional limitation of this approach is that the inference will depend on the errors inherent in the data, although the data fit function adopted by Ruggieri *et al.* (2009) does not incorporate a data error term explicitly.

These interesting approaches are oriented to finding a 'best' model, defined in the least squares sense. Rather than an optimization based approach, here we apply the hierachical Bayes regression algorithm presented above to a suite of geochemical and geophysical measurements collected from a core in the Hongyuan peatland, eastern Qinghai-Tibetan Plateau (Large *et al.*, 2009). The aim is to identify the timing and location of the distinct changes in peat deposits. In this problem it is important to note that different types of data may respond differently to a given change. Thus one may show a large increase, another a large decrease and a third a very small increase. However, the timing of each change is the same, and identifying the timing is the goal of the inference process.

4.3.1 The data

The Hongyuan peatland bog is located on the eastern Qinghai Tibetan plateau and has therefore directly been influenced by the South East Asian and the Indian Monsoon systems during the Holocene (see Figure 4.8). To understand the evolution of this peatland and its potential to provide new insights into the Holocene evolution of the East Asian monsoon, a 6 m peat core was collected from the undisturbed

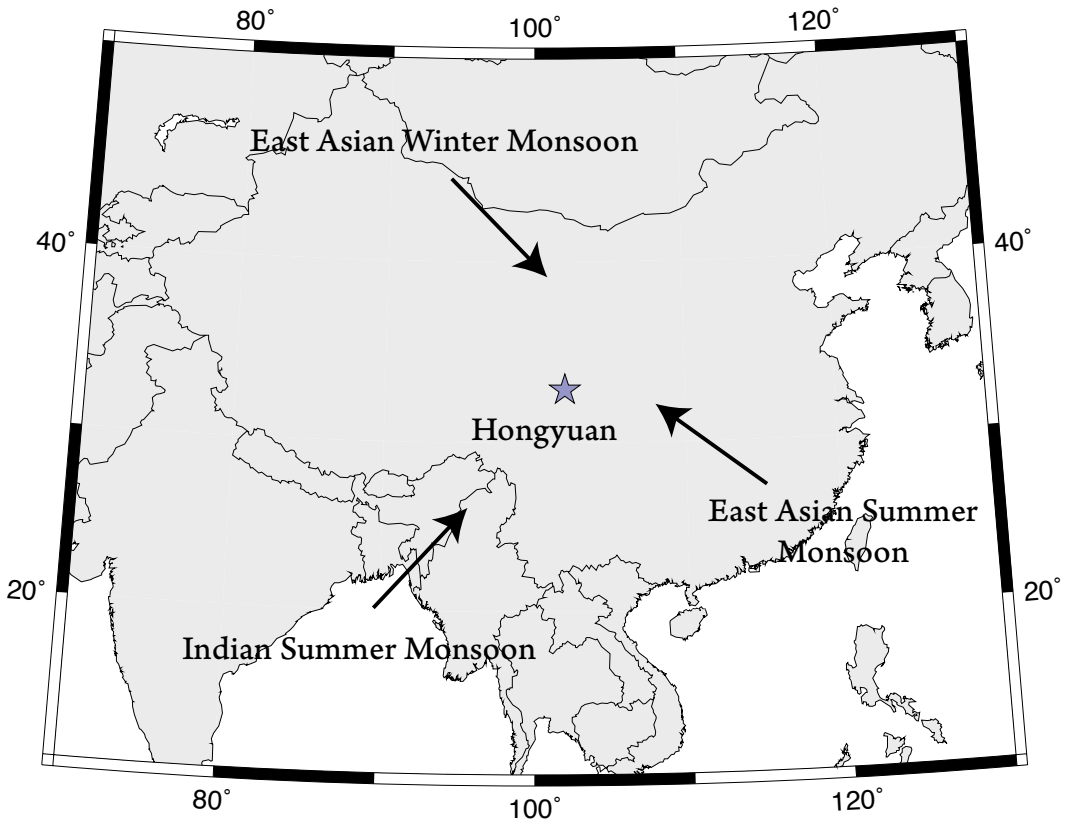


Figure 4.8: Sketch map showing the location of the Hongyuan peat bog (star). Arrows indicate the approximate wind directions associated with the summer and winter monsoons.

central part of a peat deposit near Hongyuan (Large *et al.*, 2009). The peat core was analysed for a range of environmental variables including carbon, nitrogen and hydrogen concentration, bulk density, and $\delta^{13}C$ (see Figure 4.9).

The age-depth relationship of the recovered peat sequence covers the period from 9.6 to 0.3 kyr BP and is linear indicating that the conditions governing productivity and decay varied little over the Holocene. Consequently the parameters measured are assumed to be those of the near surface peat at time of deposition (Large *et al.*, 2009). In this work, we have kept the raw data expressed in depth.

The different depth series are of variable nature, i.e. they have been collected with different instruments and record different responses to climate. The level of data noise might be quite different between records but unfortunately, nothing is known about the measurement uncertainties. The hierachical Bayes formulation is therefore relevant here as it can account for this lack of critical information. The

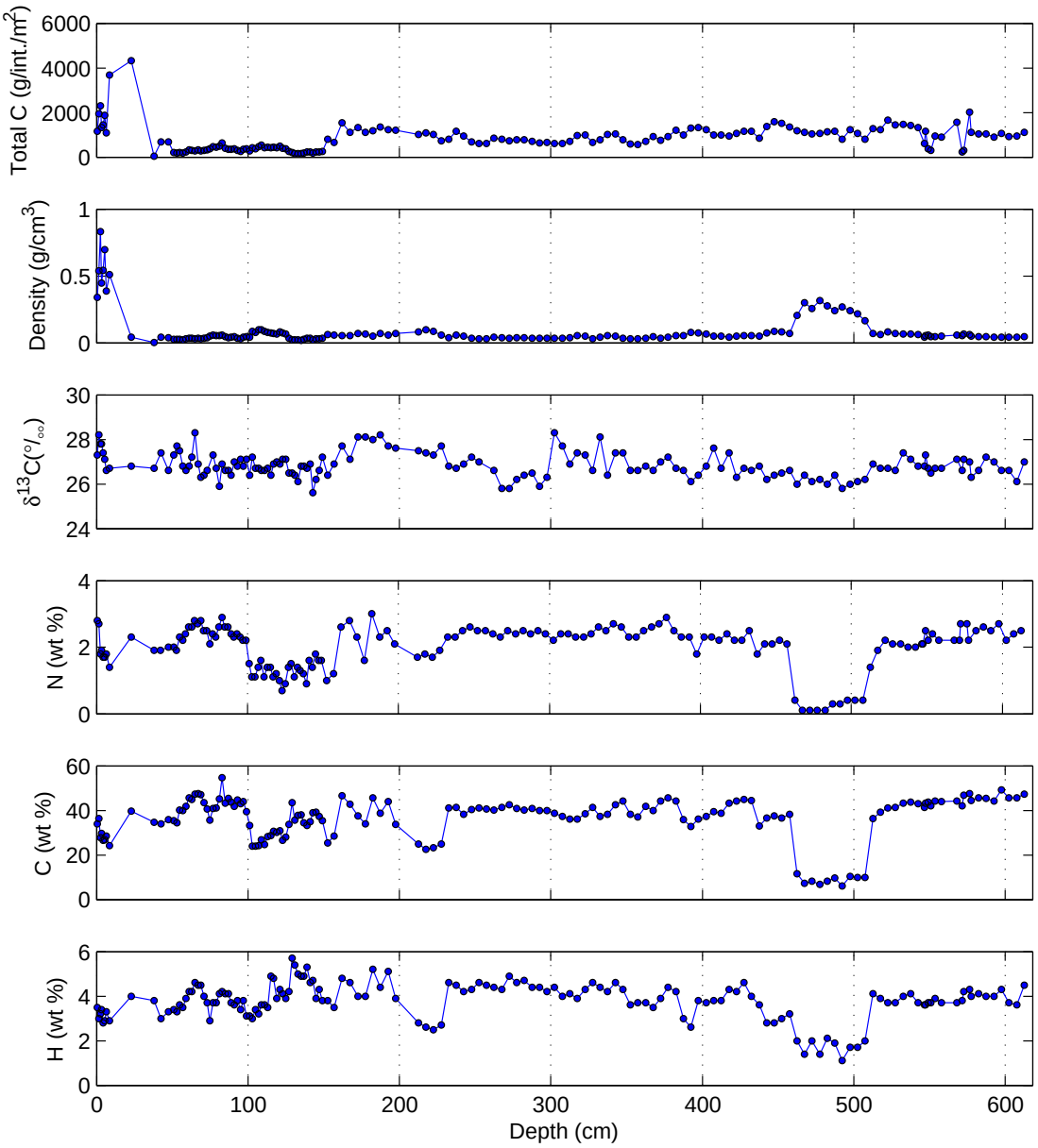


Figure 4.9: Variables measured in the peat core vs.depth.

observations can be expressed as a combined set of m data types:

$$\mathbf{d}_{obs} = \{\mathbf{d}^1, ..., \mathbf{d}^m\} \tag{4.9}$$

Each record $\mathbf{d}^j$ contains M_j data points so the total number of data points in the experiment is $N = \sum_{j=1}^{m} M_j$.

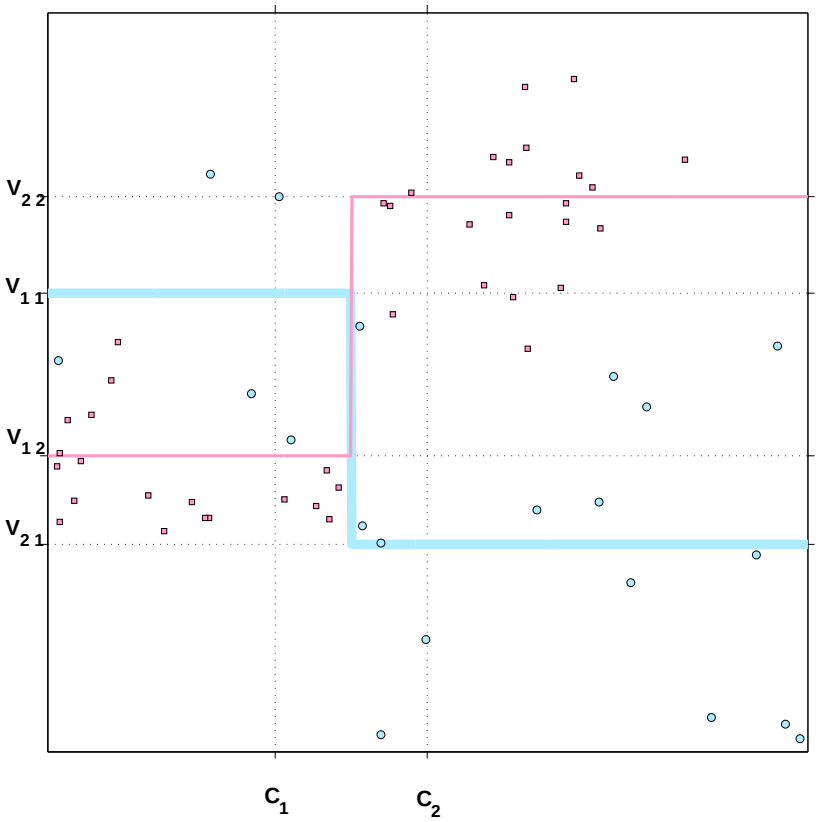


Figure 4.10: In this figure are represented 2 datasets (pink squares and blue circles), and a model $\mathbf{m} = (\mathbf{c}, \mathbf{V}, \boldsymbol{\sigma})$ that tries to fit them. Note that the number of data points is different for each data set (more pink squares than blue circles). The model is defined by two cells given by $\mathbf{c} = [c_1, c_2]$. These are Voronoi cells and the centres are not necessarily at the cell centres. For each cell i and data set j, the model gives an estimated response value $\mathbf{V}_{ij}$. From visual inspection, it can easily be seen that blue data points are more 'noisy' than the pink ones. This is quantified by the two noise hyperparameters $\boldsymbol{\sigma} = [\sigma_1, \sigma_2]$.

4.3.2 Model parameterisation

It is assumed that the underlying trends in the data between the changepoints are constant with depth. As change points here reflect changes in source contribution, all the different records are modelled simultaneously. In doing so, different reponse values for each data type are allowed for within each partition, but require that the changepoints be in the same location for each simulation considered.

The model is partitionned by a set of n Voronoi cells representing periods of time where the airborne material deposited in the peat is derived from a constant source. Each cell is defined by the location of its nucleus c_i with $i \in [1, n]$. This set of nuclei is defined by a vector $\mathbf{c} = [c_1, ..., c_n]$ of variable dimension. Each cell c_i is

assigned a set of response values $\mathbf{V}_{i1}, ..., \mathbf{V}_{im}$, one for each data type (where m is the number of data records). Therefore, the model response values are represented in the $n \times m$ matrix $\mathbf{V}$. Figure 4.10 shows an example of a model with 2 cells in a problem having 2 datasets. In the Hierarchical Bayes formulation used here, the data noise is a model parameter to be determined by the data. Therefore, an estimated value of the data errors is associated with each recorded signal. We represent the combined set of model parameter as $\mathbf{m} = (\mathbf{c}, \mathbf{V}, \boldsymbol{\sigma})$ where $\boldsymbol{\sigma}$ is the vector of the estimated noise standard deviation for each data type ($\boldsymbol{\sigma} = [\sigma_1, ..., \sigma_m]$). Between and within datasets, the data errors are supposed to be uncorrelated so the data covariance matrix is diagonal and $\boldsymbol{\sigma}$ represents the square roots of its diagonal elements. Therefore, the size of the model space is $n + (m \times n) + m$.

4.3.3 Hierarchical Bayes reversible jump algorithm

The Hierarchical Bayes algorithm is used to infer the change points location common to all datasets, as well as a predictive regression function and a noise level for each dataset being considered. Here we jointly invert different datasets, and thus the model and noise parameterisations, as well as the form of the likelihood and proposal distributions differ from the simple regression problem presented in the first section. The implementation of the algorithm is more complex and here we summarise the key points.

4.3.3.1 The likelihood function

The misfit function associated with each data type j takes the form

$$\phi^j(\mathbf{m}) = \left\| \frac{g^j(\mathbf{m}) - \mathbf{d}^j}{\sigma_j} \right\|^2 \tag{4.10}$$

where $g^j(\mathbf{m})$ a vector of the estimated data for dataset j. The total misfit function is given by

$$\phi(\mathbf{m}) = \sum_{j=1}^{m} \phi^j(\mathbf{m}). \tag{4.11}$$

Hence, the multidimensional Gaussian likelihood function takes the form

$$p(\mathbf{d}_{obs} \mid \mathbf{m}) = \prod_{j=1}^{m} \left[(2\pi\sigma_j^2)^{-M_j/2} \right] \times \exp\left\{ \frac{-\phi(\mathbf{m})}{2} \right\}. \tag{4.12}$$

4.3.3.2 Proposal distributions

At each iteration, one type of move is uniformly randomly selected from the 5 following possibilities :

1. Change a response value in a partition. Randomly select a cell i from a uniform distribution over the range $[1, n]$. Randomly select a data set j from a uniform distribution over the range $[1, m]$. Propose a new value $\mathbf{V}_{ij}$ using a Gaussian probability density centred on the current value.

2. Change the estimated data noise. Randomly select a data set j from a uniform distribution over the range $[1, m]$. Propose a new value for σ_j using a Gaussian probability density centred on the current value.

3. BIRTH. Add a new Voronoi nucleus. Its location is drawn from the uniform prior distribution for the nuclei location (see appendix A). Then, m new response values need to be created for the new cell. For each data type, the value is proposed according to a Gaussian probability density with mean and variance equal to the mean and variance of the data points within the cell. If there are no data points in the new cell, an intuitive choice would be to simply reject the proposal. However, in the reverse step, the proposal ratio would prevent a cell containing no data points to be deleted. To avoid this, the reponse value is drawn according to the prior distribution when there are no data points in the new cell.

4. DEATH. Remove at random one cell by drawing a number from a uniform distribution over the range $[1, n]$. The response values of the neighbooring cells remain unchanged.

5. MOVE. Randomly pick one cell and perturb the position of its nucleus according to a Gaussian distribution centred on the current position. The response values of the cells remain unchanged.

Note that in a birth step, the proposal for the response value is made according to the distribution of data in the new cell. This has two advantages. First, it directly drives the random walk towards the mode of the posterior distribution and hence increases the chances of accepting the new model. Secondly, the proposal does not depend on the current model, which simplifies greatly the form of the acceptance term.

Note also that the model parameters are changed almost one by one (i.e. the proposal distributions are only 1D probability functions). When perturbing $\mathbf{V}$, after having chosen a cell, we could have changed all the response values at the same time (i.e. to use a m-dimensional proposal probability function). Instead, only one of the m values is changed. Similarly, for $\boldsymbol{\sigma}$, only one component of the vector is changed at the time. The advantage of perturbing one parameter at the time is that we can separately monitor the acceptance rates for all the parameters, and 'tune' the proposal distribution in a more efficient way. For example, when changing a response value, the variance of the proposal might need to be different between data types. The variance of the proposals are adjusted manually by mean of a 'trial and error' routine in such a way that the acceptance rate for each parameter is around 44%. Indeed, Rosenthal (2008) showed that if all the parameters are perturbed at each step, the optimal acceptance rate is 24%, whereas when only one model parameter is perturbed at the time, it is 44%.

A detailed description of the form of the prior and proposal distribution, as well as the Jacobian and the acceptance rate are given in Appendix A.

4.3.4 Synthetic experiment

This multiple datasets regression method is first tested on a simple problem consisting of $m = 4$ datasets. A synthetic model $\mathbf{m}_{true} = (\mathbf{c}_{true}, \mathbf{V}_{true}, \boldsymbol{\sigma}_{true})$ has been generated with $n = 9$ cells, with different response values in each cell (Figure 4.11). Although the values of the true model in each of the 9 cells differs, they are tied together with common change points. Each dataset has been generated with a different value for the random noise. That is,

$$\boldsymbol{\sigma}_{true} = [2 \quad 4 \quad 6 \quad 8]. \tag{4.13}$$

The locations of data points are irregularly spaced and are different between datasets. We use these data, assuming the noise levels are unknown, to infer the distributions on the number and locations of changepoints, the regression functions and the noise. We first invert each data set separately as in section 4.2.3 and then carry out the joint inversion described above.

4.3.4.1 Individual Inversion of datasets

To demonstrate the influence of different datasets, we ran each data set independently as in section 4.2.3. Results for individual inversions are showed in Figure

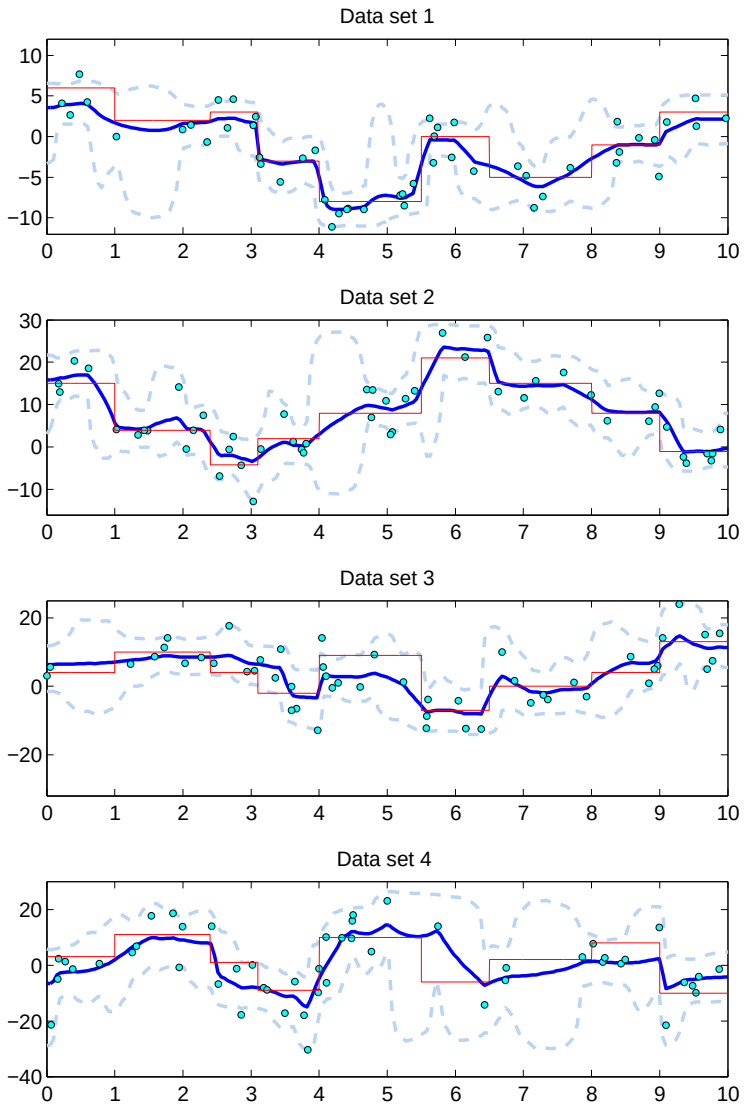


Figure 4.11: Synthetic data experiment. Inversion of individual datasets. The 4 panels correspond to our 4 synthetic datasets. The true model (red lines) consists of four partition models sharing the same Voronoi nuclei but with different responses values. For each of the 4 partition models, 50 data points (blue circles) have been generated. The position in x of each data points is drawn randomly from a uniform distribution over the range $[x_{min} = 0, x_{max} = 10]$. The value y equals the response value given by the synthetic model plus a random Gaussian error which value is unknown in the inversion. Datasets have been inverted individually, i.e. a separate reversible jump algorithm has been applied to each dataset. The blue line is the average solution model and the dashed light blue lines represent the 95% credible interval.The posterior distribution on location of change points for each dataset is shown in Figure 4.12

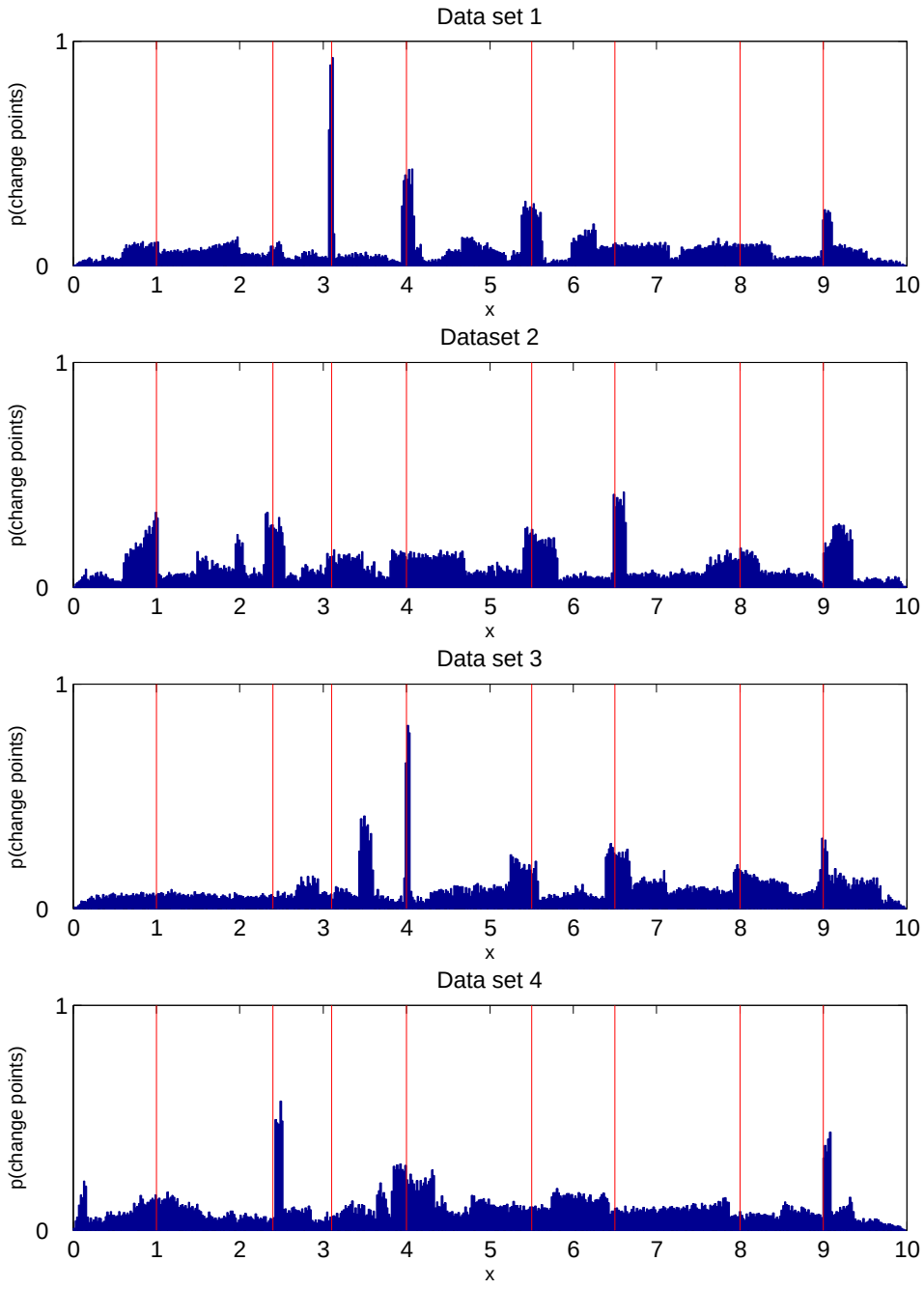


Figure 4.12: Synthetic data experiment. Inversion of individual datasets. Posterior distribution for the location of change points. Red lines show the position of change points in the true model $\mathbf{m}_{true}$. A separate reversible jump algorithm has been applied to each dataset. The 4 panels correspond to our 4 synthetic datasets shown in Figure 4.11.

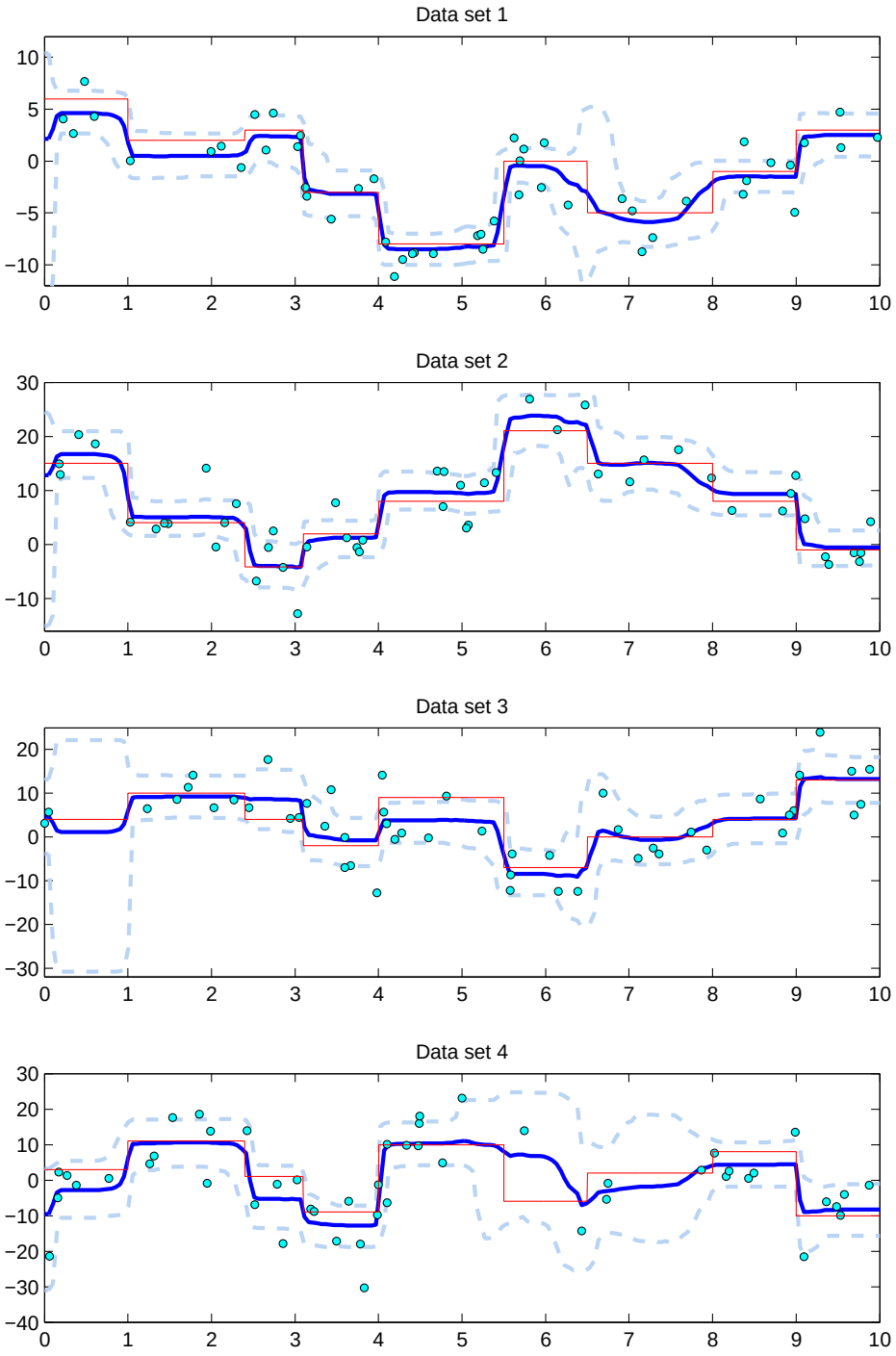


Figure 4.13: Synthetic data experiment. Joint inversion. The 4 panels correspond to our 4 synthetic datasets. The red line is the true model, and blue dots are the data points. The blue line is the average model and the dashed light blue lines represent the 95% credible interval.

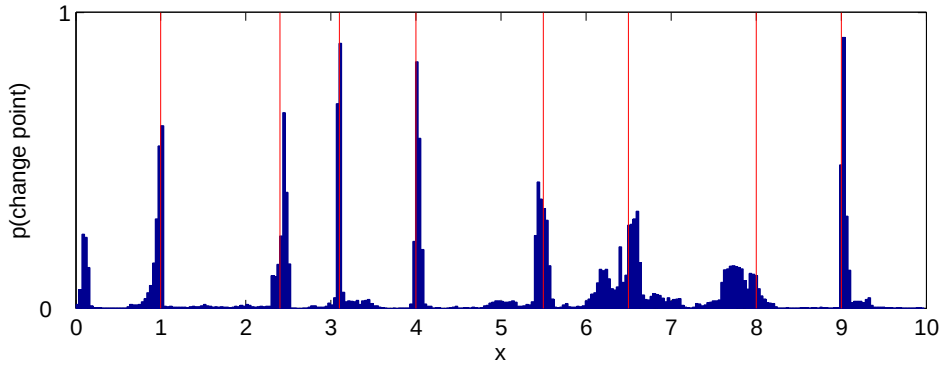


Figure 4.14: Synthetic data experiment. Joint inversion. Posterior distribution for the change points. Red lines represent the true change points in the synthetic model.

4.11. The dashed light blue line represent the 95% credible interval. That is, at each location in x, 95% of sampled models take a value between the lower and the upper bound. As expected, the credible interval increases where there are few or no data. The average solution model is quite smooth and the change points (i.e. cell boundaries) have not really been clearly identified. This can be seen by plotting a histogram of the location of change points obtained for each data set (Figure 4.12). Even if those are not direct model parameters, they give useful information about the true model and are needed for interpretation in real data experiments. Some change points are better constrained in some datasets than in others. For example, the change point at $x = 3$ is well constrained by the first dataset but invisible in the fourth dataset. Conversely, the changepoint at $x = 2.5$ has been identified by the second data set but is missing after inversion of the first dataset. By inverting the 4 datasets together, we expect to get a more precise estimation of change points and therefore of the true model.

4.3.4.2 Joint Inversion

The 4 datasets have been inverted together as described in section 4.3.3 and results are shown in Figure 4.13. Although the values of the regression function for each of the 4 signals differs, they are tied together with common changepoints Clearly, the expected model is closer to the true model than in Figure 4.11 where each dataset has been considered separately. The 95% credible interval shows a smaller range, meaning that the solution model is much more constrained here. This is because when inverted jointly, the data put higher constraints on the common change points which improves the estimation of the function value within each cell.

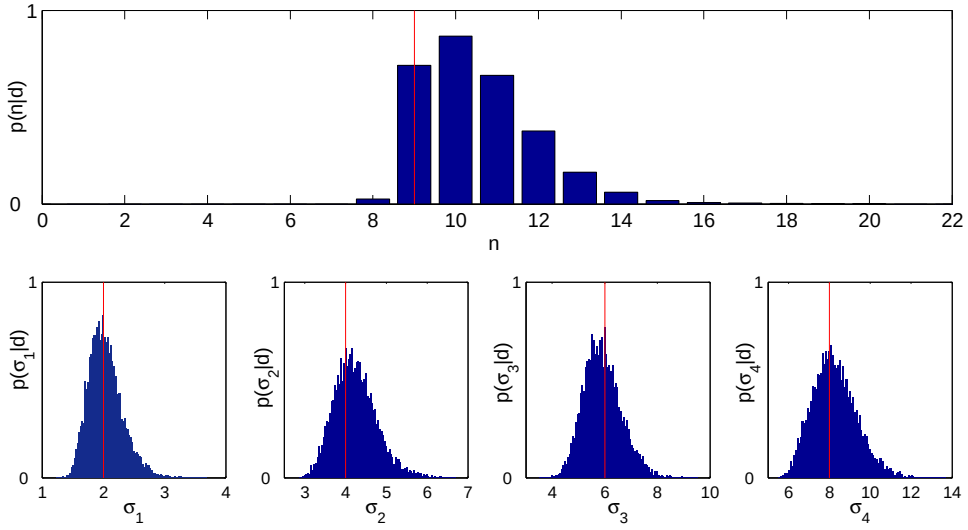


Figure 4.15: Synthetic data experiment. Joint inversion. Marginal posterior distributions for the number of nuclei (upper panel) and for the values of data noise corresponding to each data set (lower panels). Red line correspond to the values of the true model.

Figure 4.14 shows the expected changepoint structure which is also dramatically better resolved here. The joint inversion is able to take advantage of the extra information provided, i.e. that changepoints are common across each of the four signals. This example illustrates the benefit of inverting together different datasets as the knowledge obtained is larger than the sum of information resulting from individual inversions.

The posterior marginal for the hyperparameters $(n, \boldsymbol{\sigma})$ used in our formulation is plotted in Figure 4.15. The first panel shows the posterior distribution for the number of cells. The inference leads to a high probability that there are 9, which is the number of cells in the true model. This is of course, conditional on all the model assumptions (a finite number of discrete changes with constant mean values in each cell), although these are appropriate in this example. In practice, however, it is unlikely that we would be primarily interested in the absolute number of changepoints, but rather in where changes are inferred to occur. Lower panels of Figure 4.15 show the marginal for the four components of $\boldsymbol{\sigma}$. The four datasets have been generated with a significantly different value of noise (from $\sigma_1 = 2$ to $\sigma_4 = 8$) and the expanded Bayes formulation is able to recover the true values which is crucial for an accurate assessment of model complexity.

4.3.5 Results with field data

Here we apply the method to the real data series from Hongyuan showed in Figure 4.9. Figure 4.16 shows regression results for the 6 geochemical series. The average solution curves capture the most significant changepoints, as these will appear in many of the simulations, while less important or poorly resolved change points tend to be smoothed out. Figure 4.17 shows the posterior probability distribution of change points. This enables us to see the depth (and therefore the time in history) at which there have been dramatically changes in incoming airborne material, and how statistically significant are these changes in the data. It is interesting to note that changepoints are better constrained for shallower events, and as we go deeper, changepoints become less constrained.

The summary diagram of Large *et al.* (2009) (their figure 7) compares their data to previous studies, and in particular of inferred periods of cold, dry (permafrost) periods relative to warmer, wetter periods. Thus our inferred changepoints should correspond to times when these conditions switch. Apart from the relatively low amplitude probability changepoint inferred around 200 cm and the recent variations (<50cm, attributed to disturbance as a consequence of Yak grazing by Large *et al.* (2009), the changepoints in Figure 4.17 agree well with those inferred by a qualitative comparison of regional datasets from China by Large *et al.* (2009). The palaeoenvironmental history interpreted from these data can then be linked to the climate variations in northwest Pacific, the El Niño-Southern Oscillation, movement of the Intertropical Convergence Zone and the East Asian Monsoon.

Figure 4.18 also shows the posterior marginals for the hyperparameters (number of cells and data noise). Although the expected values of the hyperparameters are of little use for interpretation, they reveal statistical characteristics of the inverse problem. For example, it can be seen that the posterior expectation of the data noise is different according to data types. When data are normalised to have 0 mean and unit variance, the level of noise for the density record is expected to be around 0.15 whereas for $\delta^{13}C$ it is around 0.75, meaning the first record is 5 times less 'noisy' than the second. The expected number of cells shows the model complexity required to fit the data. It implies 24 change points while Figure 4.17 has 13 except the cluster at shallow depth. Again, this concentration of shallow changepoints might result from data being perturbed by Yaks.

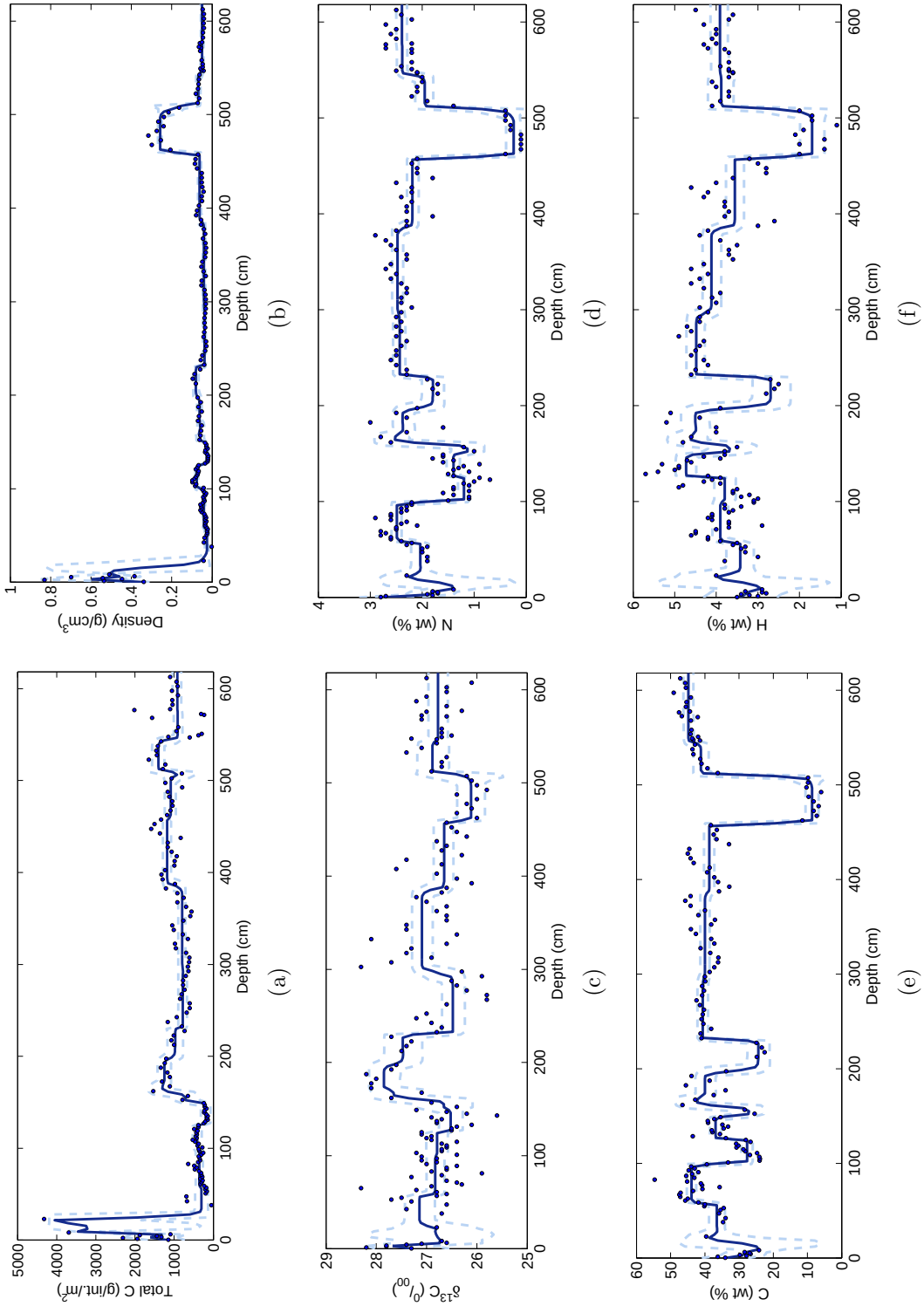


Figure 4.16: Real data experiments. Joint Inversion. Colours and lines are as in previous figures.

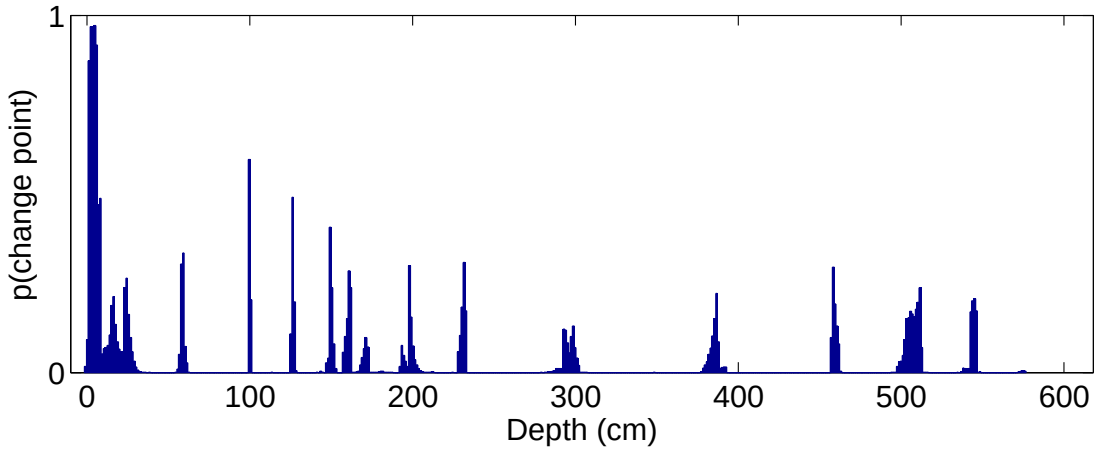


Figure 4.17: Real data experiments. Joint Inversion. Posterior distribution for the change points.

4.4 Discussion

In an inverse problem, it can be difficult to assess the correct number of unknowns needed to fit the data to the level required by the estimated data noise. However, it has been shown that the reversible jump algorithm naturally adapts the model complexity in order to fit the data to the adequate level. While this property can be seen as an advantage, the posterior solution strongly depends on the estimated data uncertainty, and hence this can be a problem if the user knows little about the measurements errors. Therefore, we have presented a method to expand the conventional Bayesian accounting of data uncertainty. In this expanded formulation, the likelihood function is defined by a hyperparameter (the standard deviation of the measurement noise) that need not be fixed but has a prior uncertainty itself. This noise parameter is 'free' during the inversion process and its posterior probability distribution is dictated by the data. We observe that this hyperparameter not only takes into account the measurements errors, but also the error present in the forward theory.

In order to illustrate results, we have applied the algorithm to a simple 1D regression problem. However, the reversible jump regression method turns out to be particularly relevant for change point modelling of palaeoclimate time series. Changepoints can be defined as abrupt changes in the mean signal, over depth or time. That is, the regression function is defined in terms of a constant value between two changepoints (i.e. a 1D equivalent to our Voronoi seismic velocity models).

Using synthetic data, we have shown that we can recover the changepoint struc-

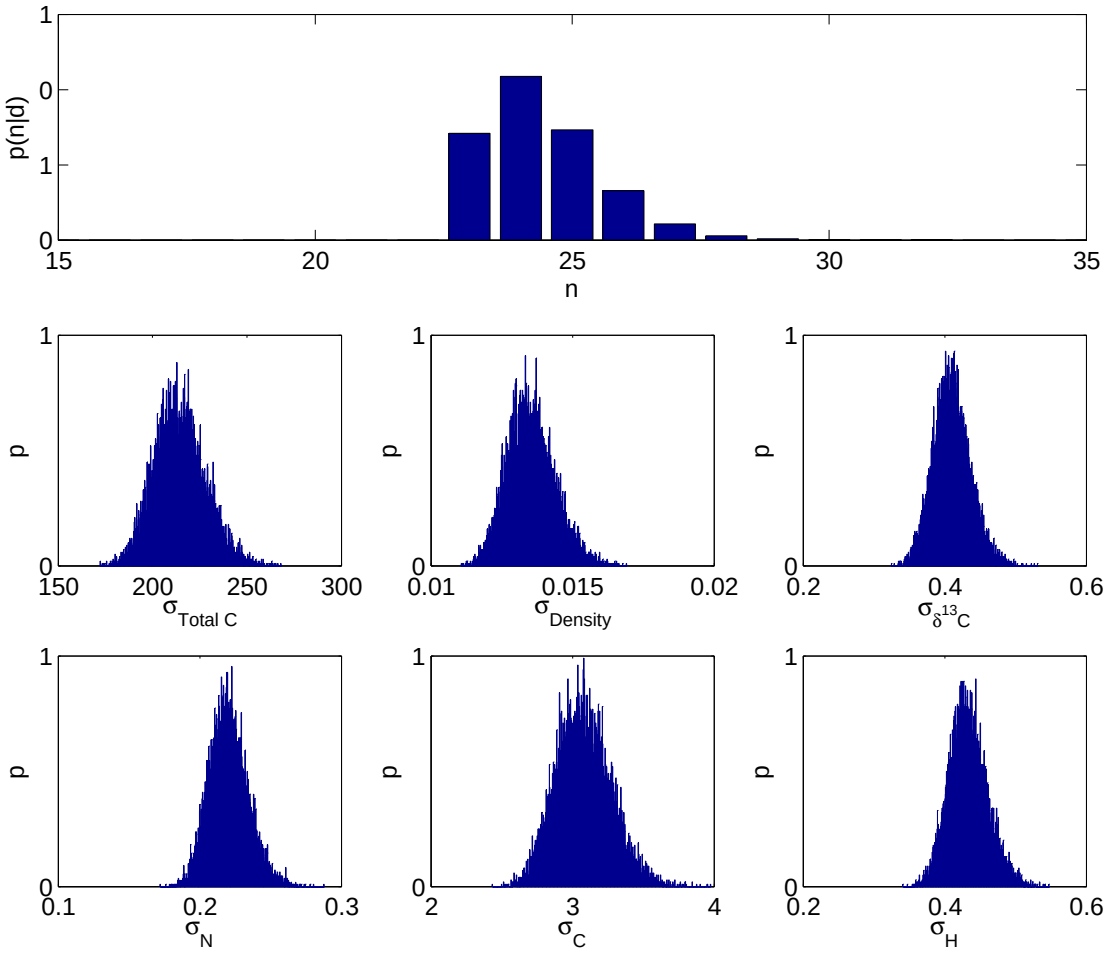


Figure 4.18: Real data experiments. Joint Inversion. Normalised data. Marginal posterior distributions for the number of nuclei (upper panel) and for the values of data noise corresponding to each data set (lower panels)

ture and the noise level reliably. When dealing with multiple datasets, we assume that all datasets contain the same changepoint locations, but the response, and noise level are different. This allows us to infer a changepoint structure, common to all datasets, in terms of distributions for number and location of changepoints, as well as the the regression function and the noise level for each data type.

The approach we present can be generalised readily to allow for different noise levels between partitions, if required. Additionally, the different datasets can be irregularly spaced in depth (or time) and there is no need for the data to be sampled at the same depths. The details of the solution will depend on which datasets are used (i.e. singly or jointly) and we recommend using joint modelling if the assumption of common changepoints is considered valid. This assumption is perhaps best as-

sessed from the understanding of geochemical behaviour in different environmental systems. Certainly, the results are more coherent and generally easier to interpret than by combining results from individual dataset modelling.

Application of the method to real datasets from peat core in Eastern Tibet provide results in agreement with previous qualitative interpretations. However, our approach provides probability distributions on all parameters and so allows us to assess quantitatively the relative importance of the inferred change point structure. The method is clearly an effective tool for identifying changes in the source regions for airborne material and to the study of climate change using such deposits.

Directions for future work would be to consider transdimensional regression functions (for example we would estimate the order of a polynomial in each partition and these could be different between partitions) and to allow for uncertainty in depth to age conversions (which is likely to be important when comparing records from different locations either regionally or globally).

Chapter 5

Multiscale Seismic Tomography With the Hierarchical Bayes Methodology

5.1 Introduction

As shown in chapters 2 and 3, the main advantage of reversible jump tomography is to have a parameterisation that is able to naturally adapt to the distribution of information provided by the data but also to the underlying structure of the earth. The parameterisation not only adapts to the density of rays but also to the structures to be imaged (e.g. discontinuities). For example, the parsimonious nature of the approach means that in a large homogeneous region of constant wave speed, it will result in a single large cell, even if the ray coverage is dense there (a feature that will be shown in this chapter). This is why reversible jump tomography can be particularly useful to a problem with multiple scales, i.e where the spatial sampling of the velocity field is highly heterogeneous or where the velocity field itself has variables scale lengths.

It is known that fixed grid optimisation schemes that use regularisation procedures give limited results in such situations. The grid size required to resolve densely sampled areas may introduce small-scale artefacts in regions where the velocity field is much less constrained. This effect is usually avoided by adding non data-driven constraints on the model. However, smoothing and damping procedures are global (responsive to the entire model) and while averaging over large areas, they make resolvable small-scale details difficult to see, or even hide them all together. Sharp discontinuities present in the model are typically blurred by smoothing processes

into gradual transitions.

To avoid these problems, the use of wavelet decomposition in seismic tomography is becoming popular. Chiao and Kuo (2001) used Harr wavelets on a sphere to parameterise a 2D model of shear wave velocities in D". They inverted S-SKS travel times and showed that the wavelet provides a natural regularization based on density of rays. Tikhotsky and Achauer (2008) also used Haar wavelet to build a 3D model constrained by both controlled source seismic and gravity data. Loris *et al.* (2007) used a more complex wavelet parameterization and minimized an L1-norm measure of the wavelet coefficients in order to produce a solution model that only has details where required by the data. However, their method is an optimization scheme that requires tuning of a regularization parameter and the complexity of the solution model (i.e. the norm of the wavelet coefficients) fully depends on the value given to this parameter by the user. Many other data-driven adaptive parameterization methods exist in the literature and are presented in chapter 1.

As an alternative to wavelet parameterization, in this chapter we address the multiscale nature of seismic tomography with the reversible jump algorithm. We simultaneously invert different datasets that span the Australian crust at various scales. The Rayleigh wave group traveltime dataset from Saygin and Kennett (2008) presented in chapter 3 is inverted in conjunction with similar cross-correlation travel times obtained from the WOMBAT experiment which sample the southeast part of the continent with a much higher resolution (Rawlinson *et al.*, 2008; Arroucau *et al.*, 2009, 2010). The datasets used here are different both in size (i.e. the number of picked travel times) and in the scale of the region they sample (see Figure 5.1). Although considerable detail might be resolvable for densely sampled areas, only relative large-scale information is available for sparsely sampled regions. The idea is to see whether additional information can be revealed by simultaneously inverting some data that can't conveniently be inverted together by a standard approach.

In chapter 3, the reversible jump tomography methodology was presented and tested on a simple synthetic problem. Then, in order to show the feasibility of the method in seismic tomography, an application to ambient noise data was also presented. Applying the method to a real dataset gave us additional insights into the mechanisms involved in the inversion. Indeed, we found out that the complexity of the sampled models (i.e. their number of cells) was extremely dependent on the estimated level of data uncertainty given by the user. For example, when the data noise is under-estimated, the required level of data fit is then over-estimated and the algorithm automatically adds more cells into the model and provides a solution

that is too complex with an artificially high data fit. As seen in chapter 4, this is because in a transdimensional Bayesian inversion, the level of data noise directly determines the required level of data fit, and hence the number of model parameters in the inversion.

In this chapter, the data errors are assumed to be independent and normally distributed with zero mean. Hence, each time we use the phrases 'level of noise', 'magnitude of errors', or 'data uncertainty', we always refer to the standard deviation of an independent random Gaussian noise.

As shown in the previous chapter, most tomographic models obtained with optimization schemes do not directly depend on the magnitude of data noise but only on relative uncertainties. This is why the definition of errors in seismology has been rather qualitative than quantitative. For example, Steck *et al.* (1998) used relative weights (1,2,3 and 4) to describe data noise. However, a Bayesian method requires absolute uncertainties and it therefore becomes essential to have reliable estimates of the noise magnitude prior to the inversion.

Seismologists have evaluated seismic travel times between pairs of stations in correlations of ambient seismic noise and tomographically constructed impressive maps of seismic wave velocity (Shapiro *et al.*, 2005; Sabra *et al.*, 2005; Yao and Van der Hilst, 2009; Yang *et al.*, 2006; Villasenor *et al.*, 2007; Brenguier *et al.*, 2008; Arroucau *et al.*, 2010). Although such techniques are now well established and widely used, the uncertainties associated with these travel time are still poorly understood. While recent studies propose to quantitatively measure the bias due to different effects in ambient noise measurments of surface waves dispersion curves (Weaver *et al.*, 2009; Yao and Van der Hilst, 2009; Hubans *et al.*, 2010), this is ongoing work and there is presently no consensus on the way to quantify errors.

In this multiscale study, inter-station distances, ambient noise recording period and processing techniques differ with each dataset. Therefore the uncertainty on the computed travel times might vary considerably between datasets as well as within a dataset.

Without any reliable estimation of data noise, it appears difficult to directly use the conventional reversible jump algorithm described in chapter 3. Hence, we propose to apply the Hierarchical Bayes formulation developed in chapter 4 to perform multiscale tomography of the Australian continent that does not depend on noise estimates given by the user.

The purpose of this chapter is twofold. First, we show that the reversible jump algorithm can be used to jointly invert different datasets in a multiple scale tomog-

raphy problem. Secondly, we show that using a broad prior distribution for the data uncertainties and letting the data infer its own level of noise is not only relevant but indispensable.

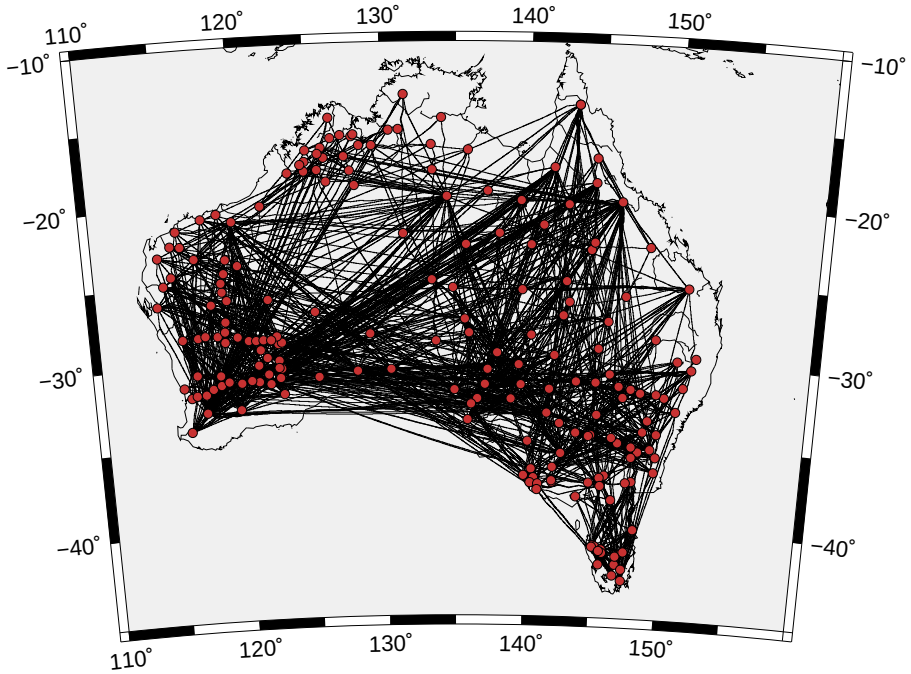
We first present our three ambient noise datasets, and then establish the necessity of the Hierarchical Bayes formulation. A synthetic experiment is presented before simultaneously inverting the three real datasets to construct a detailed map of Rayleigh wave group velocity at 5 s for Australia.

5.2 The data

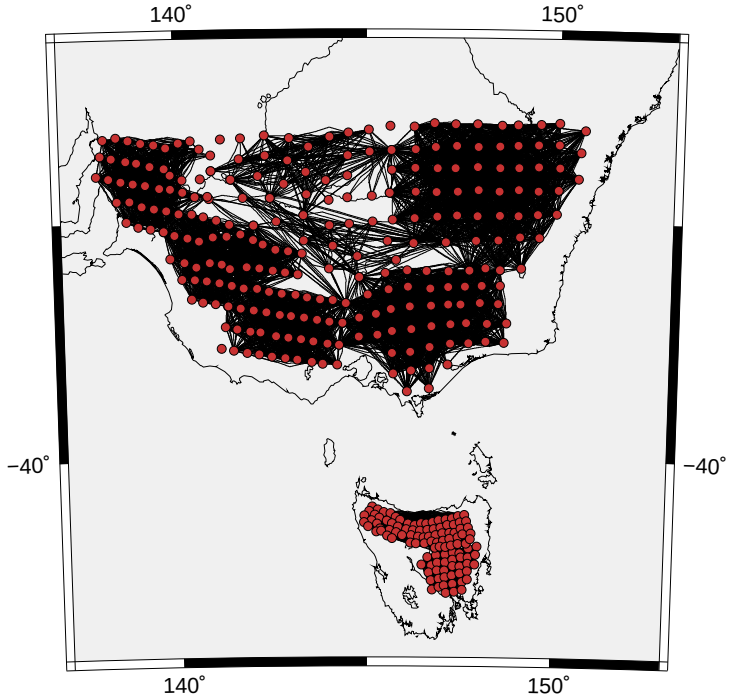
We use the 5 s Rayleigh wave group traveltimes dataset computed from ambient noise by Saygin and Kennett (2008) (this dataset will be referred as the large scale dataset) in conjunction with similar cross-correlation travel times obtained from the WOMBAT experiment (see Figure 5.1). WOMBAT is an extensive program of temporary seismic array deployments throughout southeast Australia and Tasmania. Each array consists of between 30 to 60 short period instruments that continuously record for between five to ten months. Over the last decade, a total of over 500 sites have been occupied resulting in a very large passive seismic dataset that has been used for several studies (e.g. Graeber *et al.*, 2002; Rawlinson *et al.*, 2006; Rawlinson and Urvoy, 2006; Clifford *et al.*, 2007; Rawlinson and Kennett, 2008). Recently, Arroucau *et al.* (2010) used the large volume of recorded noise, which comes from diffuse sources of seismicity such as oceanic or atmospheric disturbances, to construct travel times between stations that recorded simultaneously. Black lines in Figure 5.1 join station pairs for which a travel time is available.

Arrays were not deployed at the same time in Tasmania and in southeast Australia. As a consequence, there are no apparent travel times available between Tasmanian and mainland stations. Furthermore, the average interstation distance for the mainland arrays is about 50 km, while it is only 15 km in Tasmania. Hence the WOMBAT data can be divided into two subsets corresponding to 2 separate regions with different scales: southeast Australia and Tasmania.

Arroucau *et al.* (2009, 2010) performed ambient noise tomography using the WOMBAT arrays (Fig. 5.1(b)). The technique used to compute the cross-correlation of the vertical component of the background noise was slightly differrent from the one used for the large scale dataset by Saygin and Kennett (2008). Rayleigh wave group traveltimes were determined from the obtained cross-correlograms in a two-stage approach. In the first stage, preliminary dispersion curves for periods ranging



(a)



(b)

Figure 5.1: Ray coverage for the 3 datasets sampling the Australian crust at different scales. Black lines indicate rays between station pairs (red circles) that recorded ambient noise at the same time and for which a travel time has been measured. Top: Large scale dataset (Saygin and Kennett, 2008). Bottom: Wombat Experiment that can be divided into two sub-arrays (southeast Australia and Tasmania)

from 1 to 20 s were constructed and averaged in order to build a phase-matched filter which was subsequently applied to the seismograms prior to a second round of traveltime picking. Relative uncertainties for the picked travel times were obtained following the procedure presented by Cotte and Laske (2002) and Harmon *et al.* (2007). We note here that there is no information available on the data uncertainty for the large scale dataset in Figure 5.1(a).

The travel times obtained using WOMBAT arrays were used to generate tomographic maps for each frequency with the subspace method (Kennett *et al.*, 1988; Rawlinson *et al.*, 2006). Due to the difference in ray densities, data from mainland arrays and Tasmania were inverted separately with two different grid sizes. The internode distance for mainland regions (20 km) was four times as large than for Tasmania (5 km). Furthermore, the regularization parameters (smoothing and damping) were different for the two inversions (Arroucau *et al.*, 2009) .

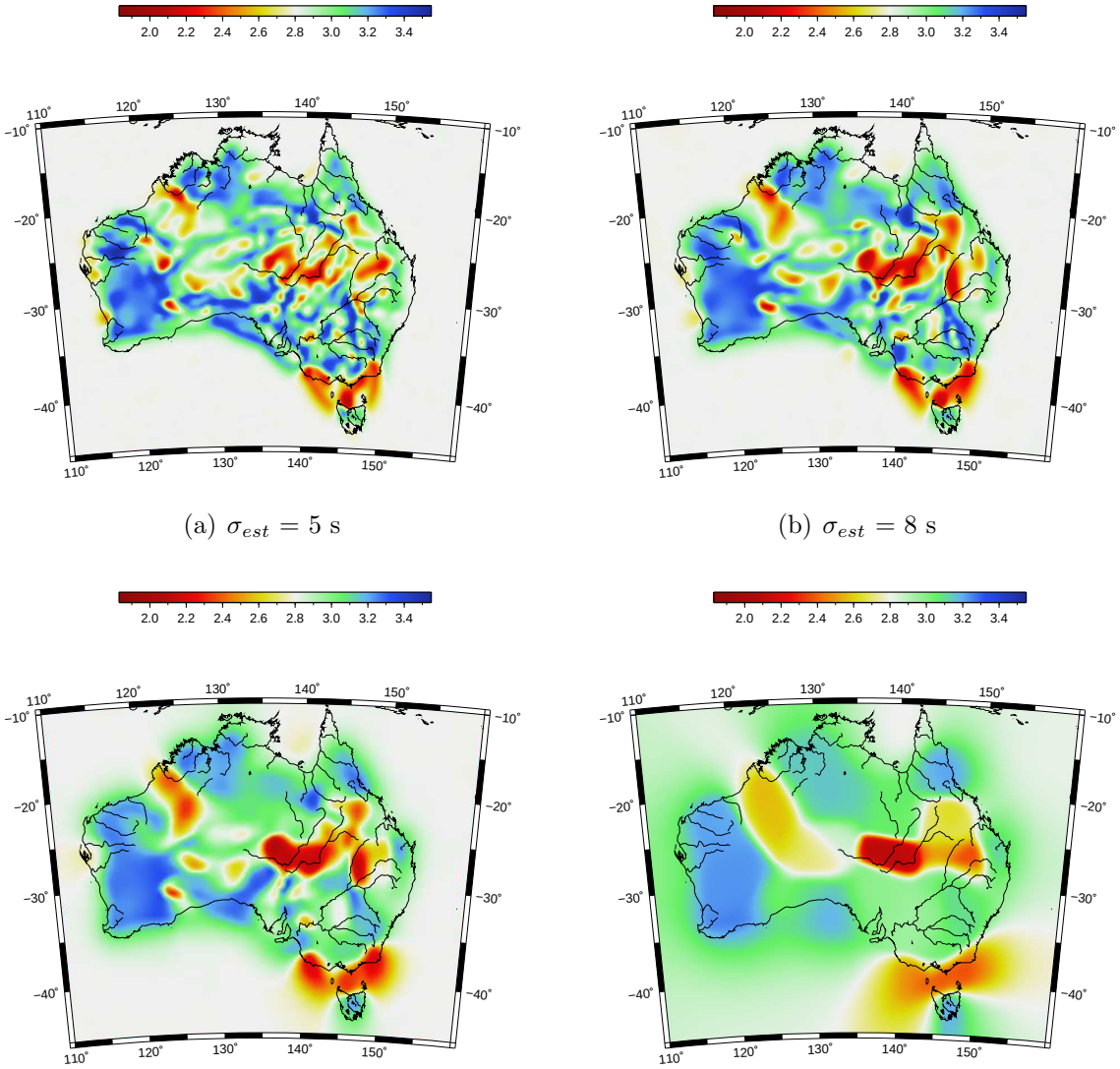
We carry out a joint inversion of these three datasets which have different scales and have been processed separately. Each dataset samples a particular region (whole of Australia, southeast Australia and Tasmania) and is characterised with a scale length defined by its average ray length (respectively 200km, 50 km, 15km). We note that, while the two WOMBAT datasets independently sample two separate regions (Figure 5.1(b)), the largest scale dataset spans the whole country and includes areas sampled by the smaller arrays (Figure 5.1(a)). Therefore, the joint inversion should reveal more information because some areas are sampled simultaneously by different datasets.

5.3 Necessity of Hierarchical Bayes

In chapter 3, the largest scale dataset was inverted with the conventional reversible jump algorithm. The data errors were supposed to be independent and normally distributed with zero mean and a fixed standard deviation σ_{est}. Thus the likelihood probability distribution is

$$p(\mathbf{d}_{obs} \mid \mathbf{m}) = \frac{1}{(\sqrt{2\pi}\sigma_{est})^N} \times \exp\left\{ \frac{-\left\| g(\mathbf{m}) - \mathbf{d}_{obs} \right\|^2}{2\sigma_{est}^2} \right\}. \tag{5.1}$$

The model vector $\mathbf{m}$ has a variable number of components, and as seen previously, the number of Voronoi cells needed to explain the data is directly determined by the estimated level of data noise σ_{est}. The level of data uncertainty effectively quantifies the information present in the data and here it naturally determines the quantity



(a) $\sigma_{est} = 5$ s

(b) $\sigma_{est} = 8$ s

(c) $\sigma_{est} = 12$ s

(d) $\sigma_{est} = 20$ s

Figure 5.2: Conventional reversible jump tomography with ambient noise cross correlation for the Australian continent (large scale dataset). Average solution model (kms^{-1}) obtained with four different (but realistic) values for σ_{est}.

of information that consequently should be present in the model. Since the level of detail in the solution is automatically adjusted as a function of the required data fit, this has been seen as an advantage over optimization based inversions where the level of data noise is not accounted for in the inversion and the level of smoothing manually adjusted by the user.

However, in surface wave dispersion analysis and ambient noise cross-correlation techniques, assessment of measurements errors are not straightforward (Bensen

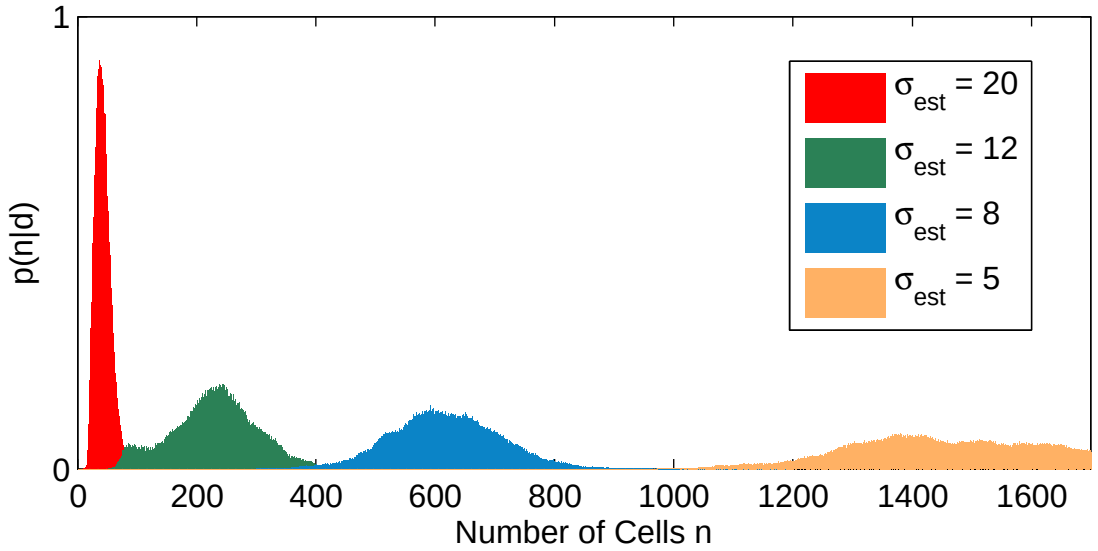


Figure 5.3: Posterior distribution on the number of Voronoi cells for four values of estimated level of data noise. As estimated uncertainty on data decreases, the complexity of sampled model increases.

et al., 2007). Here the large scale dataset has been inverted with the conventional reversible jump algorithm and the solution maps obtained for four different, yet reasonable, values of σ_{est} are shown in Figure 5.2. The posterior distribution on the number of cells for each solution is shown in Figure 5.3. There is clearly a strong correlation between σ_{est} and the number of cells of the sampled models. As the estimated data error decreases, more cells are added in the model and the solution maps show more complexity. The number σ_{est} effectively behaves as a smoothing parameter. In absence of information about the data noise, it is impossible to give a preference to any of these four solutions. The details appearing in 5.2(a) could be unreal and noise induced and data might be 'over-fitted'. Conversely, in Figure 5.2(d), data might have been considered too noisy and uninformative and consequently the solution model might be missing some detail. This shows that by choosing σ_{est}, we also choose the model complexity.

By adding the Hierarchical Bayes formulation to the algorithm and by treating σ_{est} as an unknown, we let the data infer its own degree of uncertainty without imposing any fixed value for the required data fit. In Figure 5.2 the model complexity is spatially variable but the algorithm operates within a fixed data noise level. By freeing up this constraint and treating the data variance as an unknown, we allow the overall model complexity level to be driven by the data.

In our multiscale problem, data noise may vary between datasets resulting from

variations of spectral and azimuthal characteristics of the noise field on different regions and at different scales. Furthermore, travel times have been computed differently between datasets. The Hierarchical Bayes formulation can account for this by independently treating the uncertainty on each dataset. That is, the level of noise of the three data type can be treated as an individual parameter to be inverted for. In this way, the inversion procedure can discriminate between datasets that do not have the same level of uncertainty and information provided by each dataset is naturally weighted.

5.4 Synthetic test

5.4.1 Experimental setup

A synthetic multiscale checkboard velocity model (Fig. 5.4) is constructed in which the square size is approximatively proportional to the spatial sampling of the data shown in Figure 5.1. The three regions of equal area A, B and C are examples of regions with different square size. The areas in orange have a velocity of 2.5 km/s and the green have a velocity of 3.1 km/s which makes the tomographic problem fairly non-linear. Three synthetic travel time datasets corresponding to the same configurations as shown before are constructed by using the Fast Marching Method (Sethian and Popovici, 1999; Rawlinson and Sambridge, 2004).

As explained in detail below, we think that it is realistic to have smaller uncertainties at shorter interstation distances. Consequently, some random Gaussian noise has been added to the synthetic travel times with a standard deviation of 4 s for the large scale dataset and of 1 s for the WOMBAT arrays. The hierarchical formulation will attempt to recover these noise values along with the model and its complexity.

5.4.2 Data noise hyperparameters

We use the Hierarchical Bayes formulation and invert for two noise hyperparameters: σ_1 for the large scale dataset and σ_2 for the WOMBAT arrays (southeast Australia + Tasmania). Hence, the likelihood function takes the form:

$$p(\mathbf{d}_{obs} \mid \mathbf{m}) = \frac{1}{\prod_{i=1}^{N}(\sqrt{2\pi}\sigma_i')} \times \exp\left\{\sum_{i=1}^{N} \frac{-(g(\mathbf{m})_i - \mathbf{d}_i)^2}{2(\sigma_i')^2}\right\} \tag{5.2}$$

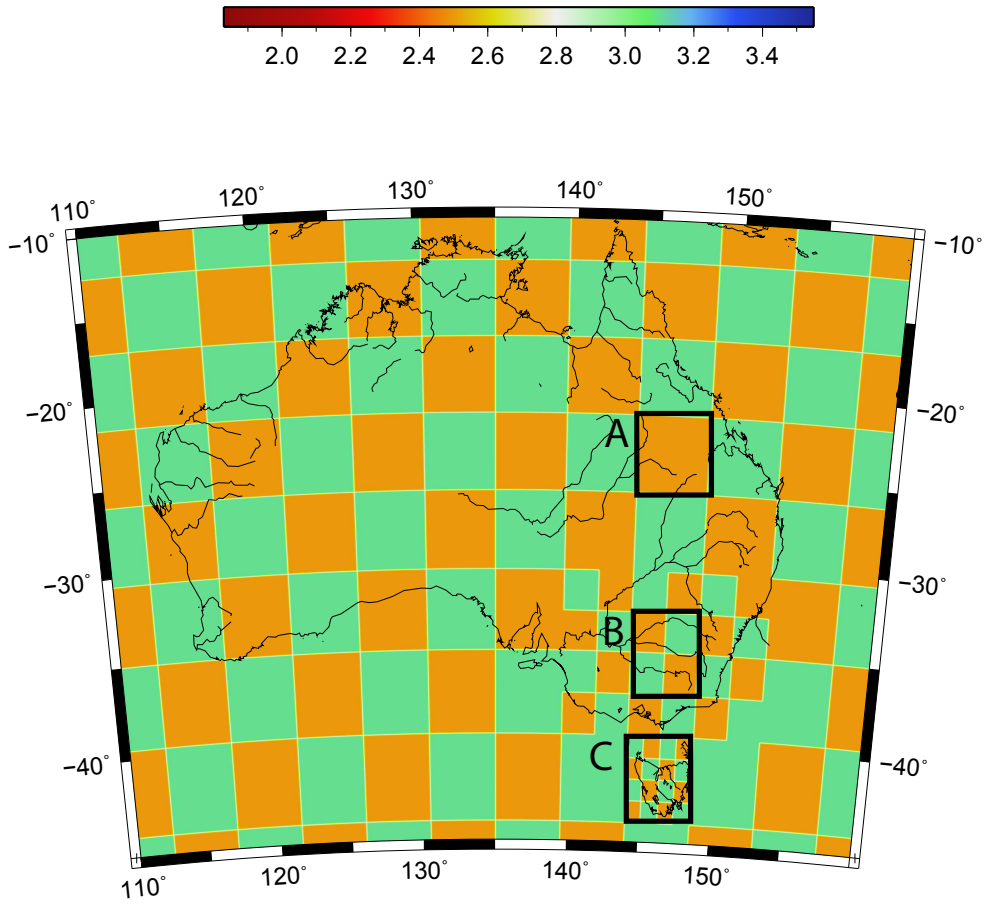


Figure 5.4: Synthetic velocity model (km/s) with a multiscale resolution.The three region A, Band C have equal area but a different complexity of the velocity structure.

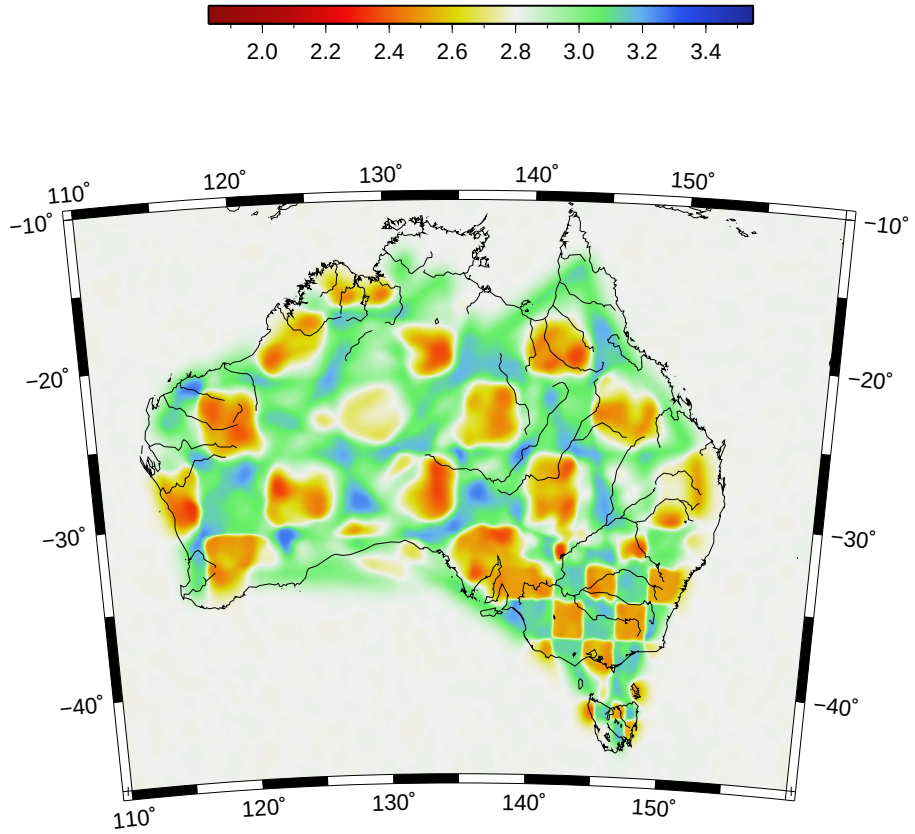
where

$$\begin{cases} \sigma_i' = \sigma_1 & \text{for data belonging to the largest scale set.} \\ \sigma_i' = \sigma_2 & \text{for WOMBAT data.} \end{cases} \tag{5.3}$$

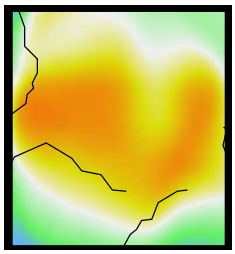
The model parameters $(\mathbf{m}, n, \sigma_1, \sigma_2)$, where n is the number of Voronoi cells, will be successively perturbed along the Markov chain in order to collect an ensemble of models that samples the posterior probability distribution.

5.4.3 Results

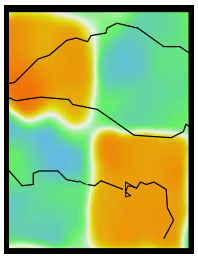
Posterior inference was made using an ensemble of around about 4×10^5 models. A total of 96 Markov chains were run in parallel. Each chain was run for 2×10^6 steps. The first million were discarded as burn-in steps, only after which the sampling



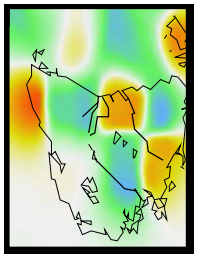
(a)



(b) Region A



(c) Region B



(d) Region C

Figure 5.5: Average solution model obtained with the Hierarchical Bayes reversible jump algorithm by jointly inverting the three datasets. The three lower panels show details for Regions A, B and C.

algorithm was judged to have converged. Then, every 250^{th} model visited was taken in the ensemble. Three passes were made around the 'outer loop' of the reversible jump algorithm with an update of the ray geometry for each pass (see Figure 3.2 of chapter 3).

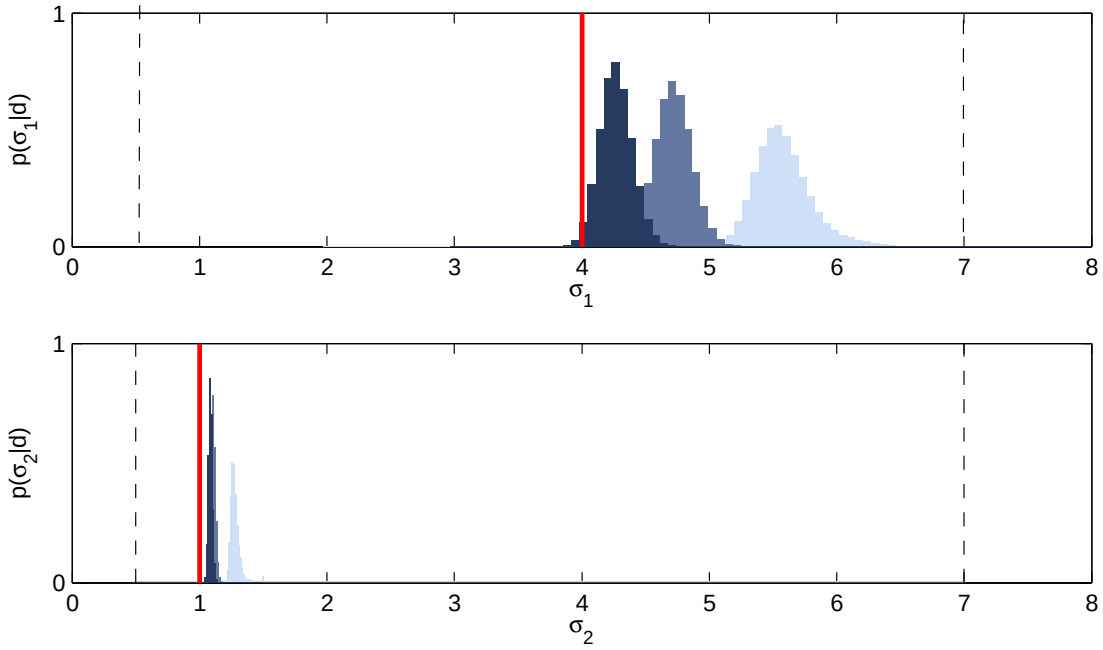


Figure 5.6: Posterior distribution for hyperparameters σ_1 (top) and σ_2 (bottom) for three 'outer-loop' iterations. The Posterior distribution after the first iteration is shown in light blue, second iteration is in blue, and third iteration in dark blue. Red lines indicate the true values of noise added to the data. The bounds of the uniform prior distribution on the noise parameters are shown with dashed lines.

The spatial average of the post burn-in samples collected during the last iteration is shown in Figure 5.5 and clearly recovers features of the true velocity field at different scales. By retrieving the different sizes of the chequerboard squares, the parameterisation has been able to adapt to the underlying structure of the model as well as to the spatial distribution of rays, overcoming the effects of a global regularization procedure. The solution model varies smoothly in regions of limited coverage like region A without sacrificing any sharp or small-scale features in well-sampled regions like region C.

The uniform prior distributions on both hyperparameters σ_1 and σ_2 were defined over the range [0.5 7] s. Their posterior distribution after each 'outer-loop' iterations are shown in Figure 5.6. For the last iteration, the posterior maxima are close to the true value of data noise (relative to the wide range of the prior). The posterior maxima for σ_1 is about four times larger as σ_2 which is the ratio between the two true values of data noise. Note that the posterior maxima for both hyperparameters would be closer to the true value if a Jeffreys prior (i.e. $1/\sigma$)was used instead of a uniform prior. This is because the standard deviation of the noise is a scale

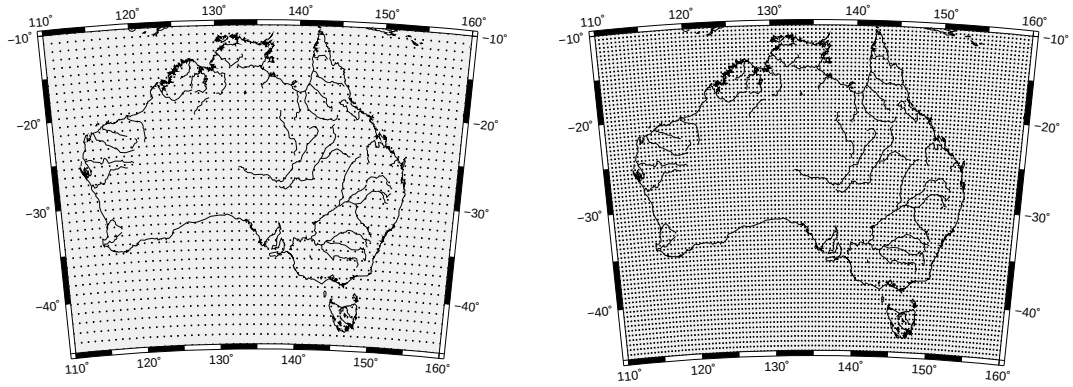
parameter, and a Jeffreys prior would be the proper non-informative prior to use. (Note that the same is true for the number of model parameters). However in this thesis, for simplicity matters, we always use uniform priors for hyperparameters.

With scant information on the data noise prior to the inversion, The Hierarchical Bayes procedure recovers the standard deviations of the true data noise (4 s and 1 s) and therefore provides a parsimonious solution model with a complexity and resolution that varies spatially and that locally conforms to the level of information provided by the data. Indeed, the low gradients in regions sampled by the largest dataset (e.g. region A in Figure 5.5) indicate that observations are less fitted there than under the WOMBAT arrays (e.g. regions B and C) where the discontinuities are better recovered. This can be seen quantitatively in Figure 5.6. Notice that inferring the level of data noise for each dataset is tantamount to inverting for the weighting factor between datasets in a joint inversion. This is because the estimated noise given to a particular dataset directly weights the contribution of this dataset to the total data fit.

5.4.4 Hyperparameters and uncertainty on the forward model

In seismic tomography, incorrect estimates of ray geometries implies an incorrect forward model, i.e. an incorrect function g in equation (5.2). This is because by keeping fixed ray paths, the tomographic problem is linearised in slowness around a reference model (see chapter 3). This linear approximation on the forward model may also contribute to the misfit. For example, let us imagine a tomographic problem where the data are perfect (no noise) but where the ray geometries used in the inversion are completely wrong. In this case it would be impossible to perfectly fit the data and the data misfit would not come from data errors (which are null) but from incorrect ray geometries. Therefore, the data misfit is not only due to the measurements uncertainty but also to the uncertainty on the forward model (i.e. the function g). In an inverse problem the data noise is inevitably defined as the difference between the observed and the predicted measurements $(\mathbf{d}_{obs} - g(\mathbf{m}))$. Therefore, the noise necessarily contains both random and modeling errors. In practice, "noise" is whatever component of the measured data that $g(\mathbf{m})$ cannot account for. See Gouveia and Scales (1998) and Scales and Snieder (1998) for a discussion.

Figure 5.6 illustrates how the hyperparameters σ_1 and σ_2 take into account this 'modelling' error. The Hierachical Bayes tomography algorithm was run over three iterations. During the first iteration, travel times for all the sampled models were computed assuming straight rays between station pairs. Then the average of all



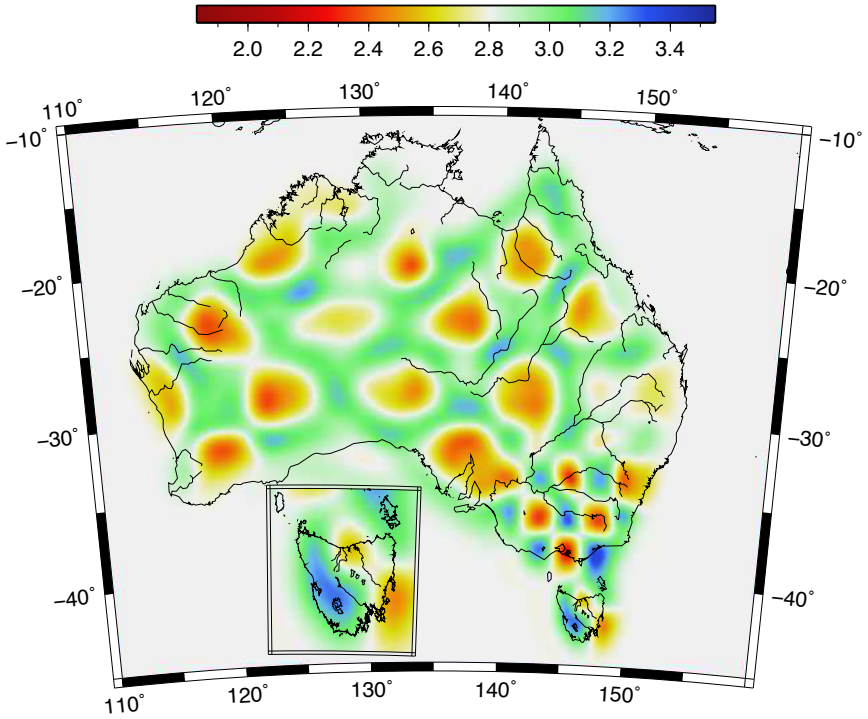
(a) Coarse grid: 56× 39 nodes. (b) Fine grid: 112× 78 nodes.

Figure 5.7: Fixed B-spline node parameterisation used for the subpsace inversion.

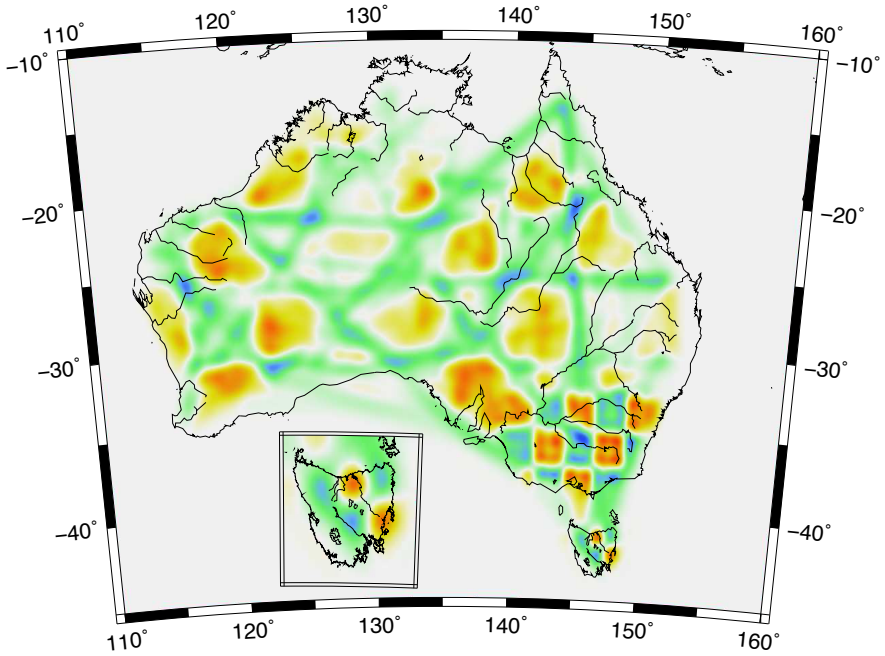
collected models was used to update ray geometries which were used in the next iteration. We show in Figure 5.6 the posterior distribution on the two hyperparameters obtained after each of these three 'outer-loop' iterations, that is for three different approximations of the forward model. The values taken by the hyperparameters during the random walk are clearly higher that the true data noise and seem to decrease as the iterations progress. By retracing the rays and iterating the process, the true rays geometries can be better approximated and the expected posterior value of the noise parameters decrease and converge towards the true value of data noise. As ray paths are better approximated, the error present in the forward model g decreases and contributes less to the data misfit. Hence the hyperparameters σ_1 and σ_2 effectively quantify the ability of the model to fit the data and therefore take into account all contributions to the misfit.

5.4.5 Comparison with the Subspace inversion

In order to compare our result to a fixed grid optimization based inversion, we jointly inverted the three datasets with the subspace method (Kennett *et al.*, 1988; Rawlinson *et al.*, 2006, 2008). Note here that the level of data noise is not accounted for as the problem is regularised using "ad-hoc" damping and smoothing and a scale factor in C_d simply absorbs into the regularization parameters. Several grid sizes were tried, and for each an iterative L-curve method (Aster *et al.*, 2005) that successively 'tuned' both damping and smoothing parameters was carried out as in Rawlinson *et al.* (2006).



(a) Coarse grid: 56× 39 nodes. $\epsilon = 4.5, \eta = 6$



(b) Fine grid: 112× 78 nodes. $\epsilon = 3, \eta = 5$

Figure 5.8: Solution models obtained with the Subspace inversion for the two different grid sizes shown in Figure 5.7. The smoothing and damping regularization parameters have been chosen by successively finding the maximum curvature of the L-curve.

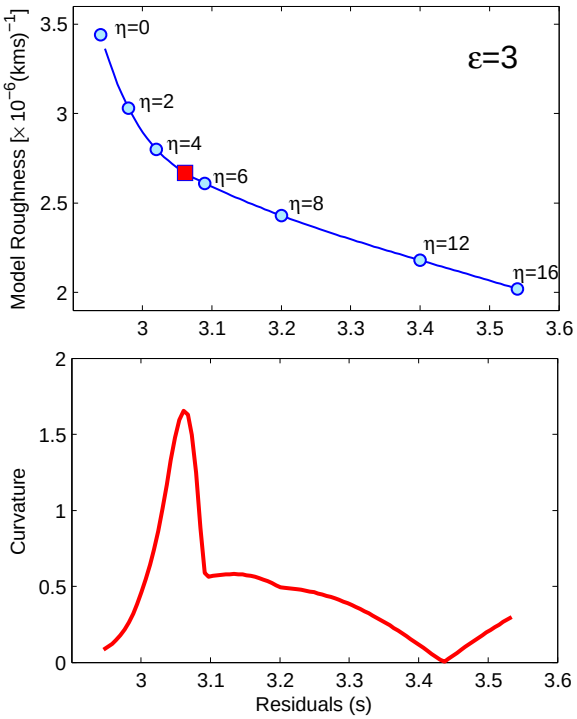


Figure 5.9: Upper panel: L-curve for the 112×78 nodes grid, ϵ is kept constant and η is changed Lower panel: curvature of the L-curve. The maximum curvature gives the corner of the L-curve and provides the optimum η.

Here we show two solutions obtained with two grid sizes as shown in Figure 5.7. A coarse parameterisation was defined with 56×39 nodes, that is with a distance separating the B-spline nodes of about 1 degree. A finner Parameterisation was defined with a finer grid of 112×78 nodes, i.e. an internode separation of 0.45 degrees The two solutions obtained with the two grid sizes are shown in Figure 5.8. The coarse grid solution shown in Figure 5.8(a) recovers the amplitudes relatively well of the true model but misses the sharp discontinuities and the small scale features under WOMBAT arrays. By using a finer grid, in Figure 5.8(b) the velocity field in south-east Australia is better recovered but small artefacts are introduced elsewhere and overall the amplitudes are worse. In both cases, information seems to be lost compared to the reversible jump solution. The coarse grid appears to be adequate for the large scale dataset whereas the finner grid is best for WOMBAT arrays.

The level of residuals (i.e. the data misfit) for both solution is relatively similar with an 'rms' value of 3.3 s for the coarse grid solution and of 3.1 s for the finer grid. In the absence of information on the measurements errors, there is no way to objectively quantify which of these two solutions better describes the true velocity

model.

Figure 5.9 shows the L-curve and its curvature obtained for the finer grid in 5.8(b) for a fixed damping value ($\epsilon = 3$) and with a varying smoothing parameter η. The curve shows the model roughness against the data fit for several values of η. The sharpness of the corner is not very well-defined, resulting from the fact that different datasets with different properties (error and scale) are inverted together. The Subspace inversion only uses a single global smoothing value although the optimal regularization parameter may be different according to datasets. This L-curve solution gives a mean of residual of 3.1 s (which is a weighted average between the two true values of data noise 1 and 4), and there is no way to discriminate between different data types.

Smoothing constraints are applied equally to all parts of the model regardless of the actual resolution capability which depends on the ray coverage. This smoothing prevents unconstrained artefacts in the poor-sampled areas but also suppresses model details in the well-sampled areas.

Note that other methods can be applied to find the 'optimal' regularization parameters like the discrepancy principle (Aster *et al.*, 2005). However, the discrepancy principle is based on the level of measurement noise which is unknown here.

5.5 Field data application

In this section we apply the Hierachical Bayes reversible jump algorithm to the three real datasets presented above and construct a multiscale tomographic image of Rayleigh wave group velocities at a period of 5 s.

5.5.1 Data noise parameterization

As previously, errors are assumed to be independent and normally distributed with zero mean and standard deviation σ_i. As a consequence, the data covariance matrix is diagonal and σ_i represent its diagonal elements. There are 5142 path averaged velocities available and it would be infeasible to invert for the uncertainty σ_i of each of these data. Instead, the errors are modelled by only a few numbers, the definition of which we call the noise parameterization.

In the synthetic example shown above, the data noise was parameterised with only two hyperparameters σ_1 (for the large scale dataset) and σ_2 (for WOMBAT arrays). Here, different arrays are also modeled separately but the nature of the data may require the noise parameterisation to be slightly more complex rather than a

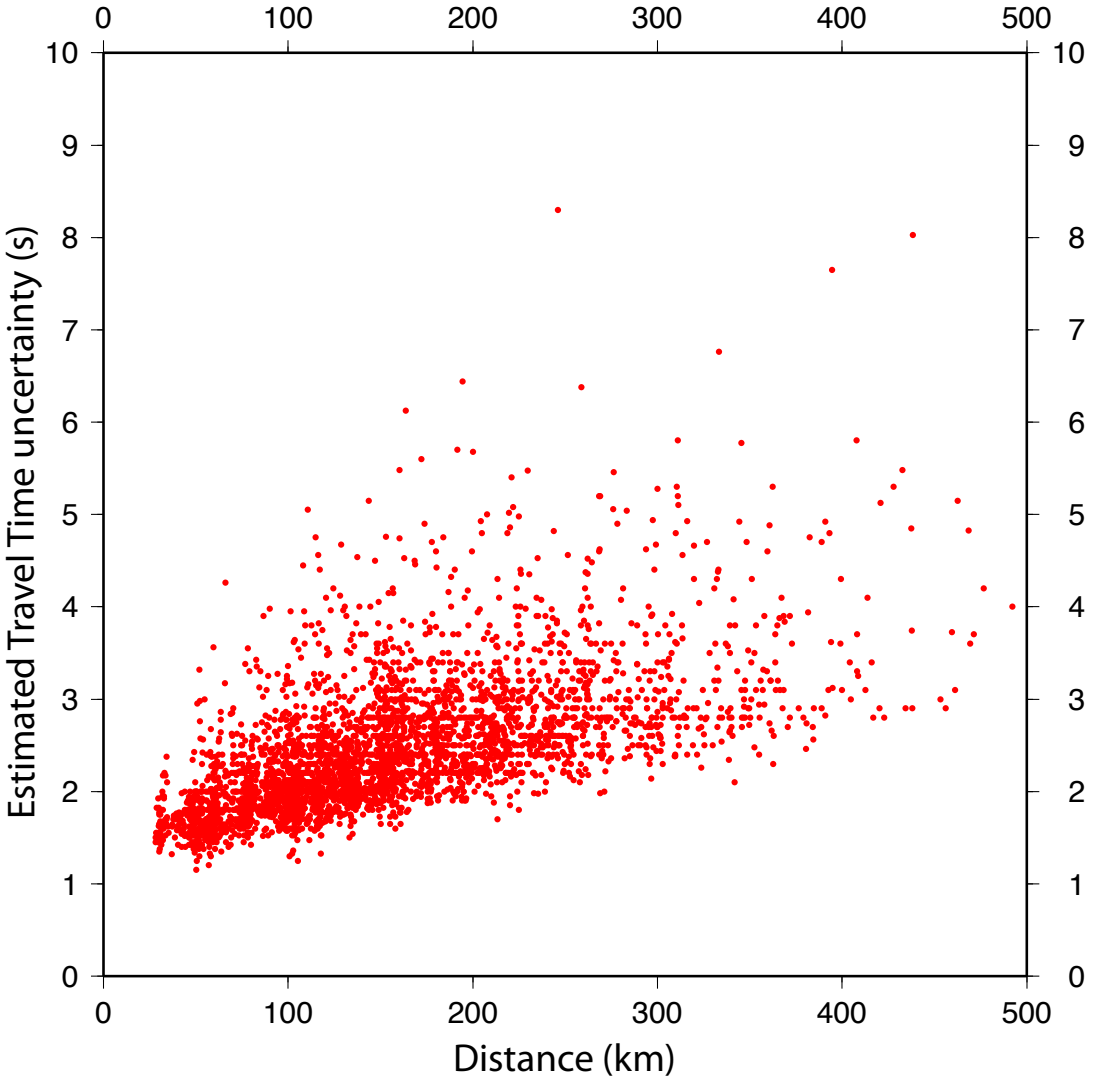


Figure 5.10: Estimated travel time picking error as a function of interstation distance for WOMBAT arrays. According to Cotte and Laske (2002) and Harmon *et al.* (2007), this method of estimating uncertainties only provides relative errors between data points.

single constant number for each array type. This subsection presents the noise parameterization chosen for each array and explains how it is based on physical arguments.

5.5.1.1 WOMBAT arrays

Arroucau *et al.* (2009) produced traveltime picking uncertainties for WOMBAT arrays following the procedure presented by Cotte and Laske (2002) and Harmon

et al. (2007). That is, uncertainties were defined as the half-width of the time interval during which the amplitude of the envelope was 50% of its peak amplitude. The aforementioned authors state that this choice for error bars sometimes results in a large scatter that may not reflect the actual precision of the measurements. Thus their error estimates need to be empirically rescaled, i.e. to be multiplied by a constant factor.

These travel time errors estimates for WOMBAT arrays are shown as a function of the angular distance in Figure 5.10. It is clear that the data noise increases and is more scattered with increasing interstation distance. Short travel times are very well constrained with an estimated uncertainty around 1.5 s while picked travel times for stations that are far apart are less precise. For an angular distance of three degrees, the estimated noise is around 3 s.

In an optimization based inversion, this estimation of relative errors would be sufficient to weight the information brought by each data. However, as we are here interested in absolute values, these uncertainties must be rescaled and we invert for a scaling factor applied in front of the data covariance matrix. Therefore, the σ'_i in (5.2) take the form

$$\sigma'_i = \lambda \times \sigma^i_{rel} \tag{5.4}$$

where σ^i_{rel} corresponds to the relative uncertainty given by the procedure described above and shown in Figure 5.10, and λ is a model hyperparameter to be inverted for. In this way, the algorithm uses the information available on relative errors in conjunction with posterior inference on a scaling factor provided by the Hierarchical Bayes procedure.

5.5.1.2 Large scale dataset

No information at all is available on data uncertainties for the large scale dataset. Instead of parameterising the data noise with a single hyperparameter as in the synthetic example, here we treat the noise as a linear function of the interstation distance and use two noise parameters. Hence, for each model sampled along the Markov chain, the likelihood in equation (5.2) is computed with:

$$\sigma'_i = a \times d_i + b \tag{5.5}$$

where d_i is the intersation distance, and a and b are hyperparameters to be inverted for during the inversion and that randomly vary along the chain. Modelling uncertainties as a linear function of distance is a choice that can be justified with several

arguments both empirical and theoretical.

The above trend is a common phenomenon called a proportional effect (Aster *et al.*, 2005). It occurs when the size of measurements errors are proportional to the measurement magnitude. It is clearly observable in Figure 5.10 for the relative uncertainties measured for the WOMBAT arrays. Another way to detect this effect is to examine the behaviours of residuals as a function of the interstation distance. We plot in Figure 5.11 the residuals obtained with conventional reversible jump tomography (i.e. with a given fixed constant σ_{est}) for the large scale dataset. These residuals are simply the differences between estimated travel times from Figure 5.2(c) and observations. Although they appear random, the absolute value of the residuals clearly seems to increase as the interstation distance increases. As expected, the "rms" value almost equals the given estimated noise $\sigma_{est} = 12$. However, short rays tend to be over-fitted relatively to this value whereas long rays are under-fitted. This shows that 'long' rays are more difficult to fit and hence could carry more noise than 'short' rays.

They are also some theoretical arguments to explain the proportional effect in ambient noise data.

At longer distances, the coherent part of the noise between two stations is more attenuated, and hence more recording time may be needed to construct the Green's function (Bensen *et al.*, 2007). There is also a decrease in signal to noise ratio with distance due to the smaller range of azimuths of propagating surface waves that contribute constructively to the cross correlation and to the scattering and multipathing along the great circle path between stations.(Harmon *et al.*, 2007; Weaver *et al.*, 2009).

5.5.2 Results

The Hierarchical Bayes algorithm was run using two different definitions for the likelihood. Datasets were first inverted with an L_2 misfit measure as in (5.2), this is the most common definition used in inversions. We also tried to define the misfit with an L_1 norm, that is :

$$\phi_{L_1}(\mathbf{m}) = \sum_{i=1}^{N} \frac{|d_i - g(\mathbf{m})_i|}{\sigma_i'} \tag{5.6}$$

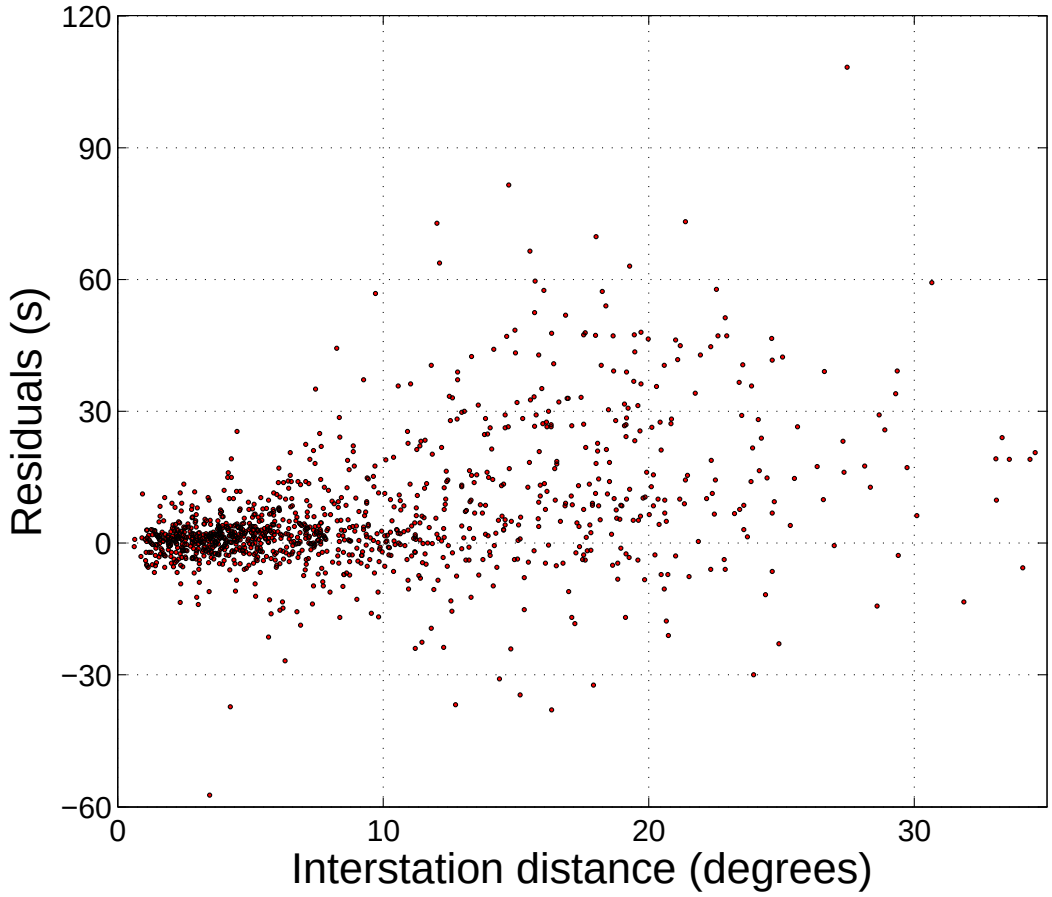


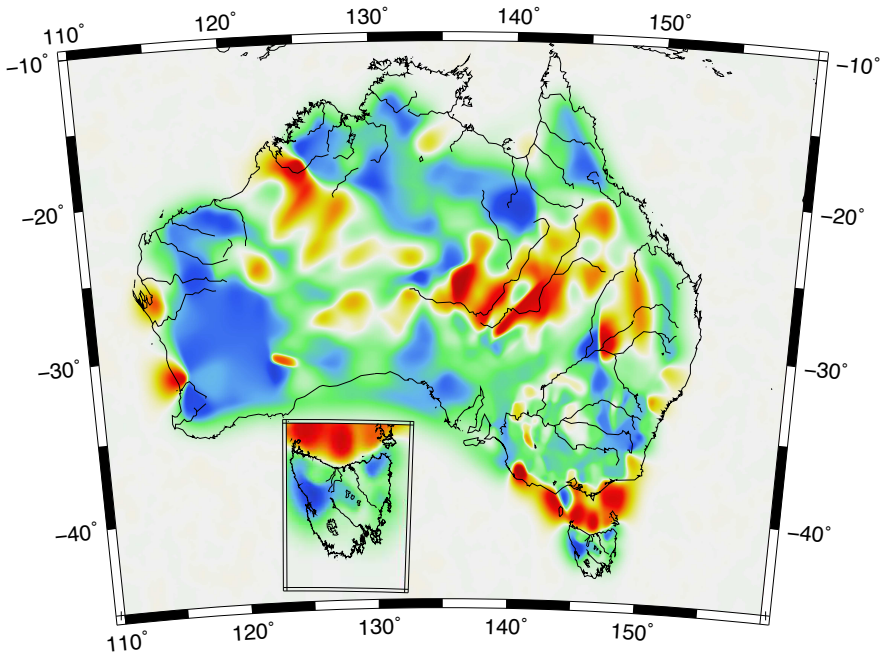
Figure 5.11: Residuals against interstation distance for the solution map showed in Figure 5.2(c), which was produced by inverting the large scale dataset with the conventional reversible jump. The purpose of this figure is to show the correlation between residuals and ray length.

The likelihood function associated with the L_1 norm is a double sided exponential distribution, which is comonly named a Laplacian probability distribution :

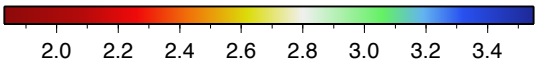
$$p(\mathbf{d}_{obs}|\mathbf{m}) = \prod_{i=1}^{N}\left[(2\sigma_i')^{-1}\right] \times \exp\{-\phi_{L_1}(\mathbf{m})\}. \tag{5.7}$$

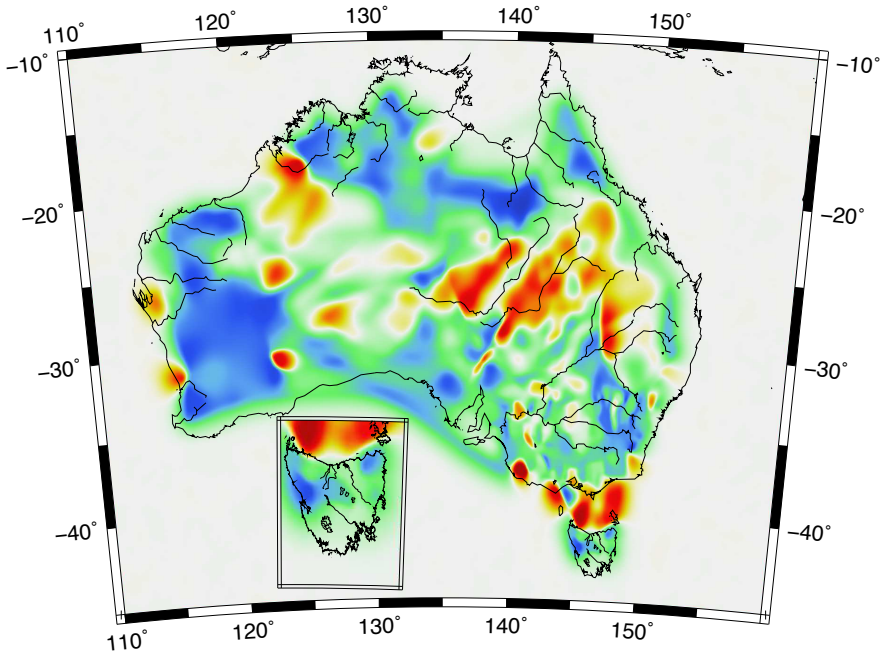
The advantage of using a Laplacian likelihood distribution is that the average solution will be more outlier resistant, or robust, than the expected Earth model obtained with a Gaussian likelihood (Aster *et al.*, 2005).

In an optimization framework, finding the L_1 norm solution is complicated be-



(a)





(b)

Figure 5.12: Hierarchical Bayes average solution (kms^{-1}). top: data misfit defined with an L_1 norm, i.e. the likelihood is Laplacian distribution. Bottom: data misfit defined with an L_2 norm, i.e.the likelihood is a Gaussian distribution.

cause the likelihood function is then a non differentiable function of $\mathbf{m}$ at any point where one of the residuals $d_i - g(\mathbf{m})_i$ is zero. The general problem of finding a solution model that maximises equation (5.7) becomes relatively complex. However, Monte Carlo schemes do not use derivatives and sampling a Laplacian distribution is as straightforward as sampling a Gaussian distribution.

Figure 5.12 shows the average solution maps for both L_1 and L_2 misfit definitions. Both maps were obtained after four 'outer-loop' iterations (i.e. four updates of the ray paths). For each iteration, a total of 96 independent Markov chains were run independently on separate processors and posterior inference was made using an ensemble of 57600 models. Each chain was run for 3×10^6 steps, the first half of which were discarded as burn-in. Then, every 250th model visited was taken in the ensemble. In our case, it turns out that the solution model obtained with an L_1 norm is relatively similar to the one obtained with an L_2 norm.

The travel times used here are mostly sensitive to the structure in the first 3 km of the crust and the solution models in Figure 5.12 resolve shallow crustal structure, and clearly discriminate between sedimentary and hard rock regions. The results indicate the variations in wavespeed between Archaean cratons in west Australia and sedimentary zones in the central and eastern Australia. The seismic images are in good agreement with results from receiver function and earthquake tomography (Fishwick *et al.*, 2005; Zielhuis and Hilst, 2007; Fishwick and Reading, 2008). As the aim of this study is to concentrate on the mathematical aspects of the inverse problem, it is beyond our scope to give a detailled geological and tectonic interpretation of our results, and hence we only show results for a period of 5 s. This results are in agreement with the first order geological map showed in Figure 3.24.

The posterior information on hyperparameters (n, λ, a, b) is shown in Figure 5.13. Note that the collected velocity models in the ensemble solution have an average of 1200 cells. Each cell is defined by a 3 parameters (2D location of the nucleus + velocity) which makes the dimension of the model space around 3600. The Monte Carlo integration of such a space is normally impossible given "the curse of dimensionality". However here, due to the Voronoi parameterization, the posterior distribution is highly redundant which reduces dramatically the apparent size of the model space. 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 field does not change at all, although we are at a different point in the parameter space. This shows that any velocity model can be described by a large number of points in the model space. Furthermore, the

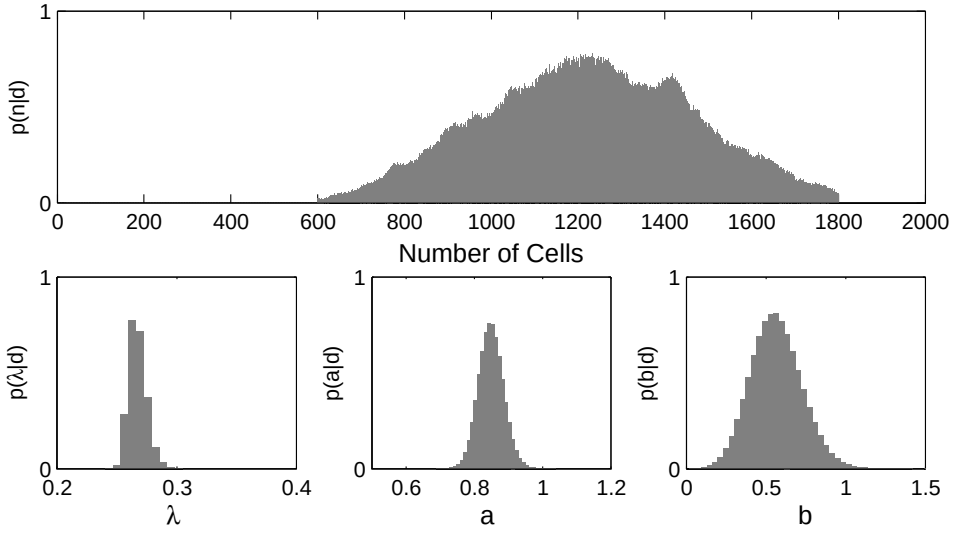
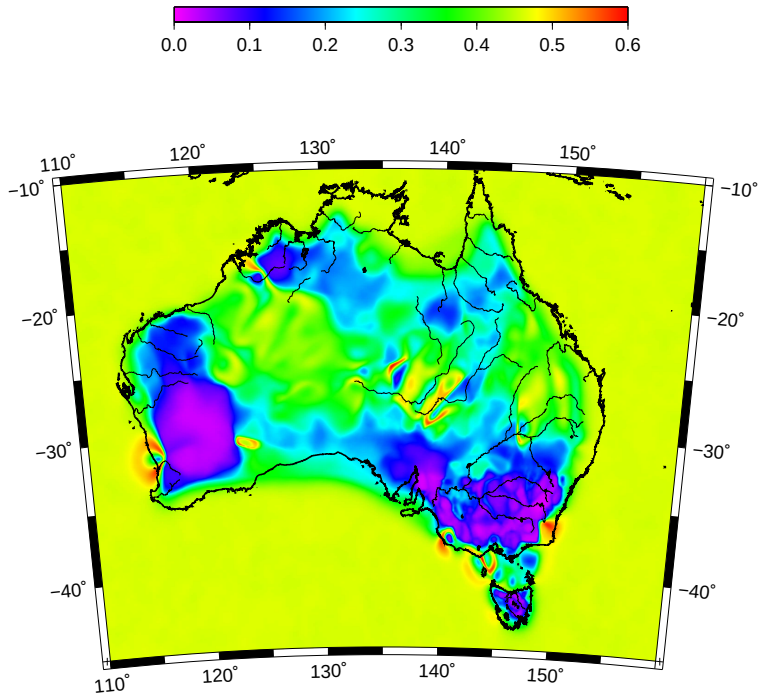


Figure 5.13: Posterior probability distribution on hyperparameters for the L_1 misfit solution shown in Figure 5.12(a).

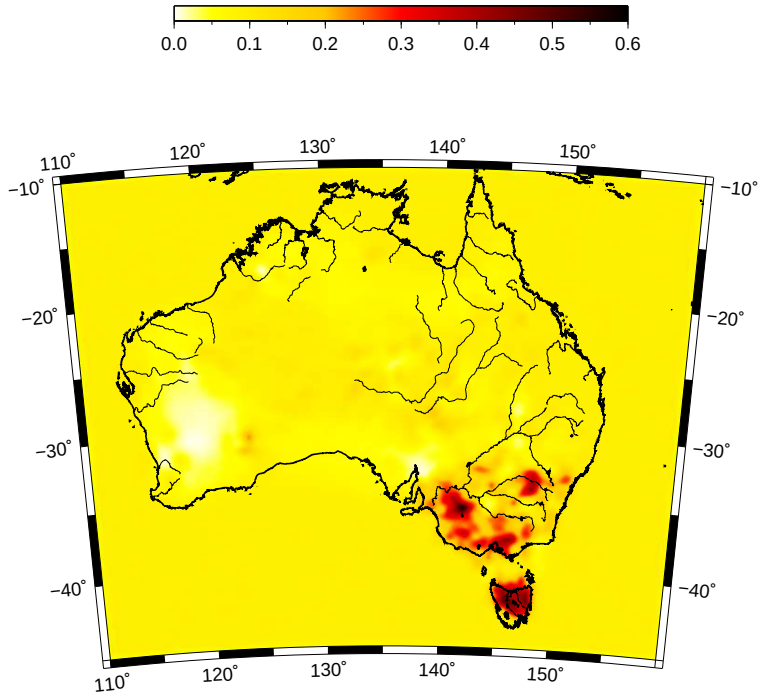
Monte Carlo integration is carried out thanks to parralelization of the code on a super-computing facility like a 'Beowulf'' cluster of computers. To give the reader an idea of the computational cost of such an inversion, each 'outer-loop' iteration of the algorithm needed approximatively 5 days, so each panel of figure 5.12 represents about 15 days of computational time. The inferred information on the level of noise in travel times indicates that the uncertainties provided for the WOMBAT arrays have been rescaled to around $\lambda = 0.28$. The posterior value on the hyperparameter a indicates that the data noise for the large scale is expected to increase 0.85 s each time the interstation distance increases by 1 degree with an expected data error of 0.6 s at 0 degrees (see lower panels in Figure 5.13).

Figure 5.14(a) shows the error map for the L_1 solution, which is constructed by taking, at each point of the velocity field, the standard deviation of the ensemble of sampled partitionned models. This locally shows how well the solution model in 5.12(a) is constrained. As expected, well sampled areas in Western Australia, South East Australia and Tasmania show a lower velocity uncertainty.

It is also interesting to look at the spatial density of Voronoi nuclei across the ensemble of models collected. The velocity field was discretised with constant square cells of size 0.5×0.5 degrees, and Figure 5.14(b) shows the average number of Voronoi nuclei per cell over the ensemble of collected models. Hence this map directly shows the average size of Voronoi cells at each point of the velocity field. A large number of small Voronoi cells are concentrated at WOMBAT arrays while there are larger



(a)



(b)

Figure 5.14: Top: Error map (kms^{-1}) associated with the L1 norm solution in 5.12(a). Bottom: Density of Voronoi nuclei across the ensemble of sampled models. The colour scale represents the expected number of nuclei per pixel

cells elsewhere. This example demonstrates the adaptive character of the Voronoi parameterization.

It is interesting to see that the estimated error on the model is not necessarily correlated with the density of cells. The Archaean cratons in Western Australia are indeed well constrained without need of small cells (This area shows low values for model uncertainty in 5.14(a) with the lowest density of cells in 5.14(b)). There is good ray coverage in Western Australia, and the algorithm would be expected to introduce a lot of small cells to provide details there. However, the velocity field seems to be quite homogeneous and there is no need to introduce complexity to properly image it. This example shows the parsimonious nature of the algorithm and indicates that the Voronoi parameterization not only adapts to the density of rays but also takes into account the underlying velocity structure.

5.6 Conclusion and future work

We have shown here that the reversible jump algorithm is particularly suited for inversion of multiple datasets that sample the Earth at different scales. Synthetic and real data examples have illustrated the adaptive character of the parameterization which enables us to image small scale features in well sampled areas without introducing spurious artefacts elsewhere. The level of smoothing is spatially variable and is naturally determined by the data. Contrary to other multiscale tomography methods, it is not only regulated by the density of rays but also by the inferred data noise and by the structure of the underlying velocity field.

As the complexity of the model is variable, the estimated level of data noise takes an important role in the inversion as it directly determines the number of model parameters needeed to fit the data to the required level. We have shown that an extended Bayesian formulation called Hierarchical Bayes can take into account the lack of knowledge on the level of data uncertainty. When assessment of measurements errors is difficult to achieve a priori (as in ambient noise tomography), this procedure treats the standard deviation of data noise as an unknown and makes a joint posterior inference on both model complexity and data uncertainty. The Hierarchical Bayes procedure turns out to be particularly useful when dealing with multiple data types having different unknown levels of noise. With scant prior knowledge on data errors, the algorithm is able to infer the level of information brought by each data type and to naturally adjust the fit to different datasets.

In our ambient noise tomography application, the data noise from WOMBAT

arrays was naturally rescaled while the noise for the large scale dataset was parameterised as a linear function of the interstation distance. The inversion resulted in a parsimonious velocity map with a spatial resolution adapted to the quantity of information present in the data.

Uncertainty assessment on apparent travel times from ambient noise cross corelation is an active area of research and Hierarchical Bayes could be used as a tool to quantify the behaviour of noise with different parameters like inter-station distance, azimuthal source distribution, or recording time.

It will be soon possible to incorporate new data from on-going deployments at different scales in Australia. Saygin and Pozgay (2010) recently processed additional travel times that sample the crust at the continental scale. There are aslo suplementary ambient noise data for Tasmania available (Young *et al.*, 2010). Furthermore, 67 short-period seismometers have recently been positioned across the Gawler and Curnamona Cratons in South Australia (Salmon and Arroucau, 2010). Station spacing was approximately 60 km and covers the area from the Streaky Bay in the west to the New South Wales border in the east. Stations recorded continuous three component data for a period of 6-8 months and ambient noise travel time are currently being processed.

Another study that could be carried out in the future is to invert for anisotropy. Arroucau *et al.* (2009) observed an azimuthal dependence on the path-averaged velocities extracted from WOMBAT arrays. Therefore, instead of inverting for a single velocity value within each cell, one would invert for 3 anisotropic parameters within each cell (a maximum velocity, a minimum velocity, and a direction).

Possible extension also include combining other classes of seismic data like receiver functions with the ambient noise. The Hierachical Bayes procedure is expected to be a powerful tool when used for joint inversion as it would be able to naturally weight the contribution of different data types in the misfit function, thus removing the arbitrary choice of a weighting factor.

Chapter 6

Transdimensional Inversion of Receiver Functions With the Hierarchical Bayes Algorithm

6.1 Introduction

The purpose of this chapter is to show that the class of algorithm presented in this thesis is not restricted to seismic tomography but is rather a general approach to inverse problems. Here we propose to apply our methodology to invert receiver function waveforms (RF), which is a well known highly non linear problem (Langston, 1979; Ammon *et al.*, 1990).

The coda of teleseismic P-waves contains a large number of direct and reverberated phases generated at interfaces beneath the receiver that contain a significant amount of information on seismic structure. However, they are difficult to identify as they are buried in microseismic and signal-generated noise (Lombardi, 2007). The signal to noise ratio is usually improved by stacking seismograms from different records recorded at a single station, but a major drawback is the introduction of different source time functions generated by different earthquakes.

This problem is overcome by a method developed in the 1970's following the pioneering work of Phinney (1964) which is now widely used in seismology. The idea is to deconvolve the vertical component from the horizontal components to produce a time series called a 'receiver function' (Langston, 1979). In a receiver function the influence of source and distant path effects are eliminated, and hence one can enhance conversions from P to S generated at boundaries beneath the recording site.

In this chapter we present a RF inversion methodology where the number of

layers defining the velocity model as well as the presumed magnitude and correlation of data noise are variable and treated as unknowns in the problem.

6.1.1 A brief history of receiver function inversion

The receiver function waveform can be inverted in the time domain for a 1D S-wave velocity model of the crust and uppermost mantle beneath the receiver. The RF inverse problem is highly non-linear and Ammon *et al.* (1990) used synthetic tests to show the non uniqueness of the solution. However, linear inversions were first used due to their simplicity. Owens *et al.* (1984) carried out an iterative linearised inversion where partial derivatives in the linearised equations were computed numerically with a finite difference scheme. The inversion was stabilised with truncation of small eigenvalues after singular value decomposition of the system of equations. Kosarev *et al.* (1993) and Kind *et al.* (1995) used a linearised Tikhonov inversion and stabilised the algorithm by penalising solutions far from a given reference model. Linear inversion procedures based on partial derivatives are easily trapped by local minima, and hence final models are strongly dependent on initial models. Therefore the initial model must be close to the true velocity structure for the procedure to be meaningful, which is difficult to ensure.

In order to deal with the inherent non-linearity of the problem, Zhu and Kanamori (2000) and Chevrot and van der Hilst (2000) proposed a simple 2D grid search method where only first order crustal features are inverted, namely Moho depth and mean Vp/Vs ratio. By staking RFs from different distances and directions, effects of lateral structural variations are suppressed, and an average crustal model is obtained. Thus the Earth model is simply parameterised with two model parameters describing a single layer over a half-space. Consequently, this method has been widely applied to map Moho depth and mean crustal composition (e.g. Yuan *et al.*, 2002; Vergne *et al.*, 2002; Harland *et al.*, 2009). However, this straightforward approach is limited as it ignores more complex structures that are present in the RF and could effectively be constrained.

As more computational power became available, Monte Carlo parameter search methods progressively became an unavoidable alternative for RF inversion. First, global optimization techniques such as genetic algorithms (Shibutani *et al.*, 1996; Levin and Park, 1997; Clitheroe *et al.*, 2000b; Chang *et al.*, 2004), niching genetic algorithm (Lawrence and Wiens, 2004), simulated annealing (Vinnik *et al.*, 2004, 2006), very fast simulated annealing (Zhao *et al.*, 1996), or the neighbourhood algorithm (Sambridge, 1999a; Bannister *et al.*, 2003; Nicholson *et al.*, 2005) were succes-

sively applied to avoid solutions trapped in local minima of the objective function. These approaches are able to efficiently search a large multidimensional model space and provide complex Earth models that minimise a misfit measure without need of linearisation.

As mentioned before, an inherent characteristic of RF inversion is that the problem is highly non-unique, thus two Earth models that are far apart in the model space (i.e which have different parameter values) can provide a similar data fit. Unfortunately, non-linear optimization algorithms are good at searching a large space and finding a global minima but they only provide a single solution, i.e. the best one in some sense, which leaves open the possibility that other Earth models which are far from this solution might also fit the data within errors. Hence a single solution is not representative of the information brought by the data. To avoid the problem of non-uniqueness, global optimization techniques have been used to perform an ensemble inference, where one obtains an ensemble of models satisfying some predefined criteria (e.g. the best 1000 data fitting models generated by the algorithm). The ensemble of 'acceptable' models are thus plotted together for visualisation (e.g. Piana Agostinetti *et al.*, 2002; Reading *et al.*, 2003; Hetényi and Bus, 2007).

However, the criteria defining whether or not a model can be accepted in the ensemble is chosen arbitrarily. Furthermore, the statistical distribution of models within the ensemble does not represent the objective function and therefore cannot be directly used to infer trade-off, constraints or resolution on model parameters. This is because most non-linear optimization algorithms do not perform importance sampling (i.e where the frequency distribution of sampled models is proportional to the objective function), and hence the ensemble solution strongly depends on user choices or on the class of algorithm employed.

A typical example has been the use of the neighbourhood algorithm (Sambridge, 1999a) for RF inversion (e.g. Piana Agostinetti *et al.*, 2002; Reading *et al.*, 2003; Bannister *et al.*, 2003; Frederiksen *et al.*, 2003; Hetényi and Bus, 2007). In a second paper, Sambridge (1999b) invoked the Bayesian philosophy and showed how to carry out a proper ensemble inference by means of a simple scheme that uses the ensemble of collected models to resample the model space at a relatively low computational cost. However, most studies that have used the neighbourhood algorithm only used it in an optimization context and plotted the best 1000 sampled models which give an idea of the non-uniqueness but cannot be directly used to quantify resolution and trade-offs. The distribution of the 1000 best models depends on the tuning parameters of the neighbourhood algorithm, which control the balance between

exploration and exploitation of the search.

The "Bayesian neighbourhood algorithm" (Sambridge, 1999b) was used for RF inversion by Lucente *et al.* (2005) and Piana Agostinetti and Chiarabba (2008). Subsequently, Piana Agostinetti and Malinverno (2010) expanded the Bayesian formulation to the case where the number of layers is not fixed in advance but is treated as an unknown in the problem. This appears to be the first application of transdimensional McMC to the receiver function problem. In this chapter, we follow their lead and extend their scheme to the hierachical case where noise levels are also treated as unknowns.

Instead of being an ensemble of best fitting models, in a Bayesian framework the solution is an ensemble of models that represent the posterior probability distribution quantifying the degree of belief we have about the Earth structure and composition. This probability distribution combines '*a priori*' information with information contained in the observed data. The best fitting models represent the maxima of this distribution, the tails are described by poorly fitting models in the ensemble, and the 'width' or the variance quantifies the constraints we have on model parameters, i.e the uncertainty on the inferred solution. Finally, the covariance of the posterior distribution shows the correlation or trade-off between model parameters.

6.1.2 Receiver function variance

As shown in previous chapters, the estimated data noise plays a critical role in Bayesian inference. In an optimization framework the solution does not depend on the level of data noise (the best fitting model remains the same as we rescale error bars around data). In contrast, with a Bayesian framework, the data uncertainty directly determines the variance of the posterior probability distribution. In the context of a transdimensional parameterization, the variance of data noise becomes even more important in the inversion. This is because it quantifies the level of usable information in the data, and thus the number of parameters needed to construct the inferred models. As seen in previous chapters, when given a level of data noise, our transdimensional Bayesian algorithm naturally adapts the complexity of the solution in order to fit the data to the required level (see Figures 4.3, 4.5 and 5.2). Of course, as more model parameters (e.g. more layers) are introduced, the data could be fit better, but the procedure naturally prevents the data to be fit more than the given level of noise.

Receiver functions are time series and hence the noise will be correlated from

sample to sample by the band-limited nature of the waveforms. The uncertainty in RFs has mainly four different origins : observational noise, errors involved in the deconvolution, modelling errors, and incoherence between the averaged RFs due to the different backazimuths and angles of incidence of events.

Observational errors on the seismic waveform result from background seismic noise (microseisms) and from the instrumental noise affecting the recording. The signal to noise ratio is expected to increase with the increasing magnitude of the events and can be considered uncorrelated among different receiver functions. Secondly, errors occur in the deconvolution which is an unstable process. For example, in the case of a frequency domain deconvolution, it is known that the result is heavily corrupted if the numerator has significant spectral power where the denominator is small. Then, there are assumptions made about Earth (e.g. horizontal homogeneous isotropic layers) that prevent us from reproducing the observed RFs. We call 'modelling errors' the part of the data that cannot be modelled by our physical approximation of the Earth. This type of noise is coherent and fully reproducible, and following the definition of Scales and Snieder (1998), it is a part of the signal we choose not to explain. For example, the complex structures near the receiver produce a scattered wavefield that is not taken into account in our forward model and which is thus considered as data noise in the inversion. Finally, the 'observed' Receiver function is an average over a number of events arriving at the sation from different backazimuths and at different angles of incidence. This may imply incoherence between RFs due to a range of effects like anisotropy and/or dipping layers, although this effect is typically avoided by staking waveforms from events occurring in the same region.

As shown by Di Bona et al. (1998), all these contributions to the RF variance may not be simply additive and an overall quantification of the noise in terms of magnitude and correlation may be difficult. Furthermore, those different undesirable effects are suppressed with a range of schemes where the noise reduction is even more difficult to quantify.

The background noise is reduced by the averaging proccess involved in the construction of the observed RF. The frequency domain deconvolution is stabilised with a water-level scheme, whose parameter is chosen by trial and error (Clayton and Wiggins, 1976). In addition, a Gaussian filter is applied to limit the final frequency band, in order to reduce the sensitivity to fine structure (if one is interested in developing an initial first-order model of the velocity structure beneath the receiver), or to exclude the high-frequency signals not justified by the frequency content of

the original data (Di Bona et al., 1998). Often, outlier RFs in the stack are eliminated from visual inspection and hence there is no clear quantification of the degree of noise, although Tkalčić *et al.* (2010) recently proposed a statistical approach to select coherent receiver functions from an ensemble of events.

Gouveia and Scales (1998) showed in the case of Bayesian seismic waveform inversion that it is important to consider all the contributions to the misfit for an accurate construction of the posterior probability distribution. This becomes even more crucial when the number of parameters is variable, because in this case the level and correlation of noise directly dictate the number of layers and thus the complexity of the solution. Therefore here it becomes clear that a major issue of a transdimensional RF inversion is the quantification of data noise.

6.1.3 The covariance matrix of data errors

In most Bayesian studies, the data noise is assumed to be normally distributed, and accounted for in the likelihood function by means of a covariance matrix $\mathbf{C}_d$. While most optimization inversion schemes assume no correlation, different procedures have been used to construct the covariance matrix of data errors needed in a Bayesian inversion. Ammon (1992) estimated the noise level from the power-spectral density in a pre-signal time window (the segment which precedes the direct P-wave arrival). Sambridge (1999b) simply used the statistics in the ensemble of averaged RF to produce an average, a variance and a covariance. Piana Agostinetti and Malinverno (2010) derived the noise correlation from the averaging function which is calculated by deconvolving the vertical component of motion from itself, using the chosen water-level parameter. If the water-level fraction is zero, the result is a perfect Gaussian (from the low-pass filter included in the procedure). Di Bona et al. (1998) evaluated the noise involved in the frequency domain deconvolution by using the residuals of a time domain deconvolution of the averaging functions from the computed RFs, in the portion preceding the P-pulse.

These different schemes approximate different effects, and it is clear here that there is no consensus to date on the way of measuring noise in RFs. Although these techniques can be used to infer the level of observational and processing noise, they do not estimate the modelling uncertainties such as errors due to assumption of a laterally homogeneous isotropic medium. Note however that Gouveia and Scales (1998) accounted for model discretisation errors in the case of Bayesian seismic waveform inversion.

In this work we address the issue of noise estimation with hierarchical models

which are able to consider the uncertainty the user has on errors. We use a Hierarchical Bayes formulation where both the level and correlation of data noise are treated as unknowns in the inversion (Malinverno and Briggs, 2004; Malinverno and Parker, 2006). This expanded Bayesian procedure lets the data infer the required level of data fit, and hence fully takes into account the complex combination of effects contributing to the misfit, which can be represented with the noise parameterisation employed. In the context of a variable number of layers, it was shown in the previous chapter that this is of great advantage over having fixed noise estimates that might be erroneous.

As with previous chapters, we first present the methodology, and then test it on some synthetic examples before considering field data.

6.2 Methodology

6.2.1 Model parameterisation

In this study the radial RF is assumed to be dominated by the response of homogeneous horizontal layers beneath the receiver. The geometry of layers is described by a variable number m of Voronoi nuclei as shown in Figure 6.1. Each nucleus is given a velocity value representing a homogeneous S-wave velocity in the layer. Hence the seismic structure is described with $2m$ parameters. Note that the last layer does not have a lower boundary and hence is a half space. The layer boundaries are defined as equidistant to adjacent nuclei, and as shown in Figure 6.1, the mapping of the seismic structure is non-unique. Two models with different parameterization can describe exactly the same Earth structure. If two adjacent Voronoi cells have the same Vs value (see third panel of Figure 6.1), then the two cells result in a single layer, and hence we acknowledge that there is no direct relation between the number of nuclei and the number of layers. Note that we could have easily removed this non-uniqueness of the parameterization by simply defining layers by the position of their boundaries as in Hopcroft *et al.* (2007). However in the context of an important sampling of the model space this non unique mapping can be seen as an advantage as a given seismic structure can be approached by different paths. That is, the best fitting seismic structure can be described by different points in the model space, and different chains sampling independently the model space in different regions can simultaneously approach the same solution.

As well described in Lombardi (2007), the timing of RFs are relative to the first

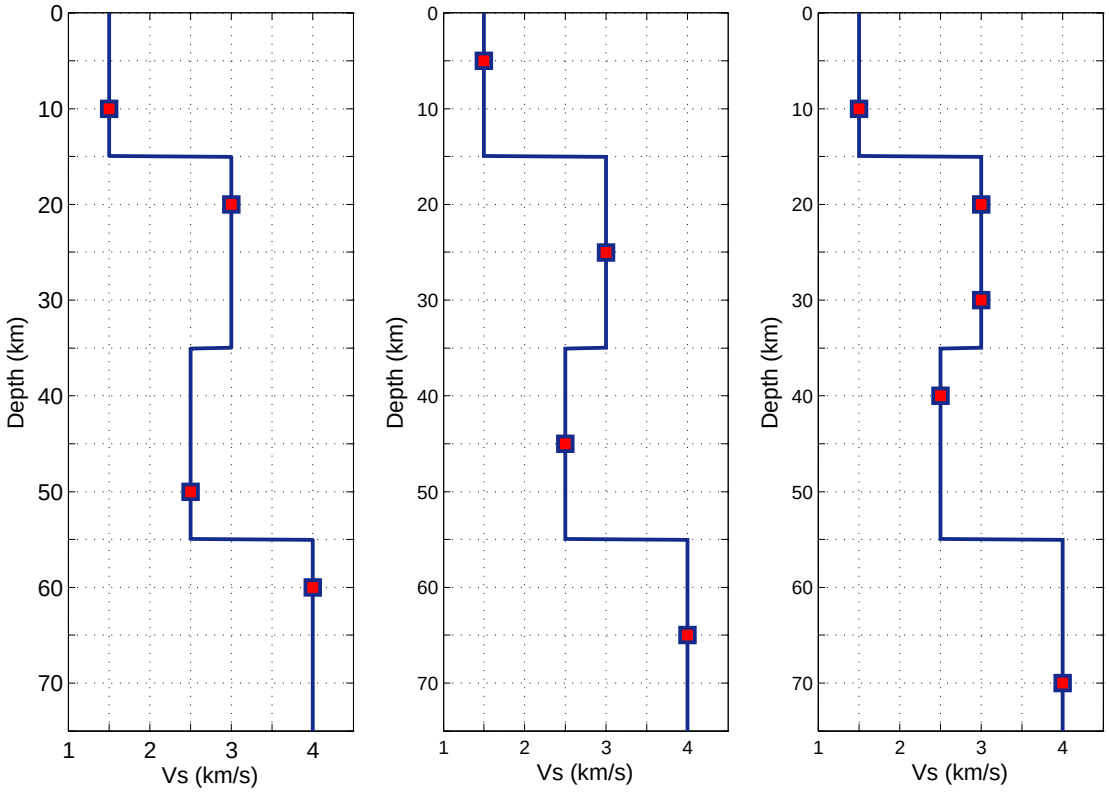


Figure 6.1: The model is parameterised with a variable number of Voronoi nuclei (red squares) which define the seismic structure (blue line). The vertical location of nuclei define the geometry of layers which Vs value is given by horizontal positioning of nuclei. Note that Voronoi nuclei are not necessary at the centre of layers but rather boundaries are defined as equidistant to adjacent nuclei. Hence the mapping is not unique, and here three models with different parameterisation and different dimensions can map exactly the same Earth structure. This makes the posterior probability distribution highly redundant and the sampling less 'rigid' as the true velocity structure can be approached by different types of models.

P arrival and thus very sensitive to the variation of Vs relative to Vp. Note that RFs are also sensitive to crustal attenuation. However in this study, we consider the Vp/Vs ratio as well as attenuation coefficients fixed to a reference model and we only invert for Vs in each layer. Note that such modelling approximations will induce data noise, that is an inability of our model to fit observations. These errors contribute to the misfit, and in a conventional Bayesian framework, they have to be taken into account in the covariance matrix of data errors which might be difficult (see Gouveia and Scales (1998) for details). For example, how would one quantify the magnitude and correlation of data noise generated by approximating a dipping

layer as horizontal? The advantage of the Hierachical Bayes formulation is that we let the data infer its own degree of uncertainty.

6.2.2 The forward calculation

Our direct search algorithm requires solving the forward problem a large number of times, that is to compute the RF predicted by a given Earth model parameterised as described above. We use the Thomson-Haskell matrix method (Thomson, 1950; Haskell, 1953) to compute the spectral response of a stack of isotropic layers to an incident planar P-wave. Multiple reflections are considered with this method. Since this way of solving the forward model is done without slowness integration, it is fast and has been widely used in Monte Carlo algorithms (e.g. Shibutani *et al.*, 1996; Sambridge, 1999a). Once synthetic seismograms have been computed for different components, receiver functions are made via frequency domain deconvolution of the vertical component from the radial component using water-level spectral division (Langston, 1979) with a water-level of 0.001.

6.2.3 The reversible jump algorithm

In a stimulating short essay, Scales and Snieder (1997) reviewed the philosophical arguments that have been invoked for and against Bayesian inversion. The principal criticism made of Bayesian inversion is that users often 'tune' the prior in order to get the solution they expect. In other words, the *a priori* knowledge of the model is often used as a control parameter to tune the properties of the final model produced. As done in previous chapters, we acknowledge this weakness and set priors to uniform distribution with relatively wide bounds.

The overall implementation of the algorithm is basically the same as presented in previous chapters and hence it is not repeated here. However, the main difference with previous applications of the Hierarchical Bayes inversion is that here the data noise is correlated. The observed receiver function can be written as

$$\mathbf{d}_{obs}(i) = \mathbf{d}_{True}(i) + \epsilon(i) \quad i = [1, n] \tag{6.1}$$

where $\epsilon(i)$ represents errors that are distributed according to a multivariate normal distribution with covariance $\mathbf{C}_d$, which may be poorly known or not known at all. We recognise that the Gaussian assumption is a very particular way of seeing the data noise that might not be appropriate. However, this is virtually the only way

the noise is accounted for in geophysical inversions.

In the case of correlated data noise, the fit to observations is no longer defined as a simple 'least-square' measure but is the Mahalanobis distance (Mahalanobis, 1936) between observed and estimated data vectors :

$$\Phi(\mathbf{m}) = (g(\mathbf{m}) - \mathbf{d}_{obs})^T \mathbf{C}_d^{-1} (g(\mathbf{m}) - \mathbf{d}_{obs}) \tag{6.2}$$

Contrary to the Euclidean distance, this measure takes in account the correlation between data. Of course, if the covariance matrix is diagonal, the Mahalanobis distance reduces to the Euclidean distance. The general expression for the likelihood probability distribution is hence:

$$p(\mathbf{d}_{obs} \mid \mathbf{m}) = \frac{1}{\sqrt{(2\pi)^k |\mathbf{C}_d|}} \times \exp\left\{\frac{-\Phi(\mathbf{m})}{2}\right\}. \tag{6.3}$$

Note that this expression requires both the inverse $\mathbf{C}_d^{-1}$ of the covariance matrix of data errors and also its determinant $|\mathbf{C}_d|$.

As explained before, our philosophy is to consider the level of data noise as an unknown in the inversion. Therefore the main issue here is to 'parameterise' the covariance matrix $\mathbf{C}_d$, i.e. to express it as a function of a given number of parameters. This is a symmetric $n \times n$ matrix that is defined with $(n^2 + n)/2$ values which are obviously impossible to estimate separately from only n data, and hence some assumptions need to be made. The covariance matrix can be written in terms of a matrix of correlation $\mathbf{R}$ and a vector of standard deviations $\mathbf{S}$

$$\mathbf{C}_d = \mathbf{S}^t \mathbf{R} \mathbf{S} \tag{6.4}$$

With this decomposition, one can separate two properties of the noise, i.e its magnitude and correlation (Piana Agostinetti and Malinverno, 2010). If the noise is considered stationary, i.e. its magnitude and correlation are constant along the time series, then $\mathbf{C}_d$ can be written:

$$\mathbf{C}_d = \sigma^2 \mathbf{R} \tag{6.5}$$

where σ^2 is the constant noise variance (i.e. the magnitude of data noise) and the

R is a symmetric diagonal-constant or Toeplitz matrix:

$$\mathbf{R} = \begin{bmatrix} 1 & c_1 & c_2 & ... & c_{n-1} \\ c_1 & 1 & c_1 & ... & c_{n-2} \\ c_2 & c_1 & 1 & ... & c_{n-3} \\ & & & \vdots & \\ c_{n-1} & c_{n-2} & c_{n-3} & ... & 1 \end{bmatrix} \tag{6.6}$$

where c_i ($i = [1, n]$) describes the noise correlation between points of the series. Thus c_1 defines the correlation between two adjacent points, and more generally c_i is the noise correlation between points that are i samples apart in the series. The key properties that we need are that the correlation function decreases with distance, with limiting values of 1 at $i = 0$ and of 0 at $i = \infty$. This is the most common kind of association found in time series. Then the magnitude of data noise is quantified by the parameter σ^2, and the main question now is how to parameterise the correlation function c_i and with how many unknowns? Below we present two types of parameterisation for the noise correlation.

6.2.3.1 1^{st} type of noise parameterization

We first propose a parameterization which is convenient to implement for our particular problem. The correlation function is simply assumed to decay exponentially and is thus given by:

$$c_i = r^i \tag{6.7}$$

where $r = c_1$ is a constant number between 0 and 1 which describes the correlation between two adjacent samples in the time series. This correlation function is plotted in Figure 6.2a for different values of r, and realisations of noise with such a correlation are shown in left panels of the same figure. Then our covariance matrix of data errors is simply described by two unknowns σ and r that represent magnitude and correlation of noise. They are called noise hyperparameters in the inversion since they do not describe Earth properties.

The major advantage of such a parameterization is that the inverse and determinant of $\mathbf{C}_d$ needed in the likelihood in (6.3) have simple analytical forms, i.e. they can be expressed in terms of σ and r. It can be easily shown with linear algebra

that the inverse of $\mathbf{C}_d$ is a symmetric tridiagonal matrix:

$$\mathbf{C}_d^{-1} = \frac{1}{\sigma^2(1-r^2)} \begin{bmatrix} 1 & -r & 0 & ... & 0 & 0 \\ -r & 1+r^2 & -r & ... & 0 & 0 \\ 0 & -r & 1+r^2 & ... & 0 & 0 \\ & & & \vdots & & \\ 0 & 0 & 0 & ... & 1+r^2 & -r \\ 0 & 0 & 0 & ... & -r & 1 \end{bmatrix} \tag{6.8}$$

See Malinverno and Briggs (2004) for a detailed demonstration (note that when $r = 1$ all the elements of $\mathbf{C}_d$ are one and the inverse does not exist). The inverse covariance matrix requires storage that is proportional to n, while computing the Mahalanobis distance in (6.2) only requires order n operations. The likelihood in (6.3) also needs the determinant of $\mathbf{C}_d$. As showed in Malinverno and Briggs (2004), an expression of this determinant can be obtained by writting the tridiagonal inverse covariance matrix $\mathbf{C}_d^{-1} = \mathbf{L}\mathbf{L}^T$, where $\mathbf{L}$ is a lower triangular matrix whose determinant is the product of its diagonal elements. The final result for the determinant of the covariance matrix is

$$|\mathbf{C}_d| = \sigma^{2n}(1-r^2)^{n-1}. \tag{6.9}$$

Hence the two hyperparameters σ and r describing the covariance matrix of data errors can be given a wide uniform prior probability distribution and posterior inference can be done to infer the magnitude and correlation of data noise. The two noise parameters will be perturbed along the transdimensional Markov chain and each time a new value is proposed, $\mathbf{C}_d^{-1}$ and $|\mathbf{C}_d|$ will be perturbed accordingly to compute the likelihood value of the proposed model.

With this form of correlation, the covariance matrix of data errors is well-conditioned and there are stable analytical solutions for $\mathbf{C}_d^{-1}$ and $|\mathbf{C}_d|$. However, as shown below, this is usually not the case if we use other forms of correlation function. Here the main advantage is that each time we want to perturb r or σ along the random walk, we directly perturb the determinant and inverse without having to numerically compute them from a perturbed $\mathbf{C}_d$, which would be too computationally expensive.

Finally, we detail the general form of the proposal function used to perturb models in our Monte Carlo algorithm. At each step of the Markov chain, one type of model perturbation is uniformly randomly selected from the following 6 possibilities, and the proposed model is either accepted or rejected according to the

usual acceptance probability:

1. Change a velocity value in one layer. Randomly select a layer and propose a new S-wave velocity using a Gaussian probability distribution centred in the current value.

2. MOVE: Randomly pick one layer and perturb the position of its nucleus according to a Gaussian probability distribution centred in the current location.

3. BIRTH: Add a new layer. Its location is drawn from the uniform prior distribution and its velocity is drawn from a Gaussian distribution centred at the current value.

4. DEATH: remove a random layer.

5. Change the noise magnitude σ. The proposed value is drawn randomly from a Gaussian distribution centred on the current value.

6. Change the noise correlation r. The proposed value is drawn randomly from a Gaussian distribution centred on the current value. The major advantage is that the likelihood of the perturbed model can be directly obtained without having to numerically estimate $\mathbf{C}_d^{-1}$ and $|\mathbf{C}_d|$.

The variances of the Gaussian proposal functions are parameters to be fixed by the user. As shown before, the magnitude of perturbations does not affect the solution but rather the sampling efficiency of the algorithm. Thus the width of proposal distributions are 'tuned ' by trial-and-error in order to have an acceptance rate as close to 44% for each type of perturbation (Rosenthal, 2000).

6.2.3.2 2^{st} type of noise parameterization

A second type of parameterization that is commonly used to model the noise in RFs is a Gaussian correlation function

$$c_i = r^{(i^2)}. \tag{6.10}$$

Here again $r = c_1$ is the correlation between two adjacent data in the series. Note also that r can be written as $r = e^{\frac{-1}{2\rho^2}}$ where ρ^2 is the variance of the Gaussian correlation function, which is shown in Figure 6.2b. Realisations of such a noise for different values of r are shown in right panels of the same figure. Compared to the first type of correlation, here there are no high frequency components, and hence this

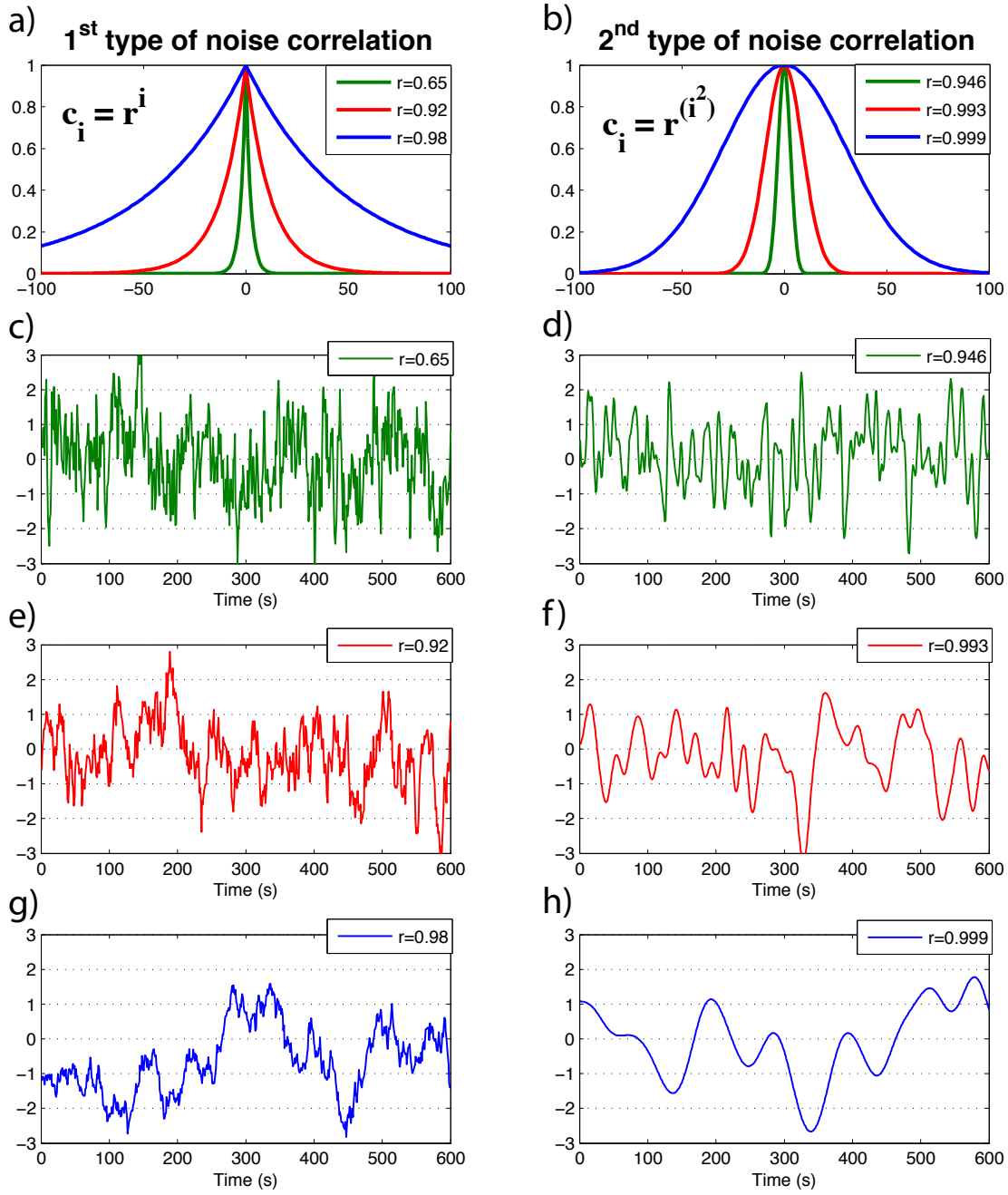


Figure 6.2: Top panels (**a** and **b**) show the two types of correlation functions c_i for different values of r. These symmetric functions represent the correlation between a point in the time series and its neighbours. In order to compare the two forms of noise assumed in this study, we plot different realisations of noise for different values of r and a fixed standard deviation $\sigma = 1$. Left panels (**c**, **e** and **g**) show different realisations with the first type of correlation (exponential decay), whereas noise vectors plotted in right panels (**d**, **f**, and **h**) are generated with a Gaussian correlation. The second type of noise in right panels seem to be closer to what is observed on RF before the first arrival, however this way of parameterising the noise turns out to be more difficult to implement for a Hierarchical Bayes inversion.

form of noise clearly seems closer to what is observed in receiver functions before the first P-arrival. This is because a Gaussian filter is used in the deconvolution process to remove high frequency noise that has high amplitude and which blurs the signal. Note that a white noise which has been convolved with a Gaussian filter will have exactly the same structure as our second type of noise.

Although this form of correlation clearly appears more relevant for our problem, it turns out that the associated correlation matrix $\mathbf{R}$ in (6.6) is highly ill-conditioned, and hence there are no stable analytical formulation for its inverse and determinant. Therefore $\mathbf{C}_d^{-1}$ and $|\mathbf{C}_d|$ have to be numerically computed with SVD decomposition and removal of a large number of small eigenvalues that destabilise the process. Unfortunately an SVD decomposition of a $n \times n$ matrix is computionally expensive and cannot be carried out each time $\mathbf{C}_d$ is perturbed along the random walk. As a result, the correlation r needs to be fixed and cannot be treated as an unknown in the inversion. However, the magnitude of data noise σ^2 can be perturbed without having to re-invert each time $\mathbf{C}_d$. This is because we have

$$\mathbf{C}_d^{-1} = (\sigma^2 \mathbf{R})^{-1} = \tfrac{1}{\sigma^2} \mathbf{R}^{-1} \tag{6.11}$$

and

$$|\mathbf{C}_d| = |\sigma^2 \mathbf{R}| = |\sigma^2 \mathbf{I}_d| \times |\mathbf{R}| = \sigma^{2n} |\mathbf{R}|. \tag{6.12}$$

In this way $\mathbf{R}^{-1}$ is computed once at the beginning and remains fixed along the Markov chain, and the level of data noise σ^2 can be treated as an unknown since each time a new value is proposed, $\mathbf{C}_d^{-1}$ and $|\mathbf{C}_d|$ can be computed from (6.11) and (6.12) without having to redo any SVD decomposition.

Therefore, here one can only invert for the magnitude of noise σ^2 whereas its correlation r is predetermined by the user. Here we set the variance ρ^2 of the Gaussian correlation function equal to the variance of the low-pass Gaussian filter (in the time domain) used in the deconvolution process.

The proposal probability distribution used to perturb the model along the random walk is the same as previously described, but with only 5 possibilities as here r is fixed. Note that this second type of inversion, where only the magnitude σ is inverted for, can be carried out with any given correlation matrix $\mathbf{R}$, provided that an approximation of its inverse is given as an input of the algorithm.

6.3 Inversion of synthetic receiver functions

We first test our algorithm with synthetic data computed from a known velocity model made of 6 horizontal layers. Figure 6.5b shows the true model (red line), which presents two major features often targeted by RF studies : a low S-wave velocity layer in the crust (between 10-20 km) and a strong velocity increase at the moho (at about 30 km depth). A synthetic receiver function is calculated from the true model and a correlated random noise is added, which results in the 'observed' receiver function (see Figure 6.6). The correlation function used to generate the synthetic noise decays exponentially, and hence inversions carried out in this section assume the first type of correlation. The algorithm is used in parallel and about 100 Markov chains (starting at different random points) sample the model space simultaneously and independently.

6.3.1 Generating a correlated random noise

The synthetic noise vector is constructed by drawing a sample from a n-dimensional normal distribution with zero mean and covariance matrix $\mathbf{C}_d$. The covariance is defined with the first type of correlation with values $\sigma_{true} = 2.5 \times 10^{-2}$ and $r_{true} = 0.85$ (these values will be unknown in the inversion). To draw values from the distribution, the covariance matrix $\mathbf{C}_d$ is diagonalised by SVD decomposition:

$$\mathbf{C}_d = \mathbf{USV}^T. \tag{6.13}$$

In this way, the Gaussian function is expressed in a coordinate system where it is uncorrelated, i.e. with a diagonal covariance matrix $\mathbf{S}$. Since $\mathbf{C}_d$ is symmetric and positive-definite, $\mathbf{U}$ and $\mathbf{V}^T (= \mathbf{U}^{-1})$ represent the bijective transformation and its inverse used to change coordinate systems. In the new coordinate system, the Gaussian distribution is uncorrelated, and then it is straight-forward to draw a random vector $\mathbf{S}\varepsilon$, where $\varepsilon = (\varepsilon_1, ..., \varepsilon_n)^T$ is a vector of independent normal random numbers with zero mean and variance one (i.e. a Gaussian white noise time series). Then, this noise vector is rotated back into the physical coordinate system.

$$\epsilon = \mathbf{US}\varepsilon \tag{6.14}$$

Our noise vector ϵ generated in this way is shown in Figure 6.4d. In order to verify the procedure, we generated in this way a large number N of noise vectors ϵ_i and numerically tested that they were statistically distributed according to our

covariance matrix, that is

$$\mathbf{C}_d = \frac{1}{N} \sum_{i=1}^{N} \epsilon_i \epsilon_i^t \tag{6.15}$$

Note that a more common way to produce correlated noise is to simply apply a low-pass filter to a white noise (e.g. a Gaussian filter). However, as we are here interested in the noise covariance, our scheme is advantageous since we are directly able to generate a noise with any given covariance.

6.3.2 Sampling the prior

Before inverting the synthetic receiver function, we tested the ability of the algorithm to sample a known transdimensional distribution. A convenient way to do that is to set the likelihood to unity and check if the collected models are distributed according to the given prior probability distribution (see section 3.1.6). In our case, we do a similar test where estimated data are set to zero and observations are equal to the generated noise plotted in Figure 6.4d. In this way, any proposed model in the Markov chain predicts a null data vector and residuals are kept equal to the noise along the random walk independently of the model parameters. Residuals are thus constantly equal to the value one would obtain if the proposed model was the true model. Hence, the algorithm is run such that the likelihood distribution is constant in regard to the Earth parameters, and it samples the prior distribution as it is directly proportional to the posterior. However, the likelihood in (6.3) depends on the noise hyperparameters and is maximised when σ and r take the true values used to generate the noise.

Figures 6.3 and 6.4 show results for this test. The frequency distribution of sampled values for S-wave velocities at depth (Figure 6.3a) shows that the algorithm samples the imposed distribution, i.e. a uniform 2D distribution over the range $[V_{min} = 2, V_{max} = 5]$ and $[D_{min} = 0, D_{max} = 60]$. The average velocity value sampled at each depth is shown in Figure 6.3b, and as expected, it is a constant line with value equal to the mean of the uniform prior distribution on the velocity. The frequency distribution of the sampled transition depths (i.e. the layer boundaries), plotted in Figure 6.3c, correctly follows their uniform prior distribution between 0 and 60 km depth. Finally, Figure 6.4a plots the histogram on the number of layers across the ensemble of collected models, which is uniform between 2 and 60 interfaces, as required by the prior.

The posterior distribution on hyperparameters σ and r are shown in Figures

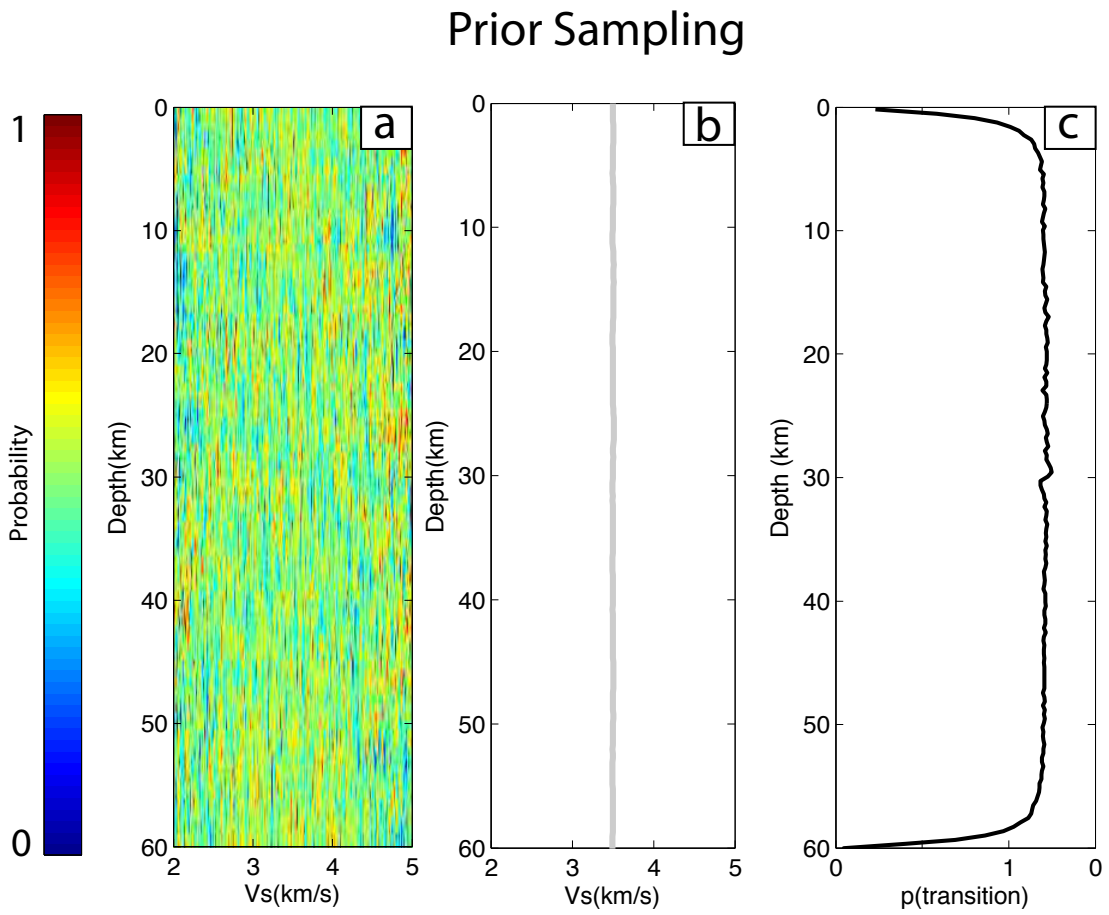


Figure 6.3: (a) Prior sampling of the S-velocity in the Depth interval considered. (b) mean value of the velocity parameter, as retrieved at the end of the sampling. (c) Prior sampling of the Depth of interfaces in the model.

6.4b and 6.4c together with the true values and with the uniform prior distributions in light blue. Note that those are equivalent to the posteriors on hyperparameters conditional on the true values of model parameters, i.e when residuals equal the data noise. Here one can clearly see that the algorithm is able to recover the magnitude and correlation of the noise from a wide poorly informative prior. Notice however that the maximum value of the posterior distribution for noise parameters are not exactly at the true values showed by red lines in Figures 6.4b and 6.4c. This is because the noise vector we have used is only a realisation of a random process. By repeating this experiment a large number of times or by simply using a much longer noise vector, we observe that the posterior expectation on hyperparameters tends towards the true values of noise magnitude and correlation.

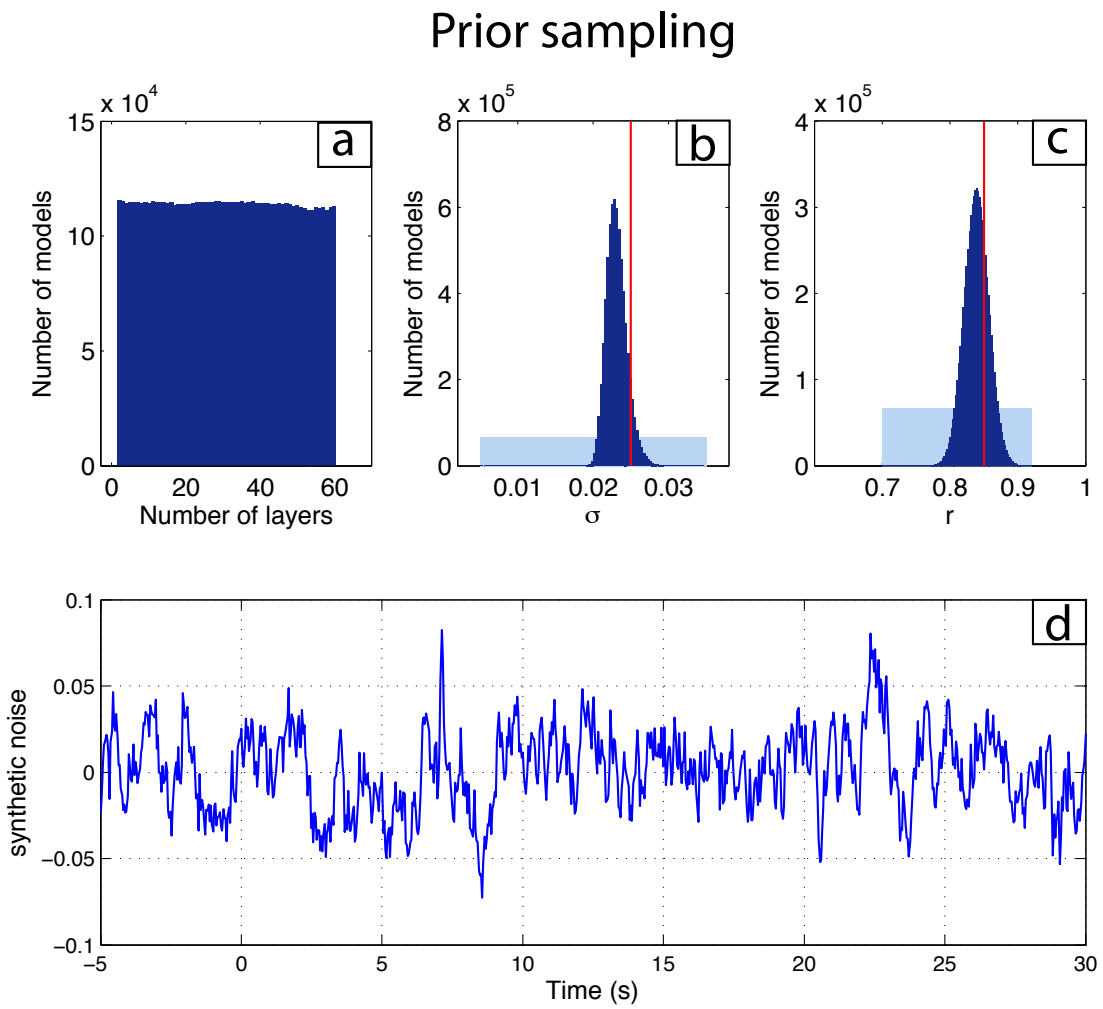


Figure 6.4: (a). Prior sampling of the number of layers in the model. (b) Sampling of the parameter σ which represents the standard deviation of noise. Light blue is the prior, dark blue the posterior and red line the true value used to generate the noise. (c) same as (b) with the parameter r which represents the noise correlation. (d) Synthetic noise used as the observed data vector in the inversion.

6.3.3 Sampling the posterior distribution

In the next step, the full synthetic receiver function is then inverted and posterior inference on model parameters is carried out. A burn-in period that allows the Markov chains to converge is used before samples start to be collected to construct the ensemble solution. As described in previous chapters, the convergence of the algorithm is monitored with a number of indicators such as acceptance rates, and sampling efficiency is optimised by adjusting the variance of the Gaussian proposal functions.

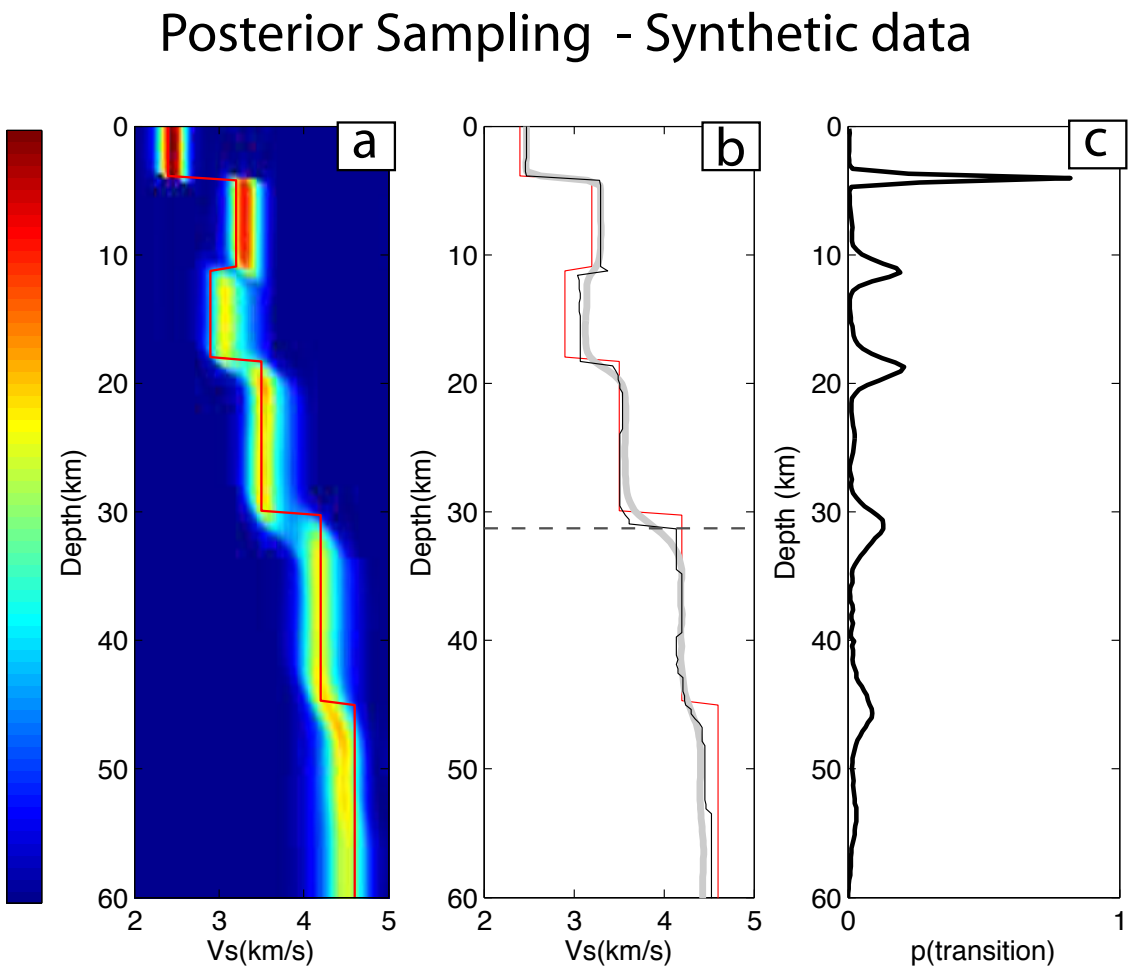


Figure 6.5: (a) Posterior probability distribution for Vs at each depth. Red shows high probabilities and blue low probabilities. The synthetic true velocity model is plotted as a red line. (b) Grey line: mean of the posterior at each depth. Black line: maximum of the posterior at each depth. Red line: true velocity model. (c) Posterior probability for the position of discontinuities.

Figures 6.5-6.8 show results for this test. The observed RF is shown in blue in Figure 6.6d, and it is constructed by adding the noise vector in Figure 6.4d to the 'true' RF (corresponding to the true model) in red in Figure 6.6e. The black line is Figure 6.6e shows the RF estimated from the best fitting model in the ensemble.

At each depth, local information about the velocity model is represented by a complete distribution which can be seen as a marginal distribution of the posterior. These marginal posteriors are shown as a colour density map in Figure 6.5a together with the true model in red (The marginal posterior at 31km depth is plotted in Figure 6.7). It is tantamount to picking a depth and asking the ensemble solution

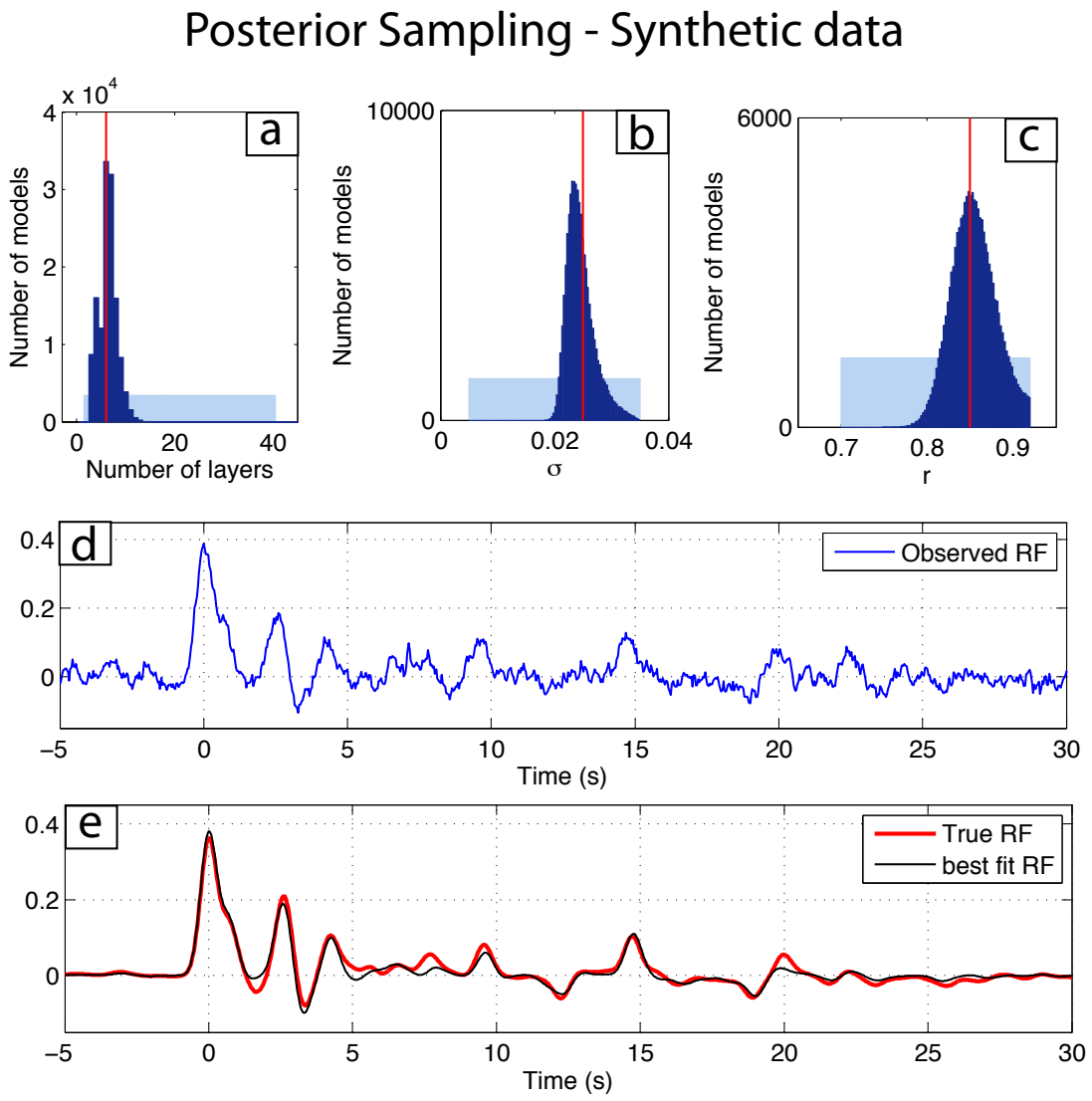


Figure 6.6: (a), (b), and (c): posterior distribution for the three hyperparameters, i.e. number of layers, standard deviation of data noise, and correlation of data noise. (d) Observed receiver function (i.e. true RF + synthetic noise). (e) Red: Synthetic receiver function. Black: RF estimated from the best fit model in the ensemble.

what velocity constraint is given by the data. The marginal posterior is constructed from the density plot (i.e. the histogram) of the ensemble of models in the solution. This density plot is used as a way to visualise the ensemble solution, and it is particularly usefull to picture the constraint we have on Vs. The Vs parameter is better constrained at shallow depths (the marginal posterior in Figure 6.5a as a low width and reaches the highest amplitudes).

Then the marginal posterior at each depth can be used to construct specific

1D models. In Figure 6.5b are plotted the posterior mean model (grey line) which follows the mean of the marginal with depth and the maximum of marginal posterior (MMP) model (black) which follows the maximum of the marginal with depth. Note that these models are merely properties of an ensemble of models that have variable parameterisations, and hence they are 'parameterless' as they cannot be constructed with our initial Voronoi parameterization. Note also that the MMP model is different from the best fitting model in the ensemble, or from the model that maximises the posterior distribution. Here it is clear that those models provide a good estimation of the true model in red.

The RF inverse problem is highly non linear and hence the posterior is far from being a unimodal Gaussian distribution. To illustrate this, we have plotted in Figure 6.7 the marginal distribution on Vs at 31 km depth. This crosssection corresponds to the dashed line in Figure 6.5b, and it is close to the Moho transition in the true model. As a result, the marginal distribution is influenced by velocity values in the two layers on each side of the Moho, and it has two maxima about this two values. Here one can see that the mean value is not representative of the true model whereas the maximum is closer to the true velocity in the lower interface. That is why the posterior mean model (grey line in Figure 6.6b) is smooth whereas the posterior maximum model is better at showing sharp transitions.

If one is interested in assessing the number and position of seismic discontinuities beneath the seismic station, it is possible to examine the ensemble solution from a different point of view and to plot the marginal posterior distribution on the location of interfaces. Figure 6.5c shows a histogram of interfaces depth in the ensemble of models. For each depth, this function represents the probability of having a discontinuity, given the data. This provides useful information on the location of transitions, which can be unclear in other plots. Note that the positions of interfaces are not direct model parameters, and hence this marginal distribution is constructed by projecting the ensemble of sampled Voronoi models into a different model space. Here the five transitions present in the true model are well recovered, and again shallower structures are better resolved than deeper ones. Posterior information on the complexity of the seismic structure is obtained by plotting the histogram of the number of layers in the sampled models (see Figure 6.6a), and most models in the ensemble have 6 layers, which is the correct number of cells present in the true model.

Figures 6.6b and 6.6c show the marginal posterior on the noise parameters together with prior distribution in light blue. The true values σ_{true} and r_{true} used to

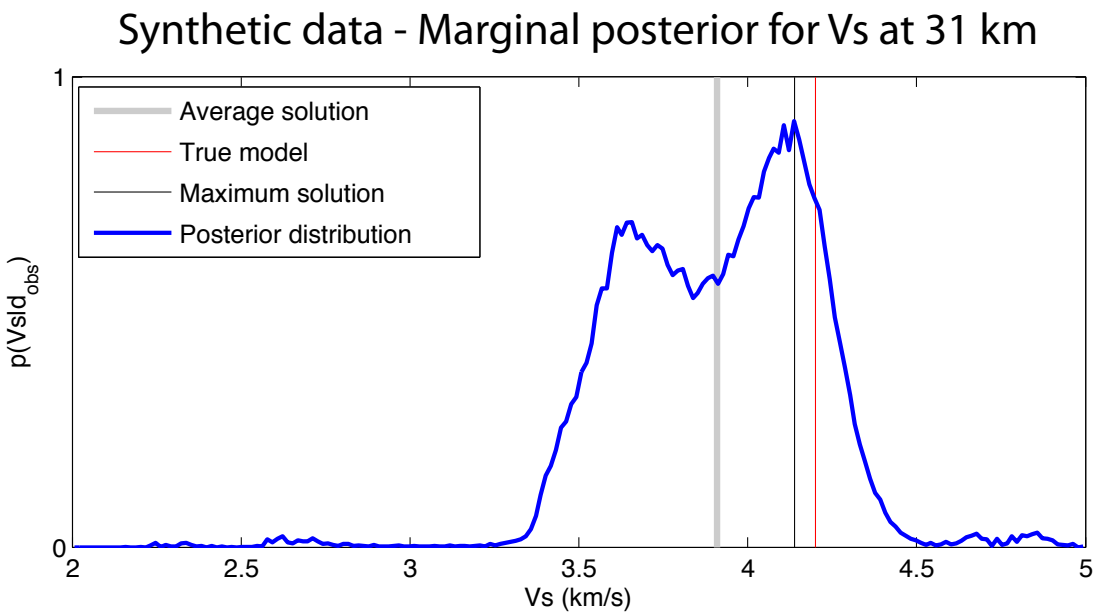


Figure 6.7: Marginal posterior for Vs at 31 km depth (i.e. slightly after the Moho discontinuity). The distribution is clearly influenced by both Vs value taken above and under the discontinuity. Grey line shows the mean value. Black line is the maximum of the marginal and the red line is the true value (i.e. the Vs value in the first layer of the mantle in the true model).

generate the synthetic noise are showed in red. With minimal information on data errors, and on the complexity of the true model prior to the inversion, the Hierarchical Bayes procedure has been able to infer the magnitude and correlation of data noise, which quantified the required level of data fit, and thus the number of model parameters needed in the inversion.

Finally, we give an example of trade-off assessment between two model parameters, that is Moho depth vs S-wave velocity in the last layer of the crust. Again, here the depth of an interface is not strictly a model parameter but a useful feature that can be picked in any sampled model. The crust-mantle transition is defined in the Voronoi models as the closest discontinuity to 30 km. We aknowledge that this definition for the Moho is loose, and instead we could choose to look at the strongest discontinuity near 30km depth. However here the main purpose it to illustrate trade-off assesment between selected seismic properties.

Figure 6.8 shows the 2D marginal posterior for the selected pair of parameters, which is obtained from the 2D histogram over the ensemble of models. White dashed lines shows the true values for both parameters. In this way one can extract accurate and quantifiable information from the ensemble about the constraints and

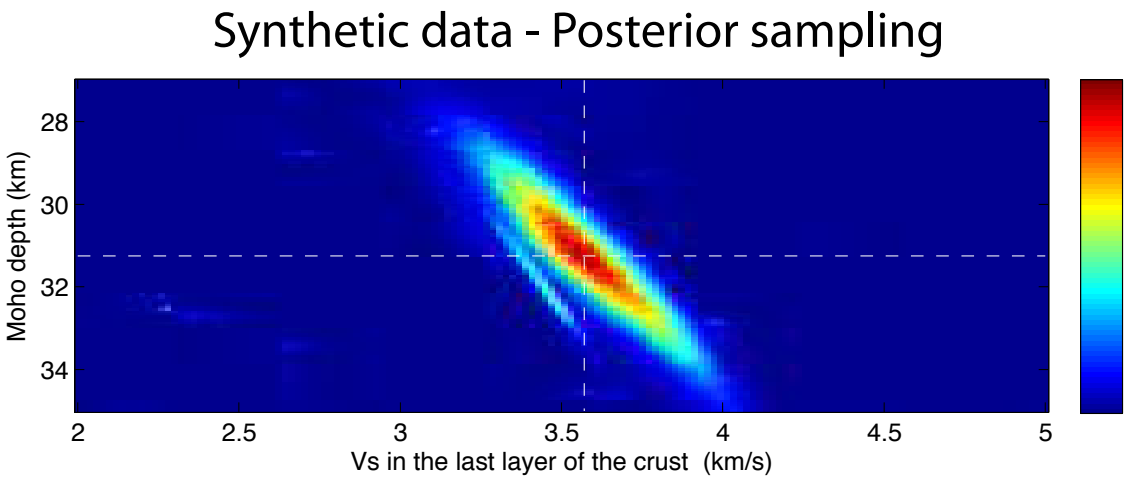


Figure 6.8: Posterior 2D marginal for the parameters representing depth of Moho and Vs in the last layer of the crust. (The Moho is defined as the closest interface to 30 km in the Voronoi models). White dashed lines show true values for both parameters.

correlation for these parameters. This trade-off means that data are fit equally well when Moho is deeper and Vs is higher or *vice-versa*. This result is not surprising in that the thicker the crust gets, the slower the entire model, which is compensated by increasing the velocity in the crust.

6.3.4 Solution with incorrectly assigned noise level

In order to demonstrate the necessity of using hierarchical models when carrying out a transdimensional inversion, we repeat the experiment of the previous section with exactly the same data vector but using a fixed covariance matrix of data errors. As shown in the introduction of this chapter, estimating the data noise in receiver functions is a difficult task and in the absence of information, practitioners may well use erroneous estimates. We place ourselves in such a situation and perform a conventional transdimensional inversion using a noise covariance matrix with incorrect values of σ and r. Here σ has been underestimated by 40% (we use 0.015 instead of 0.025), and r has been overestimated by 8% (we use 0.92 instead of 0.87). In this test the observed RF remains as previously and we observe the effects of misestimating the data noise.

Figures 6.9 and 6.10 show results for this test. Since the magnitude of noise has been underestimated, the required level of data fit is increased, and thus the algorithm automatically adds more cells than necessary and 'overfits' the observed

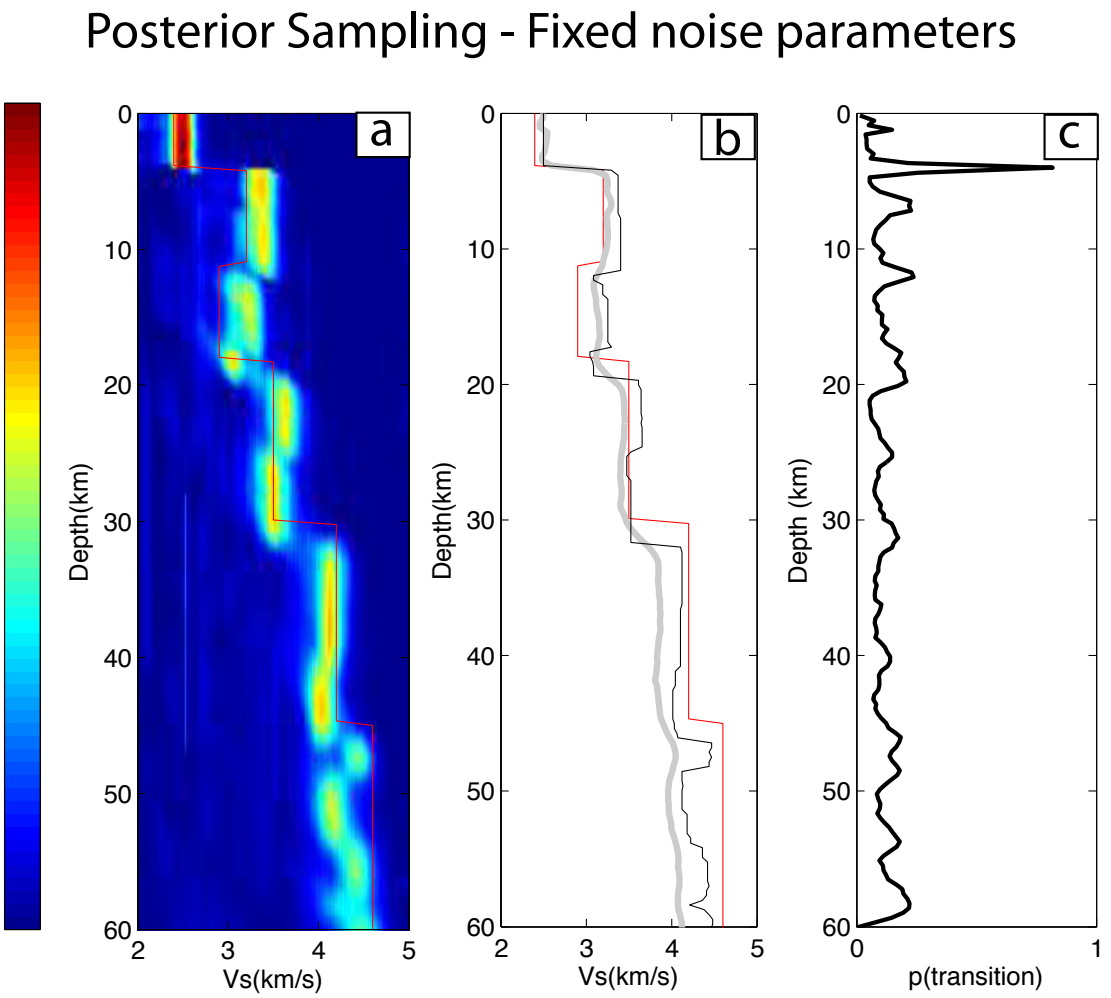


Figure 6.9: Same as Figure 6.5 but here the noise parameters σ and r used in the inversion are fixed and respectively under and over estimated. The posterior approximation of the true model is clearly worsened.(a) Posterior probability distribution for Vs at each depth. Red shows high probabilities and blue low probabilities. The synthetic true velocity model is plotted as a red line. (b) Grey line: mean of the posterior at each depth. Black line: maximum of the posterior at each depth. Red line: true velocity model. (c) Posterior probability for the position of discontinuities.

RF by fitting the noise. The expected number of layers in the model in Figure 6.10a is 15, which is the double of the true value. This results in an inferred model that is overcomplicated, and Figure 6.9c shows that location of transitions are not recovered as well as with the Hierarchical Bayes where correct values of noise are inferred by the algorithm. Even though features are generally degraded in comparison with Figure 6.5, the most resolvable elements (i.e. location of shallow discontinuities) remain resolvable. This example demonstrates that, by allowing the user to formulate the

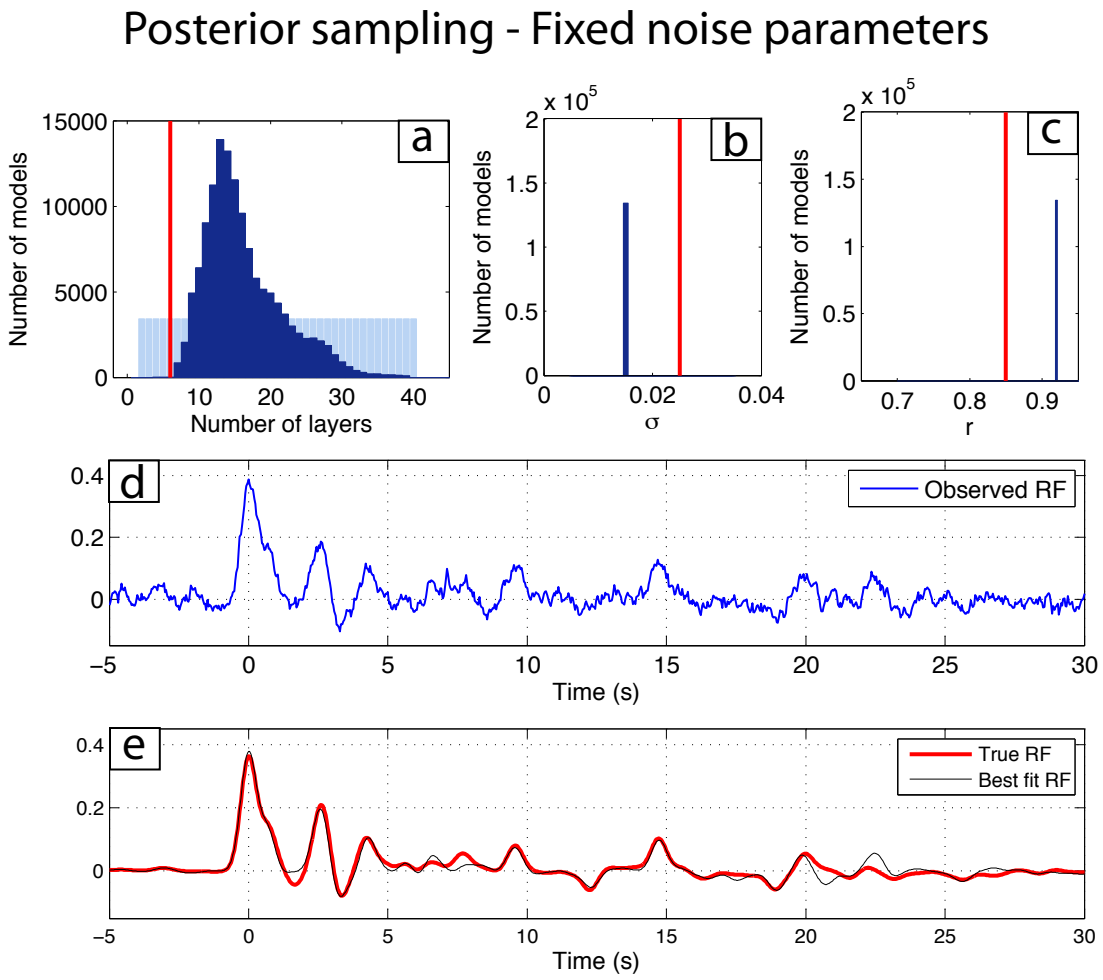


Figure 6.10: Same as Figure 6.6 but here σ is fixed and underestimated and r is fixed and overestimated. (a) Posterior distribution on the number of layers. Red line shows the true value. The model complexity is overestimated. (b), and (c): fixed values used in the inversion for the standard deviation of data noise σ, and correlation of data noise r. True values are showed in red. (d) Observed receiver function (same as in Figure 6.6d). (e) Red: Synthetic receiver function. Black: RF estimated from the best fit model in the ensemble.

full state of uncertainties he has about data noise, a hierarchical Bayesian procedure provides better estimates than when the data noise is mis-estimated.

6.4 Inversion of field measurements

To demonstrate how our transdimensional inversion algorithm fares on real data, we apply to broadband waveforms from a station located in Qiongzhong (19.039Â°N,

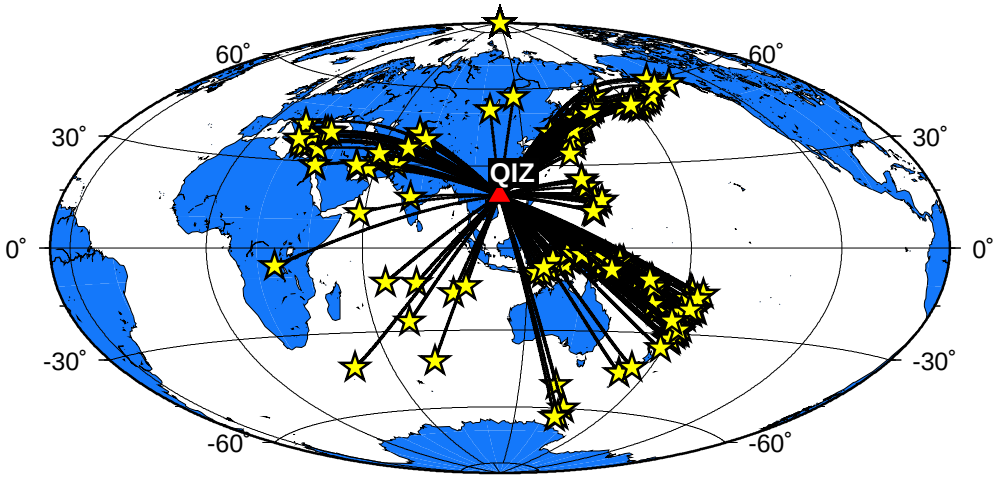


Figure 6.11: Global distribution of events for which broadband data were available at the station.

109.843Â°E) which is situated on the Hainan island in China (Qiu *et al.*, 2002). Although it is an island station, the quality of waveforms is rather good for a receiver function study. Previous studies have shown the crust beneath eastern China is as thin as 31-33 km and the underlying Moho is relatively flat and sharp (Chen *et al.*, 2010; Tkalčić *et al.*, 2010).

The map in Figure 6.11 shows the station location and the distribution of recorded earthquakes. Most events are coming from Japan and Aleutian islands from the northeast, and from Tonga-Fiji and Indonesia regions from the southwest. There is also some minor seismicity coming from the Middle East and the Indian Ocean. More than 50% of earthquakes (157 out of 302 events) are located in the South-East, i.e. between 90 deg and 180 deg. Only events from this direction are used in this study, and their ray path projections are highlighted in red in a more detailed map shown in Figure 6.12. For each event, all three components were cut for the same time window and rotated to radial and tangential. Finally, radial receiver functions were calculated using the time domain iterative deconvolution procedure proposed by Ligorria and Ammon (1999) for a low pass Gaussian filter with parameter $a = 1.0$ and $a = 2.5$ (results will be show for both values).

Compared to the widely used 'water-level' deconvolution technique in frequency domain (Clayton and Wiggins, 1976), this method can produce more stable results, but at the price of longer computation times. It is also free of complex relationships between water-level values and the resulting receiver functions.

The 157 obtained receiver functions are not all coherent, resulting from events

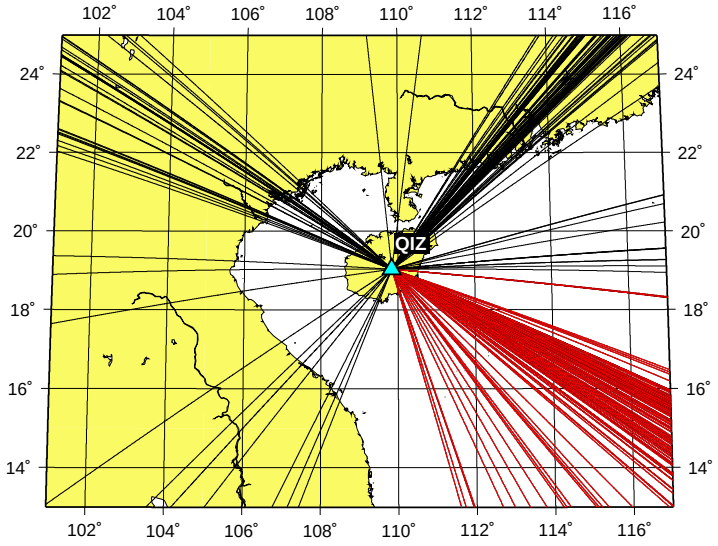


Figure 6.12: A detail from Figure 6.11, with highlighted ray paths used to obtain an averaged receiver function shown in Figure 6.13.

arriving at the station with different angle of incidence. In order to select receiver functions that are mutually coherent and could be stacked to determine the observed RF, we follow the statistical approach developed by Tkalčić *et al.* (2006), and carry out a coherency test and empirically determine 2 factors. The first factor is the cross-correlation coefficient limit, above which we call two RFs similar (for this parameter we choose 0.9). The second factor is the percentage of all other RFs from the starting group of RFs that must be correlated at 0.9 level or higher. For RFs produced with a Gaussian filter with parameter $a = 2.5$, we choose 5%. In other words, we select only those RFs that correlate at 0.9 or higher with at least 5% of all other RFs.

The starting group of RFs is showed in black in Figure 6.13, the selected are in blue, and final average of all selected RFs is in red. For $a = 2.5$, only 40 waveforms out of 157 satisfy this criterion. So, each one of 40 selected RFs correlates at 0.9 level with at least 5% of all RFs (Figure 6.13). For 8%, we would get 21 mutually coherent waveforms and for 10%, we would only get 7 coherent waveforms which are obviously not enough to create a robust average. For $a = 1.0$, waveforms are much more coherent as we get 59 mutually coherent waveforms for only a factor of 25% (Figure 6.13).

The average RF is used as the data vector in our algorithm. In the direct search algorithm, synthetic RFs predicted by proposed models are computed as mentioned

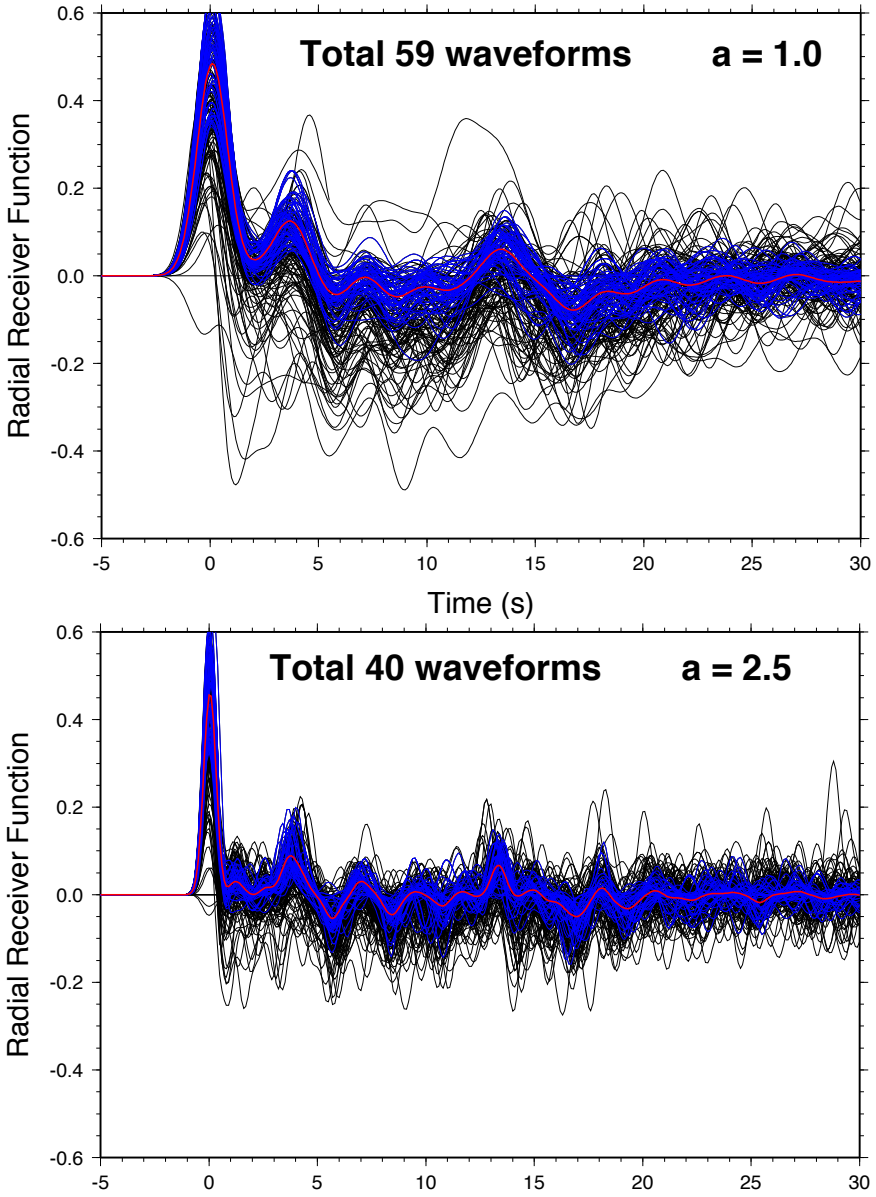


Figure 6.13: Observed receiver functions for the two types of filter used in the deconvolution process. Black lines show RFs for each of the 157 recorded events. Blue lines represent the selected coherent RFs used to produce the average RF in red.

above with a simple frequency domain deconvolution technique and a water level of 0.001 (Helmberger and Wiggins, 1971). We acknowledge here that the observed and estimated data vectors in our inversion are computed with different deconvolution methods which provide slightly different results. Ligorria and Ammon (1999) and Lombardi (2007) showed, that in the presence of noise, frequency domain deconvo-

lution shows long period instabilities with amplitude distortion visible within the first 10 s. Although it would be more appropriate to use an iterative time domain deconvolution for the forward calculation of receiver functions, the discrepancy between the two deconvolution methods, and hence the inability for estimated data to fit the observations will be treated here as errors from the forward model and automatically included in the noise covariance matrix by our noise hyperparameters.

The inversion was run with both types of noise parameterization presented in section 2.3. The first type of noise parameterization enables us to invert both for magnitude and correlation of noise, although the correlation function $c_i = r^i$ erroneously assumes high frequency components in the noise which have obviously been cut by the Gaussian filter used in the deconvolution. In the second type of inversion, a Gaussian correlation function $c_i = r^{(i^2)}$ is kept fixed and we only invert for the magnitude of data noise. In this case the correlation function is simply taken equal to the impulse response of the Gaussian low-pass filter with parameter a used in the deconvolution, thus r is empirically kept fixed equal to $r = e^{-a^2}$.

Figures 6.14 to 6.17 show results obtained with $a = 1.0$ for both types of noise parameterisations. Although results appear consistent, the two types of noise parameterization clearly result in different posterior distributions, and this demonstrates the importance of assumptions made about the data noise. Figures 6.18 to 6.21 show results for a Gaussian filter of $a = 2.5$. In this case the observed RF contains more high frequencies and it is thus modeled with more layers. Note that in the case of first type of noise parameterization, the expected values for the hyperparameter r are close to one (Figures 6.14c and 6.18), with the maximum of the marginal posterior for r asymptoting towards one. When r tends towards one, the estimated data noise becomes more and more correlated, although high frequency components remain present in the noise (see Figure 6.2). When $r = 1$ all the elements of $\mathbf{C}_d$ are one and the inverse does not exist.

In the case of the second type of noise parameterization, the parameter r has been fixed at the outset, and to validate this choice, we compare the residual waveform (observed - predicted) for the best fitting model in the ensemble solution, to a realisation of a random noise generated from the inferred expected $\mathbf{C}_d$. If the choice of r is adequate, the noise realization and residuals should have similar properties (i.e. variance and smoothness). This is because the data noise is defined as the component of the measurements that cannot be explained by $g(\mathbf{m})$. From visual inspection, one can see in Figure 6.22 that residuals and estimated noise are similar for a=2.5. While this is a qualitative test, note that posterior error validation

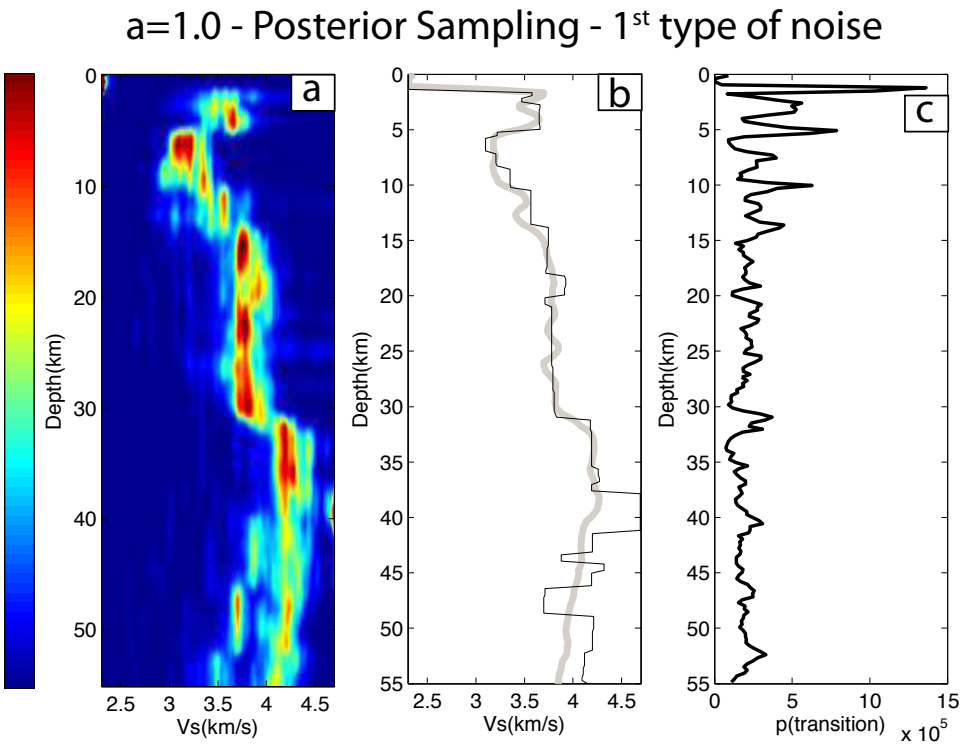


Figure 6.14: (a) Posterior probability distribution for Vs at each depth. (b) Grey line: mean of the posterior at each depth. Black line: maximum of the posterior at each depth. (c) Posterior probability for the position of discontinuities.

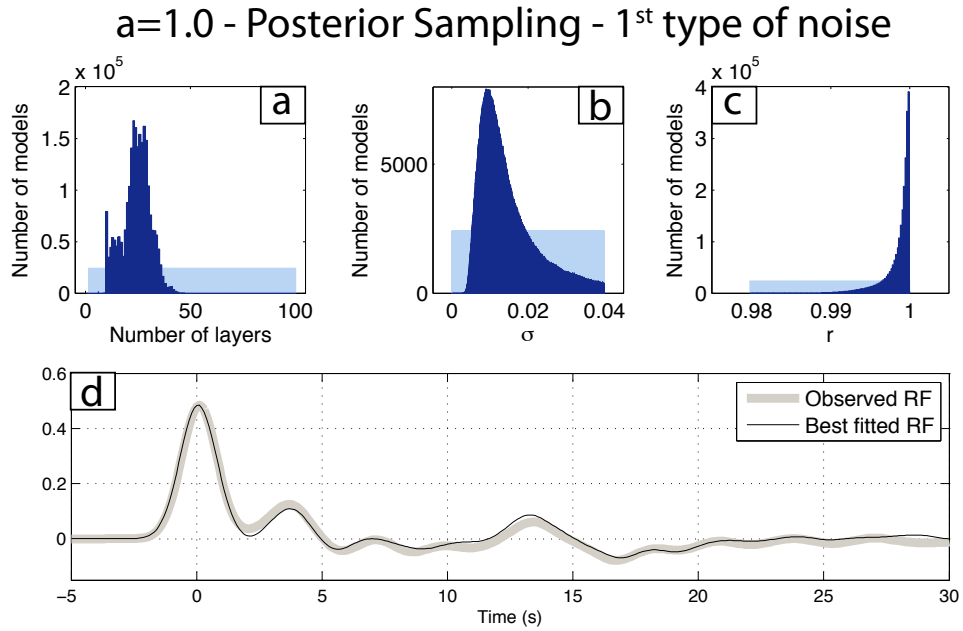


Figure 6.15: (a), (b), and (c): Posterior distribution for the three hyperparameters, i.e. number of layers, standard deviation of data noise, and correlation of data noise. (d) Thick grey line: Observed receiver function. Black line: RF estimated from the best fit model in the ensemble.

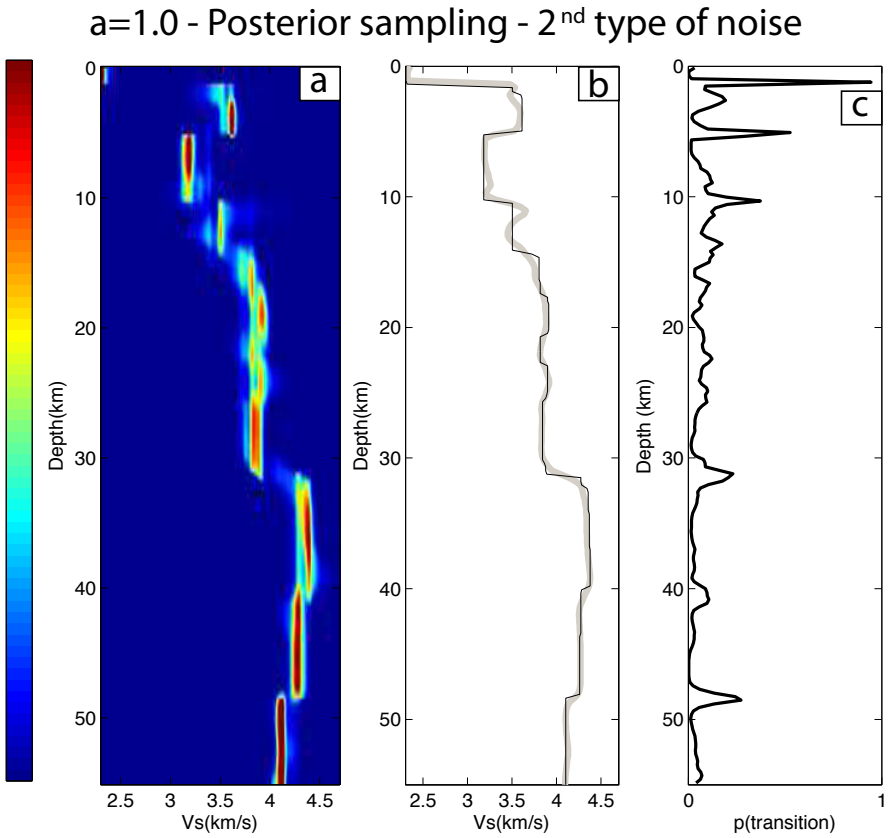


Figure 6.16: (a) Posterior probability distribution for Vs at each depth. (b) Grey line: mean of the posterior at each depth. Black line: maximum of the posterior at each depth. (c) Posterior probability for the position of discontinuities.

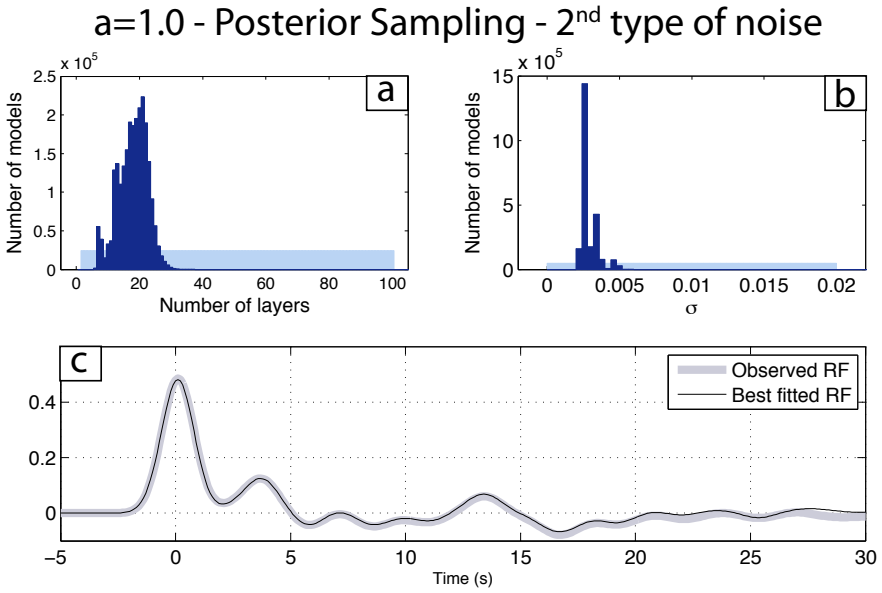


Figure 6.17: (a), (b), and (c): Posterior distribution for the two hyperparameters, i.e. number of layers, standard deviation of data noise (note that here the noise correlation r is kept fixed during the inversion. (d) Thick grey line: Observed receiver function. Black line: RF estimated from the best fit model in the ensemble.

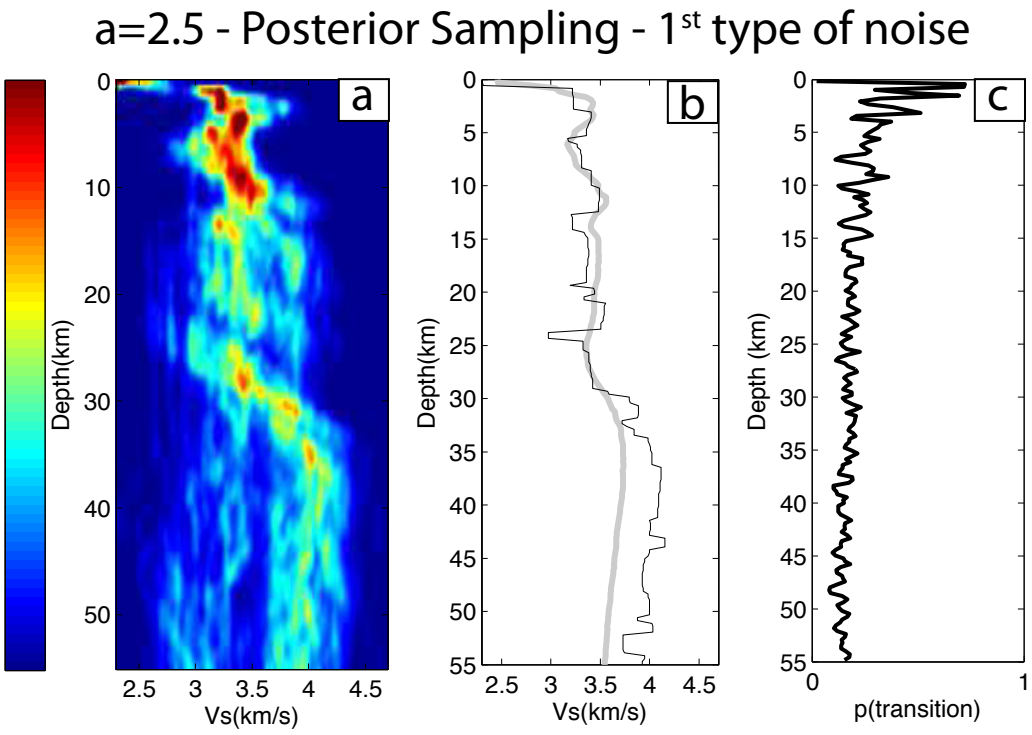


Figure 6.18: (a) Posterior probability distribution for Vs at each depth. (b) Grey line: mean of the posterior at each depth. Black line: maximum of the posterior at each depth. (c) Posterior probability for the position of discontinuities.

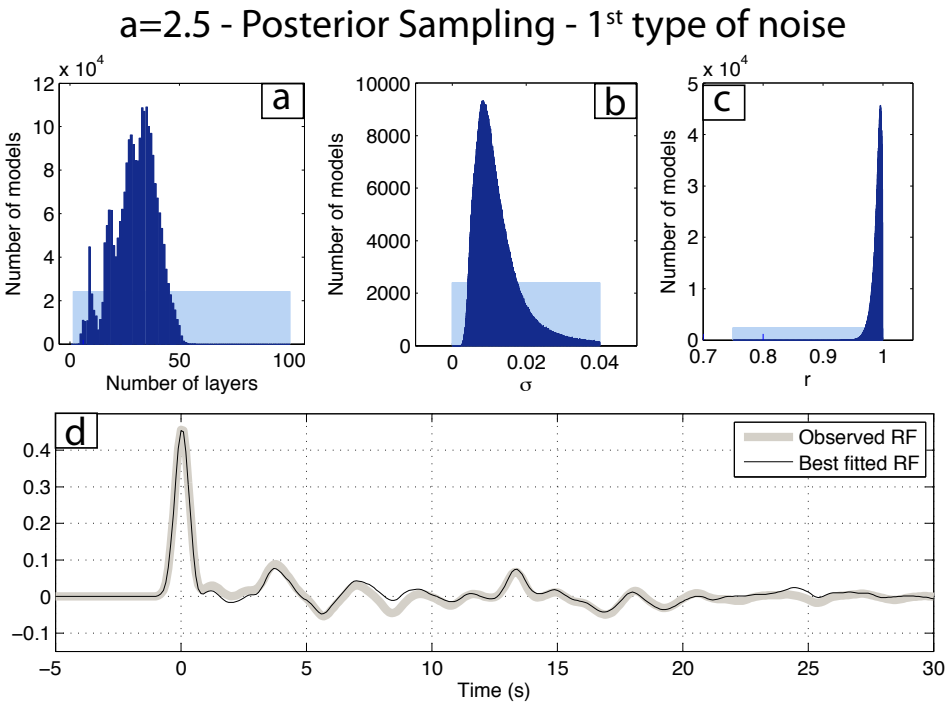


Figure 6.19: (a), (b), and (c): Posterior distribution for the three hyperparameters, i.e. number of layers, standard deviation of data noise, and correlation of data noise. (d) Thick grey line: Observed receiver function. Black line: RF estimated from the best fit model in the ensemble.

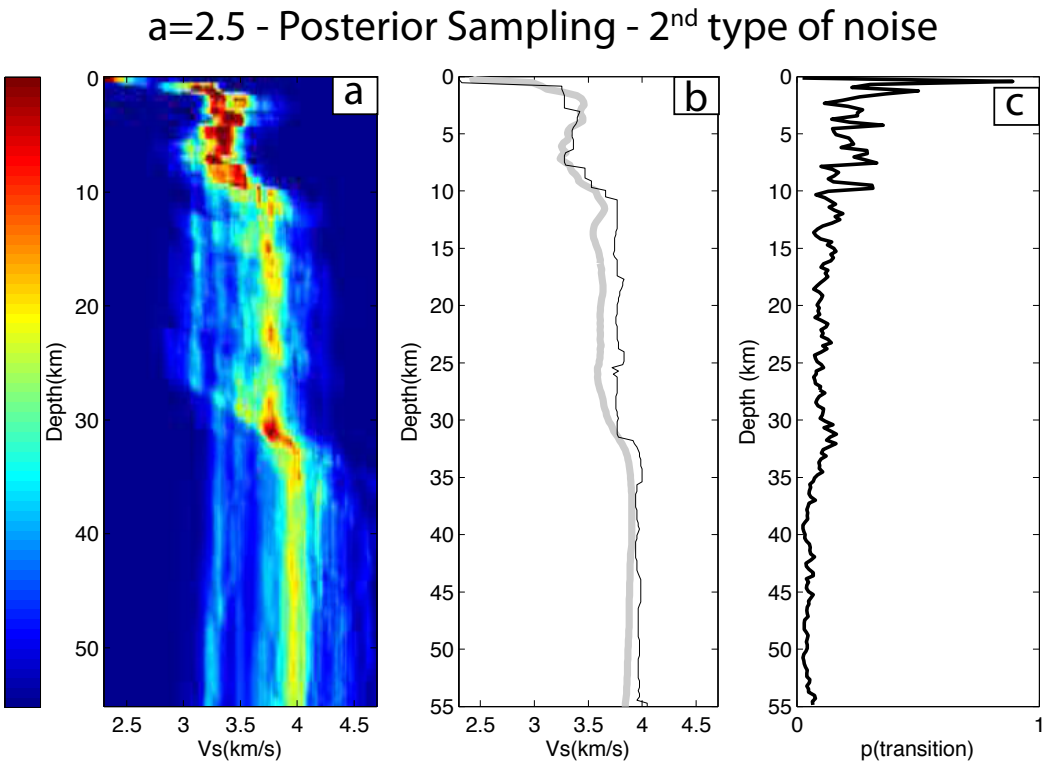


Figure 6.20: (a) Posterior probability distribution for Vs at each depth. (b) Grey line: mean of the posterior at each depth. Black line: maximum of the posterior at each depth. (c) Posterior probability for the position of discontinuities.

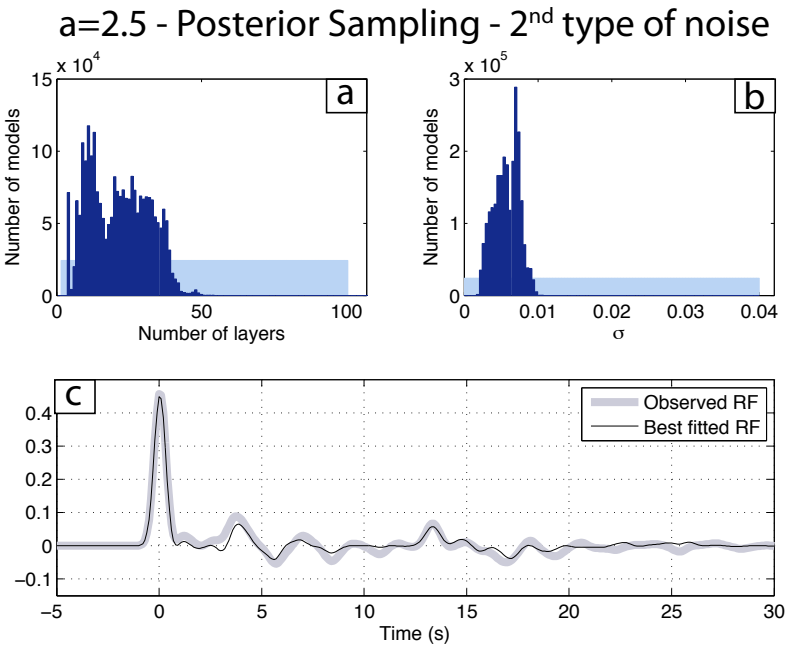


Figure 6.21: (a), (b), and (c): Posterior distribution for the two hyperparameters, i.e. number of layers, standard deviation of data noise (note that here the noise correlation r is kept fixed during the inversion. (d) Thick grey line: Observed receiver function. Black line: RF estimated from the best fit model in the ensemble.

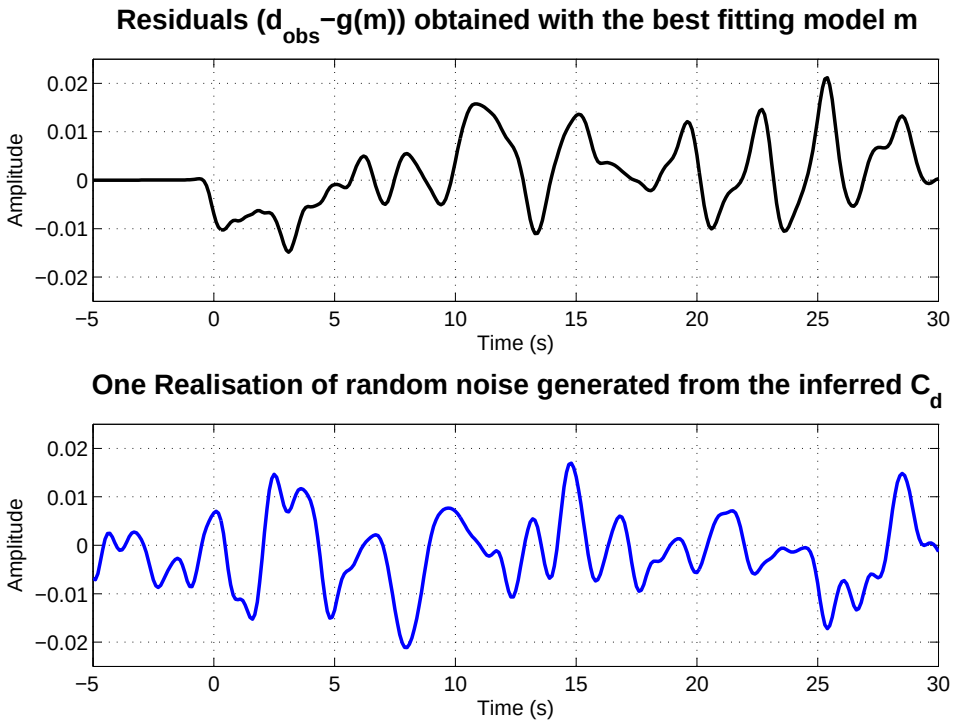


Figure 6.22: Comparison of estimated noise from residuals for the second type of parameterization and a=2.5 (Inversion results shown in Figures 6.20 and 6.21). Top: residual waveform (observed- predicted) for the best fitting model. Bottom: example of a random noise realisation generated from the expected hyperparameter σ and the fixed parameter r

can also be carried out by applying quantitative tests to residuals resulting from one model (see Dettmer *et al.*, 2009). However, whether an individual model can be considered representative of an ensemble solution is questionable. Hence, data residuals can be examined based on ensembles, that is a large random subset of the posterior can be used to compute a large sample of data residuals, which are then used to examine the assumptions about data errors (Dettmer *et al.*, 2010).

Almost all the computed S-wave velocity models are characterised by a very low velocity uppermost structure with a mean Vs = 1.75 km/s in the first kilometer of the crust. These low values are interpreted to be related to the presence at surface of either unconsolidated sediments or weathered exposed rocks. The upper-crust (0-15 km) shows complex structures and is characterised by the presence of velocity inversions. The S-wave velocity in the second part of the crust (15-30 km) is relatively constant with a mean $Vs = 3.75$ km/s, although the inversion models also show some small velocity steps at these depths. For $a = 1.0$, it is possible to

identify a relatively sharp crust-mantle transition around 31 km depth. However, for $a = 2.5$, the Moho is characterised as a gradient transition zone over a depth range of 30-35 km. The inferred seismic models are fairly consistent with the results of other studies (Qiu *et al.*, 2002; Chen *et al.*, 2010; Tkalčić *et al.*, 2010). Although our analysis of the geological implications is rather limited, the aim here is to argue that the ensemble solution produced by the transdimensional approach is in good agreement with the main geological features beneath the receiver.

6.5 Conclusion and future work

Teleseismic receiver function analysis is now a well-established seismological technique, and a large number of schemes have been implemented in last 30 years to infer seismic structure beneath broadband stations. Here we have presented a novel RF inversion method where a Bayesian formulation is employed to produce a multidimensional posterior probability distribution, and each model parameter can be described with a full probability density function. While the variance of the posterior can be used to assess uncertainty on model parameters, the posterior covariance directly quantifies the trade-offs (i.e. the correlation) between parameters. Hence the posterior can be examined from several point of views to infer different properties of the model (e.g. depth of transitions, mean Vs value at one depth, number of layers, etc).

The parameterization of the Earth is adaptive and information is extracted from an ensemble of models with a variable number of layers. While Piana Agostinetti and Malinverno (2010) published a transdimensional inversion of receiver function data, here we have extended their formulation by allowing estimation of measurement noise properties. Therefore, beyond the transdimensional character of the inversion, an original feature of this study is that little needs to be assumed about the covariance matrix of data errors. This matrix plays a capital role in a Bayesian problem, since it directly determines the form of the posterior. The noise covariance represents the level and correlation of data noise, which in the case of RFs can be difficult to quantify. However in this work, instead of using an inaccurate approximation of data noise, our philosophy is to let the data infer their own degree of uncertainty by treating the magnitude and correlation of noise as unknowns in the problem.

We have here focused on the mathematical problem and illustrated the algorithm in simple situations where a number of approximations have been made. We

only inverted for S-wave velocity structure while considering Vp/Vs ratio constant throughout the velocity model. An obvious improvement of the algorithm would be to also consider Vp/Vs ratio in each layer. In addition, layers have been assumed homogeneous and horizontal and it would be possible to treat anisotropy, slope of discontinuities, and lateral variations as unknowns in the problem. These improvement could be achieved by using densely spaced arrays, by including earthquake waveforms from a wide range of backazimuth, and using more sophisticated forward solvers.

Furthermore, there have been recent studies in which RFs have been jointly inverted with surface waves dispersion for crustal structure (e.g. Ozalaybey *et al.*, 1997; Du and Foulger, 1999; Julia *et al.*, 2000; Chang *et al.*, 2004; Lawrence and Wiens, 2004; Yoo *et al.*, 2007; Tkalčić *et al.*, 2006). These methods have the advantage of improved sensitivity to absolute velocities compared to RFs alone. However, a general drawback is the definition of the misfit function. The Hierarchical Bayes procedure is expected to be a powerful tool in this situation, as it would be able to naturally weight the contribution of differrent data types in the likelihood function, thus removing the arbitrary choice a weighting factor.

Chapter 7

Conclusions

In this work we have presented a general class of algorithm for geophysical inversion. Detailed summaries and discussion of results have been given at the end of each chapter. The goal of this concluding chapter is to view the method from a broader and more global perspective. We first present the main results of the thesis. Then we briefly describe how these objectives were obtained, and give directions for further work.

7.1 Thesis Achievements

The major issue tackled in this thesis is the need of arbitrary choices made before a geophysical inversion is carried out. These choices are multiple and can take different forms. They include the form of the parameterisation used to discretise the Earth, the number of model parameters, the level of smoothing, or the required data fit (which is directly related to the definition of data noise). These quantities define the formulation of the inverse problem and differ between specific situations. Traditionally, they are arbitrarily determined at the outset or iteratively tuned manually, and by definition affect the final solution.

The work carried out in this thesis consisted in combining different novel methodologies recently developed in the area of statistics such as hierarchical models, Partition Modelling or the reversible jump McMC algorithm, in order to construct a general geophysical inversion strategy that addresses the issue presented above. The philosophy behind our strategy consists in letting, where possible, the data naturally decide these choices that otherwise need to be arbitrarily made at the outset. Instead of rigidly defining the problem, that is precisely specifying what we want to extract from the data, here our philosophy of problem solving can be summarised

by the sentence : "Let the data themself formulate the problem and tell us what information can be inferred, and how it should be done". This includes letting the data assess which part of the data is retrievable information and which part is noise, and hence how much information should be present in the model. We acknowledge that the retrievable information in the data does not depend on the data alone, but also on the forward modeling function $g(\mathbf{m})$ and the associated assumptions (e.g., that the Earth is one-dimensional, that seismic rays are straight, etc.).

Therefore, we have built a method, where the Earth is parameterised using Voronoi cells with mobile geometry and number. The size, position and shape of the cells defining the Earth model, as well as the data uncertainty, are directly determined by the data.

We first showed in the case of 2D tomography, that the method gives promising results in situations where the ray coverage is far from ideal, as it performs better compared to standard methods that use regular parameterisations. Calculations of uncertainty estimates is also possible, and experiments with synthetic data suggested that they are a good representation of the true uncertainty. Computational cost issues have been treated and the algorithm has been optimised and parallelised.

Subsequently, the methodology was applied to seismic tomography in a situation where both the data density and the underlying structure itself contain multiple length scales. Three ambient noise datasets that span the Australian continent at different scales were simultaneously inverted to infer a multiscale tomographic image of Rayleigh wave group velocity for the Australian continent. We showed that the procedure turns out to be particularly useful when dealing with multiple data types that have different unknown levels of noise as the algorithm is able to naturally adjust the fit to different datasets and to provide a velocity map with a spatial resolution adapted to the quantity of information present in the data.

In order to show that the class of algorithm presented in this thesis is not restricted to seismic tomography but is rather a general approach, two applications to 1D problems were considered. The first was an application to palaeoclimatology, where the goal is to infer the position and number of abrupt changes in noisy geochemical records. The second was an application to receiver function waveform inversion. A particular feature of receiver functions is that they are times series, and hence the data noise is correlated. We showed how to 'parameterise' the data covariance matrix and invert for both the magnitude and correlation of noise. The algorithm was first tested on synthetic data contaminated by correlated synthetic noise, and then applied to data collected by a broadband station located on the

Haiman island in China.

7.2 The alliance of two concepts

These objectives have been achieved mainly by the combination of two approaches, each already widely used in geophysical inversion, but rarely employed together, namely adaptive parameterisation and Bayesian inference.

7.2.1 Adaptive parameterisation

Partition Modelling addresses the issue of discretising the Earth as part of the modelling process. The model is parameterised with a variable number of Voronoi polyhedra with mobile geometry throughout the inversion. The advantage of Voronoi cells is that they enable a sophisticated discretisation of the Earth model (with completely unstructured polygons) by means of an ingenuous parameterization simply made up of a few coordinates (i.e.Voronoi nuclei). To our knowledge, this is the first time in seismic tomography that the Voronoi nuclei used to discretise the velocity field are treated as direct model parameters in the inversion.

A simple analogy can be made with photography, where there is a trade-off between the quantity of light arriving at the lens and the pixel size. When a lot of light is available, pixels can be small and good resolution can be achieved. Conversely, a picture taken in the dark will require larger pixels in order to better absorb light. New cameras measure the luminosity prior to taking the picture, and automatically adapt the resolution. However the size of pixels is constant across the image. In this analogy, our scheme would locally measure the luminosity (quantity of information), and consequently locally adapt the size of pixels.

In this thesis we have shown that a Voronoi adaptive parameterization provides a flexible and powerful mechanism for tomographic problems when combined with an ensemble inference approach.

7.2.2 Bayesian Inference

At the same time, a Bayesian framework is used to formulate the problem probabilistically. We use hierarchical models which allow us to treat the number of parameters, as well as the level of data noise, as unknowns in the inversion. The Hierarchical Bayes model is called 'hierarchical' because it has two levels. At the higher level are the 'inversion parameters' such as the model complexity, or the required level

of data fit. These are not directly linked to Earth properties and are called hyperparameters. At the lower level are the 'physical parameters' that represent Earth properties, whose probability is conditional on hyperparameters. Thus, a transdimensional posterior probability distribution is defined both for hyperparameters and the Earth model.

The reversible-jump McMC algorithm (Green, 1995) is employed, and allows efficient exploration of the model space by sampling models of varying dimension. The output of the scheme is an ensemble of models from which properties like a spatial average, variance, or position of discontinuities can be extracted. In this thesis we have explored ways of extracting tomographic images from trans-dimensional ensembles of solutions. Note here that Markov chains is not the only way to carry out nonlinear Bayesian inference. Algorithms such as particle filters (van Leeuwen, 2009) or neural network modelling (Meier *et al.*, 2007, 2009; Maiti and Tiwari, 2010) can be used alternatively. Although these methods give good model uncertainty estimates and can be used in the case of adaptive parameterization, Markov chains remain the only tool to date for transdimentional inference.

7.2.3 Notable features of the methodology

The alliance of the two methodologies (i.e. the dynamic parameterisation together with the Bayesian framework) allows the emergence of some remarkable and profitable features, which were experienced in the 1D case by Malinverno (2002) and Malinverno and Leaney (2005), but which are totally novel in tomography.

First, the mesh self-adapts to the information contained in the data. Cell boundaries are free to move during the sampling process, and adapt to the structural features of the unknown model. This adaptive character of the parameterization also takes into account the spatial variability of the information provided the data, and is able to spatially adapt the mesh resolution (i.e. the cell sizes). In this way, small scale features can be imaged in well sampled areas without introducing spurious artefacts elsewhere.

Therefore the method is particularly suited to problems with multiple scales, i.e where the spatial sampling of the Earth is highly heterogeneous, or where the unknown model itself has variables scale lengths. Contrary to most adaptive-mesh tomographic schemes, here the mesh is not solely regulated by the density of rays, but instead the Bayesian procedure is able to infer the quantity of retrievable information present in the data in any region of the velocity field, and to provide a parsimonious solution. For example, a large homogeneous region of constant wave

speed, will be imaged with a single large cell, even if the ray coverage is dense there.

Second, explicit regularisation of the model parameters is not required, thus avoiding global damping procedures and the subjective process of finding an optimal regularisation value. In this sense, the approach can be viewed as a self-regularising inversion algorithm. Furthermore, the level of smoothing is spatially variable and is determined by the data.

Third, when posterior expectations (mean and variance) are computed, models with variable geometries overlap providing a smooth solution map (or curve in 1D) that has a resolution better than any single model. This allows the construction of a continuous smooth map giving accurate estimation of the Earth model uncertainty. Moreover, the model parameterisation involves fewer parameters to achieve better resolution than a fixed grid. We call this feature 'super-resolution'. The 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. The discontinuities of individual Voronoi models are smoothed out in the ensemble solution but the discontinuities required by the data are constructively reinforced. In this way, sharp transitions present in the unknown Earth model can be imaged without being smoothed.

Finally, the transdimensional approach naturally adapts the model complexity (the number of cells or layers) in order to fit the data to the adequate level given by the data noise. In this way, the procedure ensures that parsimonious model descriptions are preferred without imposing an additional simplicity (smoothing) requirement. The uncertainty due to the noise in the data is also accounted for precise identification of any non-uniqueness or poor resolution of the model parameters.

But beyond the transdimensional character of the inversion, an original feature is that little needs to be assumed about the data noise, which is often difficult to quantify. Hierarchical models can take into account the lack of knowledge on the level of data uncertainty. The magnitude and correlation of noise are treated as unknowns in the problem, and the algorithm is able to infer the level of information brought the data, and hence to naturally adjust the model complexity.

7.3 Creating and analysing ensembles of solutions

The approach developed in this work can also be viewed as the combination of two separate and independent procedures. The first is the transdimensional Bayesian inference, where the problem is to produce a large number of models with different

parameterisations and variable complexities that describe the posterior probability distribution. This distribution represents the complete solution of the inverse problem, it mathematically contains all the available information and uncertainty (lack of information) about the Earth model, as well as the correlation between parameters. However, the final goal of geophysical inversion is to produce an interpretable solution, and a single Earth model is often useful for practitioners who are not familiar with Bayesian methods. This is a reason why optimisation methods are often preferred by Earth scientists (the true Earth is unique). As shown in Figure 7.1 with the regression problem, the ensemble solution is hardly readable and cannot be directly interpreted.

Hence, the second part of the procedure consists in extracting interpretable information (e.g. a solution velocity model, a regression function or the location of discontinuities) from this ensemble of models, that is to construct a comprehensive view of the Bayesian solution. In Bayesian studies, the standard way of extracting information from the ensemble is to use marginal and conditional distributions on model parameters (Box and Tiao, 1973; Sivia, 1996). A problem with transdimensional models is that the mapping of a same model parameter may differ between dimensions. It is therefore necessary to consider combination of parameters, or even functionals of parameters which retain their interpretations as the sampler moves along the model space. For example, the location of discontinuities (i.e. the cells boundaries) can be obtained from the set of Voronoi nuclei. They are likely to be largely located at interfaces (or change points) present in the Earth model, and hence provide a physical comprehension of the Earth structure.

To obtain a 'solution model' for analysis purpose, the partition models are first projected into the 'physical Earth' domain. In tomography, sampled models are projected into the space domain, where the parameters represent the velocities in each pixel of the velocity field. Then all the models are simply averaged, i.e. a solution map is constructed by taking the mean of the distribution of values at each point across the Earth model. Instead of the mean value, another possibility is to take the mode of the distribution of values at each pixel. Figure 7.2 shows examples of different solution models that can be constructed from the ensemble solution in Figure 7.1.

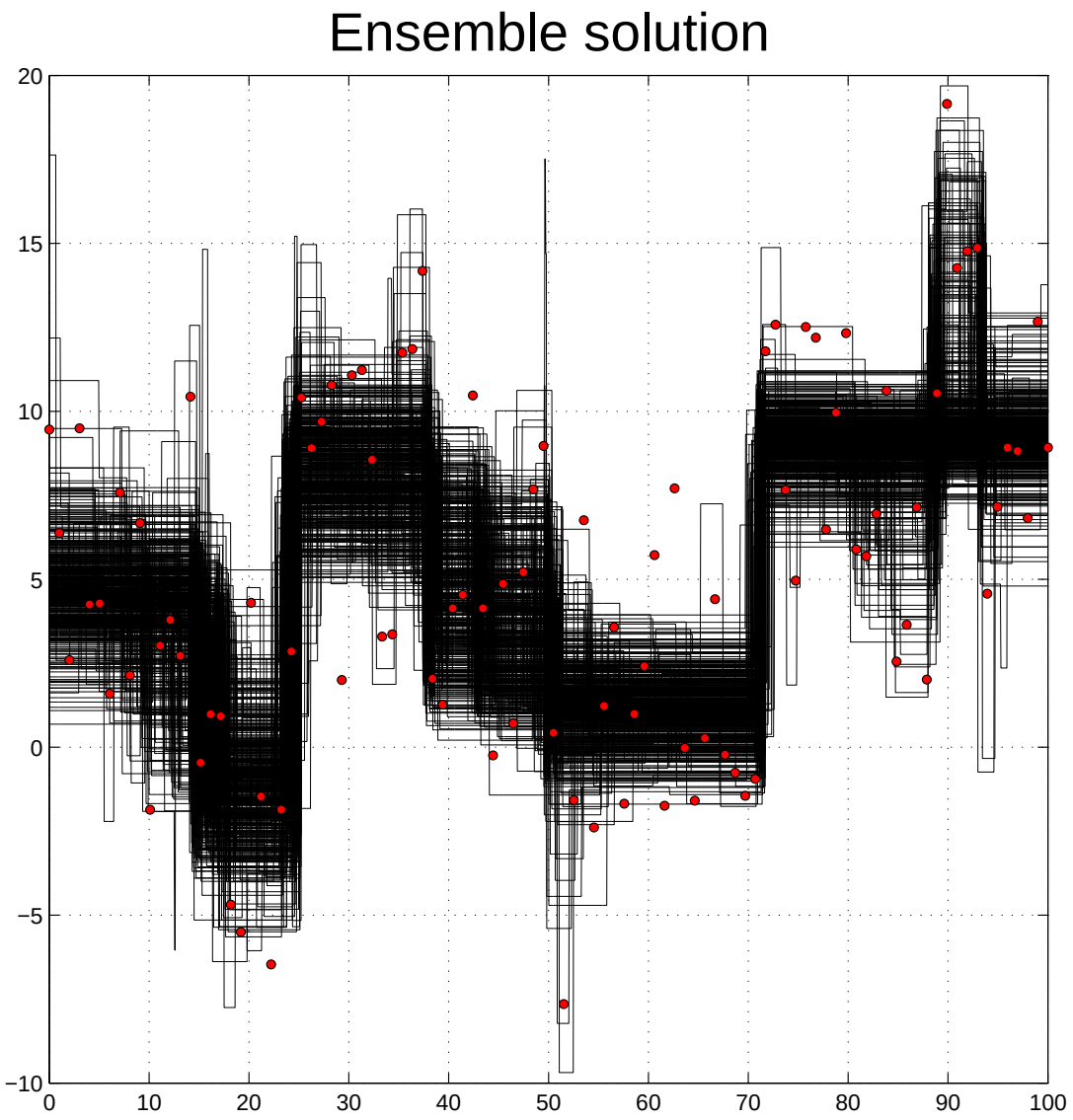


Figure 7.1: Ensemble solution representing the posterior probability distribution. This is the complete solution of the inverse problem but is difficult to interpret.

7.4 Criticisms and limits of the method

There is a relative freedom in the design of solution models. For example, instead of using the whole ensemble of collected models, one can first compute the expected value of hyperparameters (expected number of cells or expected data noise) and construct a solution map by only averaging models that take these hyperarameters values. This is known in the statistical literature as 'Empirical Bayes'. Alternatively, instead using expected values of hyperparameters, one can take as well the mode of

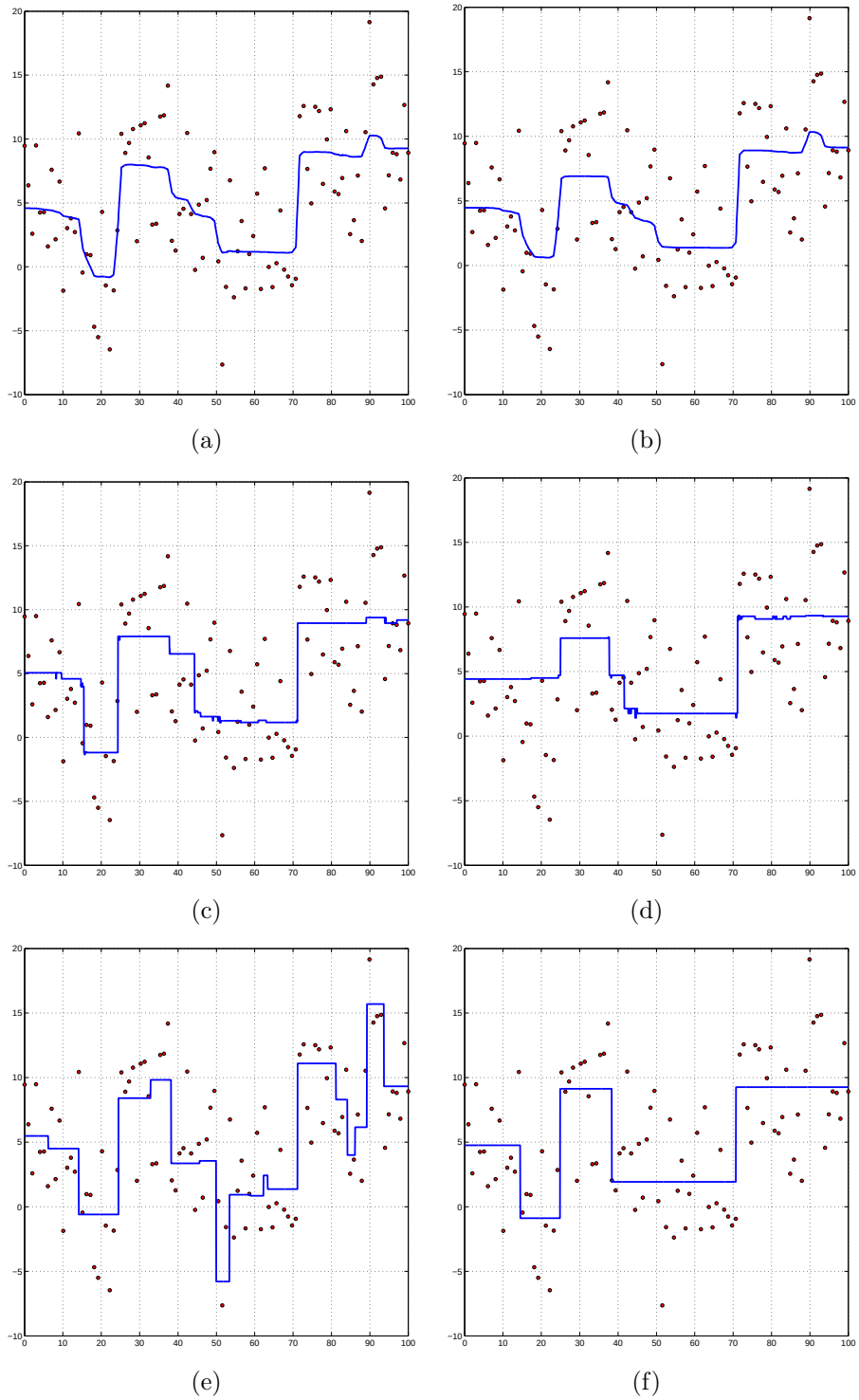


Figure 7.2: Different projections of the ensemble solution. In Left panels solution models are obtained considering the full ensemble solution. Right panels show 'Empirical Bayes' projections, where only partition models with maximum posterior complexity are considered (i.e. only Voronoi models with 5 cells). a) and b) Mean value of the marginal posterior at each depth. c) and d) Maximum value of the marginal posterior at each depth. e) and f) best fitting Voronoi model (posterior maximum).

the marginal distribution on hyperparameters.

There is a large number of ways of producing interpretable models, and the choice is purely arbitrary. Different choices may lead to different Earth models, and hence to different interpretations. Therefore it can be argued that there is an inherent contradiction in our method, as the initial philosophy is to remove any subjective choices made at the outset. Indeed, as we introduce hyperparameters and produce a parsimonious ensemble solution that accounts for all states of uncertainty, the geometry of model space becomes very complex, and then arbitrary projections of the posterior need to be chosen for interpretation. However, it is important to emphasise that in our formulation, the only veritable solution to the inverse problem is the posterior distribution, and any single model derived from the ensemble must be seen by interpreters only as a projection of the posterior solution.

Apart from the above conceptual consideration, the main criticism that can be made to our methodology is the computational price. If the Earth is defined by too many parameters, the number of partition models needed to sample the posterior distribution becomes colossal. And since the predicted data have to be computed each time a model is proposed, our algorithm become computationally prohibitive. In each chapter of the thesis, we have compared our results to solutions obtained with conventional inversion schemes, and showed that our algorithm performed better. However we did not compare computational times, and here it is necessary to recognise that, even parallelised and optimised, our method is between 1 and 3 order of magnitude slower than standard linearised inversions.

7.5 Other potential applications of the ideas in this thesis

The approach presented in this thesis is a general inversion strategy, and it has a wide range of possible applications in geosciences (given that the model space is not too large). From a statistical inversion point of view, the main difference between different geophysical inverse applications lies in the description of the forward problem. Since our method is based on a direct parameter search algorithm, the forward problem is a separate routine independent of the algorithm. Therefore our methodologies can be immediately applied to any geophysical inverse problem, provided that the Earth is parameterised with Voronoi cells, where each cell is associated with a given number of constant geophysical properties.

The method is efficient in automatically picking discontinuities present in the

data, and provides a solution model that can exhibit at the same time low gradients and sharp discontinuities. In this sense it appears to have considerable potential in Earth sciences, given the dual continuous/discrete nature of the Earth.

Geophysical inverse problems that seem appropriate here include resistivity surveying with vertical electric souding or electromagnetic surveys (EM) (Lowrie, 1997), inversion of frequency-domain airborne electromagnetic (AEM) data (e.g. Brodie and Sambridge, 2006, 2009), or seismic cross-hole tomography (Nicollin *et al.*, 2008). Furthermore, in potential fields studies, complex polyhedra are often used to describe anomalous bodies underneath the earth surface (e.g. Luo, 2010), and our Voronoi parameterisation approach could be used for imaging geophysical objects, where the number and shape of objects are unknown.

In this thesis, applications have been proposed in 1D (palaeoclimate regression models and receiver functions) and 2D (seismic tomography). However, all the algorithms presented for the 2D case are known to work in 3D, and hence the approach could be implemented in 3D, at increased computational cost.

We also anticipate that the hierarchical McMC is ideally suited for joint inversions, such as for example DC resistivity and gravity anomaly (e.g. Santos *et al.*, 2006), or surface wave dispersion and receiver functions (e.g. Ozalaybey *et al.*, 1997; Du and Foulger, 1999; Julia *et al.*, 2000; Chang *et al.*, 2004; Lawrence and Wiens, 2004; Yoo *et al.*, 2007; Tkalčić *et al.*, 2006). A general drawback of joint inversions is the definition of the likelihood function, where different likelihood functions resulting from different geophysical data are involved. The Hierarchical Bayes procedure is expected to be a powerful tool in this situation, as it would be able to naturally weight the contribution of differrent data types in the likelihood function, thus removing the arbitrary choice a weighting factor.

7.6 Potential improvements

Finally, we give three examples of potential improvements that could be brought to the algorithm.

7.6.1 An alternative measure of efficiency

A common issue of McMC is the choice of transitions that will optimise the performance of the algorithm, that is sampling efficiency and speed of convergence to stationarity. As seen before, the overall performance depends on the form of transitions, and proposal functions are either manually or automatically 'tuned' in order

to achieve the best performance. Hence the way of quantifying efficiency (how fast the model explores the model space) is an essential component of the algorithm.

Traditionally, efficiency is measured with the rate of acceptance (i.e. the ratio of accepted models to iterations). However, the acceptance rate is a monotonous function (always decreasing) of the variance of proposal functions, and hence sampling efficiency cannot be directly optimised, and the Goldilocks principle must be used (i.e. the acceptance rate must remain within certain margins, as opposed to reaching extremes).

Instead of using the acceptance rate, an alternative way of measuring efficiency would be to directly measure the speed of displacement along the model space, that is the average distance travelled per iteration. A distance between two consecutive models $d(\mathbf{m}_i, \mathbf{m}_{i+1})$ needs to be defined. For transdimensional moves, the distance can be defined as a spatial least square measure. Each time a proposal is rejected, the Markov chain remains at the same place and $d(\mathbf{m}_i, \mathbf{m}_{i+1}) = 0$. Hence, for a given number N of iterations, the 'velocity of exploration' is given by :

$$velocity = \frac{1}{N} \sum_{i=1}^{N} d(\mathbf{m}_i, \mathbf{m}_{i+1}) \tag{7.1}$$

In this way, the velocity increases with the size of perturbations, but if the variance of proposal functions becomes too large, too many models are rejected, which makes the velocity smaller as there are more zeros in the average. With this type of measure, the algorithm efficiency can be directly optimised by seeking the model transitions that maximised the velocity of exploration.

7.6.2 Treating the forward model as an unknown

In our hierarchical model, the data noise and the model parameterisation are variable and unknown in the problem. However, the forward model, i.e. the function g linking the model to the data, is given by the user and remains fixed during the procedure, although in the tomography problem ray geometries are iteratively recomputed. The forward model is the mathematical formulation of our understanding of geophysical processes (e.g. the propagation of seismic waves), it is based on assumptions and contains errors. An idea would be to push the Bayesian formulation further and take into account the uncertainty one has about the forward model. That is, the forward model could be parameterised with hyperparameters which would be given a prior probability density, and which would be perturbed along the Markov chain.

The idea of solving the forward and the inverse problem at the same time is called the 'all-at-once' approach and was first proposed in an optimization context by Haber and Ascher (2001). Subsequently, Haber *et al.* (2004) applied it to inversion of 3D electromagnetic data. In our 2D tomography problem, the ray paths could be characterised by a given (or variable) number of 'ray parameters', and one would obtain for the geometry of each ray a posterior distribution. In the case of finite frequency surface wave tomography (e.g. Yoshizawa, K. and Kennett, B. L. N, 2004), it could be interesting to compare this distribution to the 'banana doughnut' shaped sensitivity kernels.

7.6.3 Treating the data smoothing as an unknown

In the case of inversion of receiver functions, the data vector is a low pass filtered waveform. A Gaussian filter is applied to the time series in the frequency domain to remove high frequency components which have a low signal to noise ratio. This data smoothing is controlled by the width a of the Gaussian filter. It has been shown in chapter 6 that solution models strongly depend on this parameter a. A possibility could be to expand our hierarchical model and consider the data smoothing parameter a as a hyperparameter. If a is too small, the receiver function is smooth, has a low resolution and some infomation about the velocity model is lost during the filtering process. Conversely if a is too large, the signal has not been filtered enough and the receiver function is buried into high frequency noise. Hence the parameter a directly determines the form and quantity of data noise. We have seen that the hierarchical algorithm is able to descriminate between signal and noise and to use the maximum level of retrievable information in the data. Therefore, by letting the parameter a to be an unknown in the problem, we could let the data infer its own level of smoothing.

Appendix A

Hierachical Bayes regression algorithm for multiple data sets

A.1 The prior

Since we have independent parameters of different physical dimension, the prior can be separated into two terms,

$$p(\mathbf{m}, n) = p(\mathbf{m} \mid n)p(n). \tag{A.1}$$

Where $p(n)$ is the prior on the number of partitions. We choose for that a uniform distribution over the interval $I = \{n \; \epsilon \; N \mid n_{min} < n \leq n_{max}\}$. Hence,

$$p(n) = \begin{cases} 1/(\Delta n) & \text{if} \quad n \; \epsilon \; I \\ 0 & \text{otherwise} \end{cases} \tag{A.2}$$

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

Given a number of cells n, the prior probability distributions for the model parameters are independent from each other, and so can be written in separable form

$$p(\mathbf{m} \mid n) = p(\mathbf{c} \mid n)p(\mathbf{V} \mid n)p(\boldsymbol{\sigma} \mid n). \tag{A.3}$$

For the response values $\mathbf{V}$, the prior for each data type j is specified by a constant value over a defined interval $J^j = \{v \; \epsilon \; \Re \mid V^j_{min} < v < V^j_{max}\}$.

Hence we have

$$p(\mathbf{V}_{i\,j} \mid n) = \begin{cases} 1/(\Delta^j v) & \text{if} \quad \mathbf{V}_{i\,j} \; \epsilon \; J^j \\ 0 & \text{otherwise} \end{cases} \tag{A.4}$$

where $\Delta^j v = (V^j_{max} - V^j_{min})$. Since the response values in each cell and for each data type are independent,

$$p(\mathbf{V} \mid n) = \prod_{j=1}^{m} \prod_{i=1}^{n} p(\mathbf{V}_{ij} \mid n) \tag{A.5}$$

Similarly, for $\boldsymbol{\sigma}$, the prior for each data type is specified by a uniform distribution over a defined interval $K_j = \{\sigma \, \epsilon \, \Re \mid \sigma^j_{min} < \sigma < \sigma^j_{max}\}$.

Hence we have

$$p(\sigma_j \mid n) = \begin{cases} 1/(\Delta^j \sigma) & \text{if} \quad \sigma_j \, \epsilon \, K^j \\ 0 & \text{otherwise} \end{cases} \tag{A.6}$$

where $\Delta^j \sigma = (\sigma^j_{max} - \sigma^j_{min})$. Since the estimated noise for each data type is independent,

$$p(\boldsymbol{\sigma} \mid n) = \prod_{j=1}^{m} p(\sigma_j \mid n). \tag{A.7}$$

For mathematical convenience, let us for the moment assume that the Voronoi nuclei can only take place on an underlying grid of finite nodes defined by $N = n_x \times n_y$ possible positions. For n Voronoi nuclei, there are $\left(\frac{N!}{n!(N-n)!}\right)$ possible configurations on the N possible points of the underlying grid. We give equal propability to each of these configurations. Hence,

$$p(\mathbf{c} \mid n) = \left[\frac{N!}{n!(N-n)!}\right]^{-1}. \tag{A.8}$$

Therefore, after substituting (A.5), (A.7), and (A.8) into (A.3), the full prior probability density function can be expressed as

$$p(\mathbf{m}) = \begin{cases} \dfrac{n!(N-n)!}{\Delta n N! \prod_{j=1}^{m} [\Delta^j \sigma (\Delta^j v)^n]} & \text{if } (n \, \epsilon \, I \text{ and } \forall (i \, j) \, \epsilon \, [1,n][1,m], \ \mathbf{V}_{ij} \, \epsilon \, J^j \text{ and } \sigma_j \, \epsilon \, K^j) \\ 0 & \text{otherwise} \end{cases} \tag{A.9}$$

A.2 proposal distributions

At each iteration, one type of move is uniformly randomly selected from the 5 following possibilities (each having equal probabilities) :

1. Change a Value. Randomly select a cell $(i \, j)$ from a uniform distribution over

$[1\,n] \times [1\,m]$ and propose a new value $\mathbf{V}'_{i\,j}$ using

$$q_{v1}(\mathbf{V}'_{i\,j} \mid \mathbf{V}_{i\,j}) = \frac{1}{\theta_1\sqrt{2\pi}} \exp\left\{-\frac{(\mathbf{V}'_{i\,j} - \mathbf{V}_{i\,j})^2}{2\theta_1^2}\right\}. \tag{A.10}$$

Hence we have

$$\mathbf{V}'_{i\,j} = \mathbf{V}_{i\,j} + u \times \theta_1 \tag{A.11}$$

where u is a random deviate from a normal distribution $N(0,1)$ and θ_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. Change the estimated data noise. Randomly select a data set j from a uniform distribution over the range $[1, m]$. Propose a new value σ'_j using

$$q_{\sigma}(\sigma'_j \mid \sigma_j) = \frac{1}{\theta_2\sqrt{2\pi}} \exp\left\{-\frac{(\sigma'_j - \sigma_j)^2}{2\theta_2^2}\right\}. \tag{A.12}$$

3. BIRTH : create a new cell. Add a new Voronoi centre with the position c'_{n+1} found by choosing uniformly randomly a point from the underlying grid that is not already occupied. There are $(N - n)$ discrete points available. Then, m new response values $(\mathbf{V}'_{n+1\,1}, ..., \mathbf{V}'_{n+1\,m})$ need to be created for the new cell. For each data type, the new value is proposed according to a Gaussian probability density $q_{v2}(\mathbf{V}'_{n+1\,j}|\mathbf{d}^j, \mathbf{c}')$ with mean and variance equal to the mean and variance of the data points within the cell.

$$q_{v2}(\mathbf{V}'_{n+1\,j}|\mathbf{d}^j, \mathbf{c}') = \frac{1}{\Theta\sqrt{2\pi}} \exp\left\{-\frac{(\mathbf{V}'_{n+1\,j} - L)^2}{2\Theta^2}\right\}. \tag{A.13}$$

where L and Θ are functions giving the mean and standard deviation of data points of type j within cell i, given the Voronoi tessellation **c'**. If there are no data points in the new cell, the reponse value $\mathbf{V}'_{n+1\,j}$ is drawn according to the prior distribution.

4. DEATH. Remove at random one cell by drawing a number from a uniform distribution over the range $[1, n]$. The response values of the neighbooring cells remain unchanged.

5. MOVE : Randomly pick one cell (from a uniform distribution) and Randomly

change the position of its nucleus according to

$$q_c(c_i' \mid c_i) = \frac{1}{\theta_3\sqrt{2\pi}} \exp\left\{-\frac{(c_i' - c_i)^2}{2\theta_3^2}\right\}. \tag{A.14}$$

A.3 Proposal ratios

For the proposal types that do not involve a change of dimension 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

$$\begin{aligned} q_c(c_i' \mid c_i) &= q_c(c_i \mid c_i') \\ q_\sigma(\sigma_j' \mid \sigma_j) &= q_\sigma(\sigma_j \mid \sigma_j') \\ q_{v1}(\mathbf{V}_{ij}' \mid \mathbf{V}_{ij}) &= q_{v1}(\mathbf{V}_{ij} \mid \mathbf{V}_{ij}') \end{aligned} \tag{A.15}$$

and in all three cases the the proposal ratio equals one.

$$\frac{q(\mathbf{m} \mid \mathbf{m}')}{q(\mathbf{m}' \mid \mathbf{m})} = 1. \tag{A.16}$$

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 c_{n+1}' is generated independently from the new response values $(\mathbf{V}_{n+1\,1}', ..., \mathbf{V}_{n+1\,m}')$ then proposal distributions can be separated and we write

$$\frac{q(\mathbf{m} \mid \mathbf{m}')}{q(\mathbf{m}' \mid \mathbf{m})} = \frac{q(\mathbf{c} \mid \ \mathbf{m}')}{q(\mathbf{c}' \mid \mathbf{m})} . \frac{q(\mathbf{V} \mid \mathbf{m}')}{q(\mathbf{V}' \mid \mathbf{m})}. \tag{A.17}$$

Specifically we have the probability of a birth at position c_{n+1}' which is given by

$$q(\mathbf{c}' \mid \mathbf{m}) = 1/(N-n), \tag{A.18}$$

the probability of generating a set of new velocity values $(\mathbf{V}_{n+1\,1}', ..., \mathbf{V}_{n+1\,m}')$ is given by

$$q(\mathbf{V}' \mid \mathbf{m}) = \prod_{j=1}^{m} q_{v2}(\mathbf{V}_{n+1\,j}' | \mathbf{d}^j, \mathbf{c}') \tag{A.19}$$

the probability of deleting the cell at position c_{n+1}' (reverse step)

$$q(\mathbf{c} \mid \mathbf{m}') = 1/(n+1) \tag{A.20}$$

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

$$q(\mathbf{v} \mid \mathbf{m}') = 1. \tag{A.21}$$

Substituting these expressions in (A.17) we obtain

$$\left(\frac{q(\mathbf{m} \mid \mathbf{m}')}{q(\mathbf{m}' \mid \mathbf{m})}\right)_{birth} = \frac{(N-n)}{(n+1)\prod_{j=1}^{m} q_{v2}(\mathbf{V}'_{n+1\,j}|\mathbf{d}^j, \mathbf{c}')}. \tag{A.22}$$

For the death of a randomly chosen nucleus, we move from n to $(n-1)$ cells. Suppose that nucleus, c_i is removed. In this case, a similar reasoning to the birth case above leads us to a proposal ratio (reverse to forward) of

$$\left(\frac{q(\mathbf{m} \mid \mathbf{m}')}{q(\mathbf{m}' \mid \mathbf{m})}\right)_{death} = \frac{n\prod_{j=1}^{m} q_{v2}(V_{i\,j}|\mathbf{d}^j, \mathbf{c})}{(N-n+1)} \tag{A.23}$$

where $(\mathbf{V}_{i\,1}, ..., \mathbf{V}_{i\,m})$ are the response value of the cell deleted.

A.4 The Jacobian

In our case, the Jacobian only needs to be calculated when there is a jump between two models of different dimensions, i.e. when a birth or death is proposed (Green, 1995). 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 $\mathbf{m}$' can be written as

$$\mathbf{m} = (\mathbf{c}, \mathbf{V}, u_c, \mathbf{u}_v) \longleftrightarrow (\mathbf{c}, \mathbf{V}, c'_{n+1}, V'_{n+1\,1}, ..., V'_{n+1\,m}) = \mathbf{m}'. \tag{A.24}$$

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 $[0, 1, ..., N-n]$. The random numbers $\mathbf{u}_v = (u_v^1, ..., u_v^m)$ are drawn from Gaussian distributions depending on the distribution of data points.

$$V'_{n+1\,j} = L + u_v^j \tag{A.25}$$

where L is the average of data points of type j in the new cell and u_v^j is drawn from Gaussian distributions centred at 0.

Note that the model space is divided into a discrete space (nuclei position) and a continuous space (response values). u_c is a discrete variable used for the transfor-

mation between discrete spaces and $\mathbf{u}_v$ is a vector of continuous variable used for the transformation between continuous spaces. (Denison *et al.*, 2002) showed that the Jacobian term is always unity for discrete transformations. Therefore, the Jacobian term only accounts for the change in variables from

$$(\mathbf{V}, \mathbf{u}_v) \longleftrightarrow (\mathbf{V}, \mathbf{V}'_{n+1\,1}, ..., \mathbf{V}'_{n+1\,m}) = \mathbf{V}'. \tag{A.26}$$

Hence, we have

$$|\mathbf{J}|_{birth} = \left| \frac{\delta(\mathbf{V}')}{\delta(\mathbf{V}, \mathbf{u}_v)} \right| = 1. \tag{A.27}$$

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

A.5 The acceptance probability

The probability of accepting the proposed model is given by

$$\alpha(\mathbf{m}' \mid \mathbf{m}) = \min\left[1, \frac{p(\mathbf{m}')}{p(\mathbf{m})} \times \frac{p(\mathbf{d}_{obs} \mid \mathbf{m}')}{p(\mathbf{d}_{obs} \mid \mathbf{m})} \times \frac{q(\mathbf{m} \mid \mathbf{m}')}{q(\mathbf{m}' \mid \mathbf{m})} \times |\mathbf{J}|\right] \tag{A.28}$$

We now substitute expressions for each proposal ratio into (A.28) to get final expressions for the acceptance probability in each case. For the moves that do not include a change in dimension, we have seen that the proposal ratio becomes unity. Hence for the three cases the acceptance term is simply given by the ratio of the posteriors

$$\alpha(\mathbf{m}' \mid \mathbf{m}) = \min\left[1, \frac{p(\mathbf{m}')}{p(\mathbf{m})} . \frac{p(\mathbf{d}_{obs} \mid \mathbf{m}')}{p(\mathbf{d}_{obs} \mid \mathbf{m})}\right]. \tag{A.29}$$

Since the dimension of the model does not change, according to (A.9), the prior ratio is either null or unity and we have

$$\alpha(\mathbf{m}', \mathbf{m}) = \begin{cases} \min\left[1, \frac{p(\mathbf{d}_{obs}|\mathbf{m}')}{p(\mathbf{d}_{obs}|\mathbf{m})}\right] & \text{if} \quad \forall(i,j) \in [1,n][1,m],\ \mathbf{V}'_{i\,j} \in J^j) \\ 0 & \text{otherwise} \end{cases} \tag{A.30}$$

We now consider now the 5 possible moves described earlier. For changes in response

values and nuclei positions, we have

$$\alpha(\mathbf{m}', \mathbf{m}) = \begin{cases} \min\left[1, \exp\left\{-\frac{\phi(\mathbf{m}')-\phi(\mathbf{m})}{2}\right\}\right] & \text{if} \quad \forall(i,j) \,\epsilon\, [1,n][1,m],\ \mathbf{V}'_{ij} \,\epsilon\, J^j) \\ 0 & \text{otherwise} \end{cases} \tag{A.31}$$

When perturbing the estimated noise parameters $\boldsymbol{\sigma}$, the normalizing constant in the likelihood is changed and

$$\alpha(\mathbf{m}', \mathbf{m}) = \begin{cases} \min\left[1, \prod_{j=1}^{m}\left(\frac{\sigma_j}{\sigma'_j}\right)^{M_j} \exp\left\{-\frac{\phi(\mathbf{m}')-\phi(\mathbf{m})}{2}\right\}\right] & \text{if} \quad \forall j \,\epsilon\, [1,m],\ \sigma'_j \,\epsilon\, K^j) \\ 0 & \text{otherwise} \end{cases} \tag{A.32}$$

Note that $\phi(\mathbf{m}')$ and $\phi(\mathbf{m})$ incorporate σ_j and σ'_jÂ respectively.

For a birth step, according to (A.9), the prior ratio takes the form

$$\left(\frac{p(\mathbf{m}')}{p(\mathbf{m})}\right)_{birth} = \begin{cases} \dfrac{n+1}{(N-n)\prod_{j=1}^{m}[\Delta^j v]} & \text{if } \left((n+1) \,\epsilon\, I \text{ and } \forall j \,\epsilon\, [1,m], \mathbf{V}'_{n+1\,j} \,\epsilon\, J^j\right) \\ 0 & \text{otherwise.} \end{cases} \tag{A.33}$$

After substituting (4.12), (A.22), and (A.33) into (A.28), the acceptance term for the birth step reduces to

$$\alpha(\mathbf{m}', \mathbf{m}) = \begin{cases} \min\left[1, \prod_{j=1}^{m}\left[\frac{1}{q_{v2}(\mathbf{V}'_{n+1\,j}|\mathbf{d}^j, \mathbf{c}')\Delta^j v}\right]. \exp\left\{-\frac{\phi(\mathbf{m}')-\phi(\mathbf{m})}{2}\right\}\right] & \text{if} \quad A \\ 0 & \text{otherwise} \end{cases} \tag{A.34}$$

with

$$A = \left((n+1) \,\epsilon\, I \quad \text{and} \quad \forall j \,\epsilon\, [1,m], \quad \mathbf{V}'_{n+1\,j} \,\epsilon\, J^j\right) \tag{A.35}$$

For the death step, the prior ratio in (A.33) must be inverted. After substituting this with (4.12) and (A.23) into (A.28), and after simplification we get an the acceptance probability

$$\alpha(\mathbf{m}', \mathbf{m}) = \begin{cases} \min\left[1, \prod_{j=1}^{m}\left[q_{v2}(\mathbf{V}_{ij}|\mathbf{d}^j, \mathbf{c})\Delta^j v\right]. \exp\left\{-\frac{(\phi(\mathbf{m}')-\phi(\mathbf{m}))}{2}\right\}\right] & \text{if} \quad (n-1) \,\epsilon\, I \\ 0 & \text{otherwise} \end{cases} \tag{A.36}$$

where i indicates the cell that we remove from the current changepoint model. $\mathbf{c}$.

Bibliography

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

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control 19*(6), 716–723.

Aki, K. (1957). Space and time spectra of stationary stochastic waves, with special reference to microtremors.

Aki, K., A. Christoffersson and E. Husebye (1977). Determination of the three-dimensional seismic structure of the lithosphere. *J. Geophys. Res 82*(2).

Al-Awadhi, F., M. Hurn and C. Jennison (2004). Improving the acceptance rate of reversible jump MCMC proposals. *Statistics and Probability Letters 69*(2), 189–198.

Ammon, C. (1992). A comparison of deconvolution techniques: Report UCID-ID-111667. *Lawrence Livermore National Laboratory, Livermore, California 43*.

Ammon, C., G. Randall and G. Zandt (1990). On the nonuniqueness of receiver function inversions. *Journal of Geophysical Research 95*(B10), 15303.

Arroucau, P., R. N., S. M. and R. A. M. (2009). Rayleigh wave group tomography in southeast Australia and Tasmania from cross-correlation of the ambient noise wavefield recorded with WOMBAT, a rolling array experiment . In *AGU Fall Meeting Abstracts*.

Arroucau, P., N. Rawlinson and M. Sambridge (2010). New insight into Cainozoic sedimentary basins and Palaeozoic suture zones in southeast Australia from ambient noise surface wave tomography. *Geophysical Research Letters 37*(7), L07303.

Aster, R., B. Borchers and C. Thurber (2005). *Parameter Estimation and Inverse Problems*. Academic Press.

Aurenhammer, F. (1991). Voronoi diagrams-a survey of a fundamental geometric data structure. *ACM Computing Surveys (CSUR) 23*(3), 345–405.

Bannister, S., J. Yu, B. Leitner and B. Kennett (2003). Variations in crustal structure across the transition from West to East Antarctica, Southern Victoria Land. *Geophysical Journal International 155*(3), 870–880.

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

Bensen, G., M. Ritzwoller, M. Barmin, A. Levshin, F. Lin, M. Moschetti, N. Shapiro and Y. Yang (2007). Processing seismic ambient noise data to obtain reliable broad-band surface wave dispersion measurements. *Geophysical Journal International 169*(3), 1239–1260.

Betts, P., D. Giles, G. Lister and L. Frick (2002). Evolution of the Australian lithosphere. *Australian Journal of Earth Sciences 49*(4), 661–695.

Bijwaard, H. and W. Spakman (2000). Non-linear global P-wave tomography by iterated linearized inversion. *Geophysical Journal International 141*(1), 71–82.

Bijwaard, H., W. Spakman and E. Engdahl (1998). Closing the gap between regional and global travel time tomography. *Journal of Geophysical Research 103*(B12), 30055.

Bloemendal, J. and P. de Menocal (1989). Evidence for a change in the periodicity of tropical climate cycles at 2. 4 Myr from whole-core magnetic susceptibility measurements. *Nature(London) 342*(6252), 897–900.

Bohm, G., P. Galuppo and A. Vesnaver (2000). 3D adaptive tomography using Delaunay triangles and Voronoi polygons. *Geophysical Prospecting 48*(4), 723–744.

Box, G. and G. Tiao (1973). Bayesian Inference in Statistical Inference.

Bradford, R. and A. Thomas (1996). Markov chain Monte Carlo methods for family trees using a parallel processor. *Statistics and Computing 6*(1), 67–75.

Brenguier, F., N. Shapiro, M. Campillo, V. Ferrazzini, Z. Duputel, O. Coutant and A. Nercessian (2008). Towards forecasting volcanic eruptions using seismic noise. *Nature Geoscience 1*(2), 126.

Briffa, K., P. Jones, F. Schweingruber, S. Shiyatov and E. Cook (1995). Unusual twentieth-century summer warmth in a 1,000-year temperature record from Siberia. *Nature 376*, 156–159.

Brodie, R. and M. Sambridge (2006). A holistic approach to inversion of frequency-domain airborne EM data. *Geophysics 71*, G301.

Brodie, R. and M. Sambridge (2009). Holistic inversion of frequency-domain airborne electromagnetic data with minimal prior information. *Exploration Geophysics 40*(1), 8–16.

Brooks, S. and P. Giudici (1999). Convergence assessment for reversible jump MCMC simulations. *Bayesian Statistics 6*, 733–742.

Brooks, S., P. Giudici and A. Philippe (2003). Nonparametric convergence assessment for MCMC model selection. *JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS 12*(1), 1–22.

Brooks, S., P. Giudici and G. Roberts (2003). Efficient construction of reversible jump Markov chain Monte Carlo proposal distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 65*(1), 3–39.

Brooks, S. and G. Roberts (1999). On Quantile Estimation and MCMC Convergence. *Biometrika 86*, 710–717.

Brooks, S. P. and G. O. Roberts (1998). Diagnosing convergence of markov chain monte carlo algorithms. *Statistics and Computing 8*, 319–335.

Budinger, T., G. Gullberg and R. Huesman (1979). Emission computed tomography. *Image reconstruction from projection: Implementation and Applications*, 147246.

Campillo, M. and A. Paul (2003). Long-Range Correlations in the Diffuse Seismic Coda.

Cerveny, V. and M. Brown (2003). Seismic Ray Theory. *The Journal of the Acoustical Society of America 113*, 14.

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

Chang, S., C. Baag and C. Langston (2004). Joint analysis of teleseismic receiver functions and surface wave dispersion using the genetic algorithm. *Bulletin of the Seismological Society of America 94*(2), 691.

Chappellaz, J., T. Bluniert, D. Raynaud, J. Barnola, J. Schwander and B. Stauffert (1993). Synchronous changes in atmospheric CH 4 and Greenland climate between 40 and 8 kyr BP. *Nature 366*(6454), 443–445.

Chen, Y., F. Niu, R. Liu, Z. Huang, H. Tkalčić, L. Sun and W. Chan (2010). Crustal structure beneath China from receiver function analysis. *Journal of Geophysical Research 115*(B3), B03307.

Chevrot, S. and R. van der Hilst (2000). The Poisson ratio of the Australian crust: geological and geophysical implications. *Earth and Planetary Science Letters 183*(1-2), 121–132.

Chiao, L. and B. Kuo (2001). Multiscale seismic tomography. *Geophysical Journal International 145*(2), 517–527.

Clayton, R. and R. Wiggins (1976). Source shape estimation and deconvolution of teleseismic bodywaves. *Geophysical Journal of the Royal Astronomical Society 47*(1), 151–177.

Clifford, P., S. Greenhalgh, G. Houseman and F. Graeber (2007). 3-D seismic tomography of the Adelaide fold belt. *Geophysical Journal International 172*(1), 167–186.

Clitheroe, G., O. Gudmundsson and B. Kennett (2000a). Sedimentary and upper crustal structure of Australia from receiver functions. *Australian Journal of Earth Sciences 47*(2), 209–216.

Clitheroe, G., O. Gudmundsson and B. Kennett (2000b). The crustal thickness of Australia. *Journal of geophysical research 105*(B 6), 13697–13713.

Cobb, K., C. Charles, H. Cheng and R. Edwards (2003). El Nino/Southern Oscillation and tropical Pacific climate during the last millennium. *Nature 424*(6946), 271–276.

Cole, J., R. Fairbanks and G. Shen (1993). Recent variability in the Southern Oscillation: Isotopic results from a Tarawa Atoll coral. *Science 260*(5115), 1790.

Cotte, N. and G. Laske (2002). Testing group velocity maps for Eurasia. *Geophysical Journal International 150*(3), 639–650.

Cowles, M. and B. Carlin (1996). Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review. *Journal of the American Statistical Association 91*(434).

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

Dansgaard, W., S. Johnsen, H. Clausen, D. Dahl-Jensen, N. Gundestrup, C. Hammer, C. Hvidberg, J. Steffensen, A. Sveinbjorns, J. Jouzel et al. (1993). Evidence for general instability of past climate from a 250-kyr ice-core record. *Nature 364*(6434), 218–220.

Denison, D., N. Adams, C. Holmes and D. Hand (2002). Bayesian partition modelling. *Computational Statistics and Data Analysis 38*(4), 475–485.

Denison, D. and C. Holmes (2001). Bayesian Partitioning for Estimating Disease Risk. *Biometrics 57*(1), 143–149.

Denison, D., C. Holmes, B. Mallik and A. Smith (2002). *Bayesian nonlinear methods for classification and regression*. Chichester: John Wiley.

Derode, A., E. Larose, M. Tanter, J. de Rosny, A. Tourin, M. Campillo and M. Fink (2003). Recovering the Green's function from field-field correlations in an open scattering medium (L). *The Journal of the Acoustical Society of America 113*, 2973.

Dettmer, J., S. Dosso and C. Holland (2010). Trans-dimensional geoacoustic inversion. *The Journal of the Acoustical Society of America 128*, 3393.

Dettmer, J., C. Holland and S. Dosso (2009). Analyzing lateral seabed variability with Bayesian inference of seabed reflection data. *The Journal of the Acoustical Society of America 126*, 56.

Di Bona, M. et al. (1998). Variance estimate in frequency-domain deconvolution for teleseismic receiver function computation. *Geophysical journal international 134*(2), 634–646.

Du, Z. and G. Foulger (1999). The crustal structure beneath the northwest fjords, Iceland, from receiver functions and surface waves. *Geophysical Journal International 139*(2), 419–432.

Duijndam, A. (1988a). Bayesian estimation in seismic inversion. Part I: Principles. *Geophysical Prospecting 36*(8), 878–898.

Duijndam, A. (1988b). Bayesian estimation in seismic inversion. Part II: Uncertainty analysis: Geophys. *Prosp 36*, 899–918.

Esper, J., E. Cook and F. Schweingruber (2002). Low-frequency signals in long tree-ring chronologies for reconstructing past temperature variability.

Fishwick, S., B. Kennett and A. Reading (2005). Contrasts in lithospheric structure within the Australian craton. Insights from surface wave tomography. *Earth and Planetary Science Letters 231*(3-4), 163–176.

Fishwick, S. and A. Reading (2008). Anomalous lithosphere beneath the Proterozoic of western and central Australia: A record of continental collision and intraplate deformation? *Precambrian Research 166*(1-4), 111–121.

Flournay, N. and R. Tsutakawa (1989). Statistical Multiple Integration (Proc. AMS-IMS-SIAM Summer Research Conf. on Statistical Multiple Integration)(Providence, RI: American Mathematical Society).

Frederiksen, A., H. Folsom and G. Zandt (2003). Neighbourhood inversion of teleseismic Ps conversions for anisotropy and layer dip. *Geophysical Journal International 155*(1), 200–212.

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. *Geophysical Journal International 132*(1), 203–225.

Fukao, Y., M. Obayashi, H. Inoue and M. Nembai (1992). Subducting slabs stagnant in the mantle transition zone. *Journal of Geophysical Research 97*(B4), 4809–4822.

Gallagher, K., K. Charvin, S. Nielsen, M. Sambridge and J. Stephenson (2009). Markov chain Monte Carlo (MCMC) sampling methods to determine optimal models, model resolution and model choice for Earth Science problems. *Marine and Petroleum Geology 26*(4), 525–535.

Gallagher, K., M. Ramsdale, L. Lonergan and D. Morrow (1997). The role of thermal conductivity measurements in modelling thermal histories in sedimentary basins. *Marine and Petroleum Geology 14*(2), 201–214.

Gelman, A., J. Carlin, H. Stern and D. Rubin (1995). Bayesian Data Analysis. Texts in Statistical Science. *Chapman & Hall. ISBN 0 412*(03991), 5.

Gelman, A., G. Roberts and W. Gilks (1996). Efficient Metropolis jumping rules. *Bayesian Statistics 5*, 599–607.

Gelman, A. and D. Rubin (1992). Inference from Iterative Simulation Using Multiple Sequences. *STATISTICAL SCIENCE 7*, 457–457.

Gilks, W., S. Richardson and D. Spiegelhalter (1996). *Markov Chain Monte Carlo in Practice*. Chapman & Hall/CRC.

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

Gouveia, W. and J. Scales (1998). Bayesian seismic waveform inversion- Parameter estimation and uncertainty analysis. *Journal of Geophysical Research 103*(B2), 2759–2780.

Graeber, F., G. Houseman and S. Greenhalgh (2002). Regional teleseismic tomography of the western Lachlan Orogen and the Newer Volcanic Province, southeast Australia. *Geophysical Journal International 149*(2), 249–266.

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

Green, P. (2003). Trans-dimensional Markov chain Monte Carlo. *Highly Structured Stochastic Systems 27*, 179–98.

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

Haario, H., M. Laine, A. Mira and E. Saksman (2006). DRAM: Efficient adaptive MCMC. *Statistics and Computing 16*(4), 339–354.

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

Haber, E. and U. Ascher (2001). Preconditioned all-at-once methods for large, sparse parameter estimation problems. *Inverse Problems 17*, 1847.

Haber, E., U. Ascher and D. Oldenburg (2004). Inversion of 3D electromagnetic data in frequency and time domain using an inexact all-at-once approach. *GEOPHYSICS-WISCONSIN THEN TULSA-SOCIETY OF EXPLORATION GEOPHYSICISTS- 69*, 1216–1228.

Han, C. and B. Carlin (2001). Markov Chain Monte Carlo Methods for Computing Bayes Factors: A Comparative Review. *Journal of the American Statistical Association 96*(455), 1122–1132.

Harland, K., R. White and H. Soosalu (2009). Crustal structure beneath the Faroe Islands from teleseismic receiver functions. *Geophysical Journal International 177*(1), 115–124.

Harmon, N., D. Forsyth and S. Webb (2007). Using ambient seismic noise to determine short-period phase velocities and shallow shear velocities in young oceanic lithosphere. *Bulletin of the Seismological Society of America 97*(6), 2009.

Hartzell, S. and C. Langer (1993). Importance of model parameterization in finite fault inversions: Application to the 1974 Mw 8.0 Peru earthquake. *J. geophys. Res 98*, 22–123.

Haskell, N. (1953). The dispersion of surface waves on multilayered media. *Bulletin of the Seismological Society of America 43*(1), 17.

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

Helmberger, D. and R. Wiggins (1971). Upper mantle structure of midwestern United States. *J. geophys. Res 76*(14), 3229–3245.

Herbert, T. and L. Mayer (1991). Long climatic time series from sediment physical property measurements. *Journal of Sedimentary Research 61*(7), 1089–1108.

Hetényi, G. and Z. Bus (2007). Shear wave velocity and crustal thickness in the Pannonian Basin from receiver function inversions at four permanent stations in Hungary. *Journal of Seismology 11*(4), 405–414.

Hopcroft, P., K. Gallagher and C. Pain (2007). Inference of past climate from borehole temperature data using Bayesian Reversible Jump Markov chain Monte Carlo. *Geophysical Journal International*.

Hopcroft, P., K. Gallagher and C. Pain (2009). A Bayesian partition modelling approach to resolve spatial variability in climate records from borehole temperature inversion. *Geophysical Journal International 9999*(9999).

Hoppe, W. and R. Hegerl (1980). Three-dimensional structure determination by electron microscopy (nonperiodic specimens). *Computer Processing of Electron Microscope Images*, 127–185.

Hubans, F., A. Paul, M. Campillo and H. Karabulut, H. Hatzidimitriou (2010). Crustal tomography of the Aegean-Anatolian domain using noise cross-correlations. In *EGU Meeting Abstracts*.

Imbrie, J., A. McIntyre and A. Mix (1989). Oceanic response to orbital forcing in the late Quaternary: observational and experimental strategies. *NATO ASI series. Series C, Mathematical and physical sciences 285*, 121–164.

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

Jones, P. and M. Mann (2004). Climate over past millennia. *Rev. Geophys 42*(2), 1–42.

Joyce, J., L. Tjalsma and J. Prutzman (1990). High resolution planktic stable isotope record and spectral analysis for the last 5.35 MY: ODP Site 625. *Northeast Gulf of Mexico: Paleoceanography 5*, 507–529.

Julia, J., C. Ammon, R. Herrmann and A. Correig (2000). Joint inversion of receiver function and surface wave dispersion observations. *Geophysical Journal International 143*(1), 99–112.

Kennett, B., M. Sambridge and P. Williamson (1988). Subspace methods for large inverse problems with multiple parameter classes. *Geophysical Journal International 94*(2), 237–247.

Kind, R., G. Kosarev and N. Petersen (1995). Receiver functions at the stations of the German Regional Seismic Network (GRSN). *Geophysical Journal International 121*(1), 191–202.

Koren, Z., K. Mosegaard, E. Landa, P. Thore and A. Tarantola (1991). Monte Carlo estimation and resolution analysis of seismic background velocities. *Journal of Geophysical Research 96*(B12), 20289–20299.

Kosarev, G., N. Petersen, L. Vinnik and S. Roecker (1993). Receiver functions for the Tien Shan analog broadband network: contrasts in the evolution of structures across the Talasso-Fergana fault. *Journal of Geophysical Research 98*(B3), 4437–4448.

Kylander, M., J. Muller, R. Wust, K. Gallagher, R. Garcia-Sanchez, B. Coles and D. Weiss (2007). Rare earth element and Pb isotope variations in a 52 kyr peat core from Lynch's Crater (NE Queensland, Australia): Proxy development and application to paleoclimate in the Southern Hemisphere. *Geochimica et Cosmochimica Acta 71*(4), 942–960.

Langston, C. (1979). Structure under Mount Rainier, Washington, inferred from teleseismic body waves. *J. geophys. Res 84*(B9), 4749–4762.

Large, D., B. Spiro, M. Ferrat, M. Shopland, M. Kylander, K. Gallagher, X. Li, C. Shen, G. Possnert, G. Zhang et al. (2009). The influence of climate, hydrology and permafrost on Holocene peat accumulation at 3500m on the eastern Qinghai–Tibetan Plateau. *Quaternary Science Reviews*.

Larose, E., L. Margerin, A. Derode, B. van Tiggelen, M. Campillo, N. Shapiro, A. Paul, L. Stehly and M. Tanter (2006). Correlation of random wavefields: An interdisciplinary review.

Lawrence, J. and D. Wiens (2004). Combined receiver-function and surface wave phase-velocity inversion using a niching genetic algorithm: application to Patagonia. *Bulletin of the Seismological Society of America 94*(3), 977.

Levin, V. and J. Park (1997). P-SH conversions in a flat-layered medium with anisotropy of arbitrary orientation. *Geophysical Journal International 131*(2), 253–266.

Ligorria, J. and C. Ammon (1999). Iterative deconvolution and receiver-function estimation. *Bulletin of the Seismological Society of America 89*(5), 1395.

Lisiecki, L. and M. Raymo (2005). A Plio-Pleistocene stack of 57 globally distributed benthic δ 18 O records. *Paleoceanography 20*, 522–533.

Lobkis, O. and R. Weaver (2001). On the emergence of the Green's function in the correlations of a diffuse field. *The Journal of the Acoustical Society of America 110*, 3011.

Lombardi, D. (2007). Alpine Crustal and Upper-Mantle Structure from Receiver Functions.A thesis submitted for the degree of Doctor of Philosophy of ETH Zurich.

Loris, I., G. Nolet, I. Daubechies and F. Dahlen (2007). Tomographic inversion using 1-norm regularization of wavelet coefficients. *GEOPHYSICAL JOURL INTERNATIONAL 170*(1), 359.

Lowrie, W. (1997). *Fundamentals of Geophysics*. Cambridge University Press.

Lucente, F., N. Piana Agostinetti, M. Moro, G. Selvaggi and M. Di Bona (2005). Possible fault plane in a seismic gap area of the southern Apennines (Italy) revealed by receiver function analysis. *Journal of Geophysical Research 110*(B4), B04307.

Luo, X. (2010). Constraining the shape of a gravity anomalous body using reversible jump Markov chain Monte Carlo. *Geophysical Journal International 180*(3), 1067–1079.

MacKay, D. (2003). Information theory, inference, and learning algorithms.

Mahalanobis, P. (1936). On the generalised distance in statistics. Proceedings National Institute of Science. India.

Maiti, S. and R. Tiwari (2010). Neural network modeling and an uncertainty analysis in Bayesian framework: A case study from the KTB borehole site. *Journal of Geophysical Research 115*(B10), B10208.

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

Malinverno, A. and V. Briggs (2004). Expanded uncertainty quantification in inverse problems: Hierarchical Bayes and empirical Bayes. *Geophysics 69*, 1005.

Malinverno, A. and W. Leaney (2000). A Monte Carlo method to quantify uncertainty in the inversion of zero-offset VSP data: SEG 70th Annual Meeting,

Calgary, Alberta, The Society of Exploration Geophysicists. In *Expanded Abstracts*.

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

Malinverno, A. and R. Parker (2006). Two ways to quantify uncertainty in geophysical inverse problems. *Geophysics 71*, W15.

Mann, M., R. Bradley and M. Hughes (1998). Global-scale temperature patterns and climate forcing over the past six centuries. *Nature 392*(6678), 779–787.

Mayewski, P., G. Holdsworth, M. Spencer, S. Whitlow, M. Twickler, M. Morrison, K. Ferland and L. Meeker (1993). Ice-core sulfate from three Northern Hemisphere sites: Source and temperature forcing implications. *Atmospheric environment. Part A, General topics 27*(17-18), 2915–2919.

Meier, U., A. Curtis and J. Trampert (2007). Global crustal thickness from neural network inversion of surface wave data. *Geophysical Journal International 169*(2), 706–722.

Meier, U., J. Trampert and A. Curtis (2009). Global variations of temperature and water content in the mantle transition zone from higher mode surface waves. *Earth and Planetary Science Letters 282*(1-4), 91–101.

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

Metropolis, N. et al. (1953). Equations of state calculations by fast computational machine. *Journal of Chemical Physics 21*(6), 1087–1091.

Michelini, A. (1995). An adaptive-grid formalism for traveltime tomography. *Geophysical Journal International 121*(2), 489–510.

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. and A. Tarantola (1995). Monte Carlo sampling of solutions to inverse problems. *J. Geophys. Res 100*(B7), 12–431.

Mudelsee, M. (2000). Ramp function regression: a tool for quantifying climate transitions. *Comput. Geosci 26*, 293–307.

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

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

Nicholson, T., M. Bostock and J. Cassidy (2005). New constraints on subduction zone structure in northern Cascadia. *Geophysical Journal International 161*(3), 849–859.

Nicollin, F., D. Gibert, P. Bossart, C. Nussbaum and C. Guervilly (2008). Seismic tomography of the Excavation Damaged Zone of the Gallery 04 in the Mont Terri Rock Laboratory. *Geophysical Journal International 172*(1), 226–239.

Nolet, G. (2008). *A breviary of seismic tomography: imaging the interior of the earth and sun*. Cambridge University Press.

Nolet, G. and R. Montelli (2005). Optimal parametrization of tomographic models. *Geophysical Journal International 161*(2), 365–372.

Nolet, G. and G. Panza (1976). Array analysis of seismic surface waves: Limits and possibilities. *Pure and Applied Geophysics 114*(5), 775–790.

Nolte, B. and L. Fraser (1994). Vertical seismic profiles invasion with genetic algorithms. *Geophs J Int 117*, 162–179.

Okabe, A., B. Boots and K. Sugihara (1992). *Spatial tessellations: concepts and applications of Voronoi diagrams*. John Wiley & Sons, Inc. New York, NY, USA.

Owens, T., G. Zandt and S. Taylor (1984). Seismic evidence for an ancient rift beneath the Cumberland Plateau, Tennessee: A detailed analysis of broadband teleseismic P waveforms. *Journal of Geophysical Research 89*(B9), 7783–7795.

Ozalaybey, S., M. Savage, A. Sheehan, J. Louie and J. Brune (1997). Shear-wave velocity structure in the northern Basin and Range province from the combined analysis of receiver functions and surface waves. *Bulletin of the Seismological Society of America 87*(1), 183.

Petit, J., J. Jouzel, D. Raynaud, N. Barkov, J. Barnola, I. Basile, M. Bender, J. Chappellaz, M. Davis, G. Delaygue et al. (1999). Climate and atmospheric history of the past 420,000 years from the Vostok ice core, Antarctica. *Nature 399*(6735), 429–436.

Phinney, R. (1964). Structure of the Earth's crust from spectral behavior of long-period body waves. *Journal of Geophysical Research 69*, 2997–3017.

Piana Agostinetti, N. and C. Chiarabba (2008). Seismic structure beneath Mt Vesuvius from receiver function analysis and local earthquakes tomography: evidences for location and geometry of the magma chamber. *Geophysical Journal International 175*(3), 1298–1308.

Piana Agostinetti, N., F. Lucente, G. Selvaggi and M. Di Bona (2002). Crustal structure and Moho geometry beneath the Northern Apennines (Italy). *Geophysical Research Letters 29*(20), 1999.

Piana Agostinetti, N. and A. Malinverno (2010). Receiver function inversion by trans-dimensional Monte Carlo sampling. *Geophysical Journal International 181*(2), 858–872.

Press, W., S. Teukolsky, W. Vetterling and B. Flannery (1992). *Numerical recipes in FORTRAN: the art of scientific computing*. Cambridge University Press New York, NY, USA.

Prindle, K. and T. Tanimoto (2006). Teleseismic surface wave study for S-wave velocity structure under an array: Southern California. *Geophysical Journal International 166*(2), 601–621.

Qiu, X., K. Priestley and D. Mckenzie (2002). Teleseismic P-waveform receiver function analysis and its application to Qiongzhong station (QIZ) of Hainan Island, NW South China Sea. *Journal of Geosciences of China 4*(1), 1–8.

Quinn, T., T. Crowley, F. Taylor, C. Henin, P. Joannot and Y. Join (1998). A multicentury stable isotope record from a New Caledonia coral: Interannual and decadal sea surface temperature variability in the southwest Pacific since 1657 AD. *PALEOCEANOGRAPHY 13*(4), 412–426.

Raggi, D. (2005). Adaptive MCMC methods for inference on affine stochastic volatility models with jumps. *Econometrics Journal 8*(2), 235–250.

Rawlinson, N. and B. Kennett (2008). Teleseismic tomography of the upper mantle beneath the southern Lachlan Orogen, Australia. *Physics of the Earth and Planetary Interiors 167*(1-2), 84–97.

Rawlinson, N., S. Pozgay and S. Fishwick (2009). Seismic tomography: A window into deep Earth. *Physics of the Earth and Planetary Interiors*.

Rawlinson, N., A. Reading and B. Kennett (2006). Lithospheric structure of Tasmania from a novel form of teleseismic tomography. *Journal of Geophysical Research-Solid Earth 111*(B2), B02301.

Rawlinson, N. and M. Sambridge (2003). Seismic traveltime tomography of the crust and lithosphere. *Advances in Geophysics*.

Rawlinson, N. and M. Sambridge (2004). Wave front evolution in strongly heterogeneous layered media using the fast marching method. *Geophysical Journal International 156*(3), 631–647.

Rawlinson, N., M. Sambridge and E. Saygin (2008). A dynamic objective function technique for generating multiple solution models in seismic tomography. *Geophysical Journal International* (0).

Rawlinson, N., H. Tkalcic and B. Kennett (2008). New Results from WOMBAT: an Ongoing Program of Passive Seismic Array Deployment in Australia. In *AGU Fall Meeting Abstracts*, pp. 03.

Rawlinson, N. and M. Urvoy (2006). Simultaneous inversion of active and passive source datasets for 3-D seismic structure with application to Tasmania. *Geophys. Res. Lett 33*.

Reading, A., B. Kennett and M. Dentith (2003). Seismic structure of the Yilgarn Craton, Western Australia. *Australian Journal of Earth Sciences 50*(3), 427–438.

Robert, C. (1995). Convergence Control Methods for Markov Chain Monte Carlo Algorithms. *STATISTICAL SCIENCE 10*, 231–253.

Rosenthal, J. (2000). Parallel computing and Monte Carlo algorithms. *Far East Journal of Theoretical Statistics 4*(2), 207–236.

Ruggieri, E., T. Herbert, K. Lawrence and C. Lawrence (2009). Change point method for detecting regime shifts in paleoclimatic time series: Application to d18O time series of the Plio-Pleistocene. *Paleoceanography 24*.

Sabra, K., P. Gerstoft, P. Roux, W. Kuperman and M. Fehler (2005). Surface wave tomography from microseisms in Southern California. *Geophys. Res. Lett 32*, L14311.

Salmon, M. and P. Arroucau (2010). New Results from the South Australian Seismic Arrays. In *AESC Abstracts*.

Sambridge, M. (1999a). Geophysical inversion with a neighbourhood algorithm I. Searching a parameter space. *Geophysical Journal International 138*(2), 479–494.

Sambridge, M. (1999b). Geophysical inversion with a neighbourhood algorithmII. Appraising the ensemble. *Geophysical Journal International 138*(3), 727–746.

Sambridge, M., J. Braun and H. Mcqueen (1995). Geophysical parametrization and interpolation of irregular data using natural neighbours. *Geophysical journal international(Print) 122*(3), 837–857.

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

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

Sambridge, M., K. Gallagher, A. Jackson and P. Rickwood (2006). Trans-dimensional inverse problems, model comparison and the evidence. *Geophysical Journal International 167*(2), 528–542.

Sambridge, M. and O. Gudmundsson (1998). Tomographic systems of equations with irregular cells. *Journal of Geophysical Research 103*(B1), 773–782.

Sambridge, M. and N. Rawlinson (2005). Seismic tomography with irregular meshes. *Geophysical monograph 157*, 49–65.

Santos, F., S. Sultan, P. Represas and A. El Sorady (2006). Joint inversion of gravity and geoelectrical data for groundwater and structural investigation: application to the northwestern part of Sinai, Egypt. *Geophysical Journal International 165*(3), 705–718.

Saygin, E.and Kennett, B. and S. Pozgay (2010). Australian Crust and Upper-Mantle Structure from Ambient Noise Tomography. In *AESC Abstracts*.

Saygin, E. (2007). Seismic Receiver and Noise Correlation Based Studies in Australia. A thesis submitted for the degree of Doctor of Philosophy of The Australian National University.

Saygin, E. and B. Kennett (2008). Ambient seismic noise tomography of Australian continent. *Tectonophysics*.

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

Scales, J. and R. Snieder (1998). What is noise. *Geophysics 63*(4), 1122–1124.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics 6*(2), 461–464.

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

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

Shapiro, N., M. Campillo, L. Stehly and M. Ritzwoller (2005). High-resolution surface-wave tomography from ambient seismic noise. *Science 307*(5715), 1615.

Shibutani, T., M. Sambridge and B. Kennett (1996). Genetic algorithm inversion-forreceiverfunctionswithapplicationtothecrustanduppermost mantle structure beneath eastern Australia. *Geophys. Res. Lett 23*, 1829–1832.

Sisson, S. (2005). Transdimensional Markov Chains: A Decade of Progress and Future Perspectives. *Journal of the American Statistical Association 100*(471), 1077–1090.

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

Smith, A. (1991). Bayesian Computational Methods. *Philosophical Transactions: Physical Sciences and Engineering 337*(1647), 369–386.

Smith, A. and G. Roberts (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society B 55*(1), 3–23.

Spakman, W. and H. Bijwaard (1998). Irregular Cell Parameterization of Tomographic Problems. *Ann. Geophys 16*, 18.

Spakman, W. and H. Bijwaard (2001). Optimization of cell parameterizations for tomographic inverse problems. *Pure and Applied Geophysics 158*(8), 1401–1423.

Steck, L., C. Thurber, M. Fehler, W. Lutter, P. Roberts, W. Baldridge, D. Stafford and R. Sessions (1998). Crust and upper mantle P wave velocity structure beneath Valles caldera, New Mexico: Results from the Jemez teleseismic tomography experiment. *Journal of Geophysical Research-Solid Earth 103*(B10).

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

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

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

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

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

Thomson, W. (1950). Transmission of elastic waves through a stratified solid medium. *Journal of Applied Physics 21*, 89.

Tierney, L. (1994). Markov Chains for Exploring Posterior Distributions. *The Annals of Statistics 22*(4), 1701–1728.

Tierney, L. and A. Mira (1999). Some adaptive Monte Carlo methods for Bayesian inference. *Statistics in Medicine 18*(1718), 2507–2515.

Tikhotsky, S. and U. Achauer (2008). Inversion of controlled-source seismic tomography and gravity data with the self-adaptive wavelet parametrization of velocities and interfaces. *Geophysical Journal International 172*(2), 619–630.

Tkalčić, H., Y. Chen, R. Liu, Z. Huang, L. Sun and W. Chan (2010). Multi-Step modelling of teleseismic receiver functions combined with constraints from seismic tomography: Crustal structure beneath southeast China. *Submitted to Geophysical Journal International*.

Tkalčić, H., M. Pasyanos, A. Rodgers, R. Gok, W. Walter and A. Al-Amri (2006). A multistep approach for joint modeling of surface wave dispersion and teleseismic receiver functions: Implications for lithospheric structure of the Arabian Peninsula. *Journal of Geophysical Research 111*(B11), B11311.

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

Tomé, A. and P. Miranda (2004). Piecewise linear fitting and trend changing points of climate parameters. *Geophysical Research Letters 31*(2), L02207.

van Leeuwen, P. (2009). Particle filtering in geophysical systems. *Monthly Weather Review 137*, 4089–4114.

Vasco, D. (1995). A transformational approach to geophysical inverse problems. *Geophys. J. Int 123*, 183–212.

Vergne, J., G. Wittlinger, Q. Hui, P. Tapponnier, G. Poupinet, J. Mei, G. Herquel and A. Paul (2002). Seismic evidence for stepwise thickening of the crust across the NE Tibetan plateau. *Earth and Planetary Science Letters 203*(1), 25–33.

Vesnaver, B., R. Madrussani et al. (2000). Depth imaging and velocity calibration by 3D adaptive tomography*. *First Break 18*(7), 303–312.

Villasenor, A., Y. Yang, M. Ritzwoller and J. Gallart (2007). Ambient noise surface wave tomography of the Iberian Peninsula: Implications for shallow seismic structure. *Geophys. Res. Lett 34*, L11304.

Vinnik, L., I. Aleshin, M. Kaban, S. Kiselev, G. Kosarev, S. Oreshin and C. Reigber (2006). Crust and mantle of the Tien Shan from data of the receiver function tomography. *Izvestiya Physics of the Solid Earth 42*(8), 639–651.

Vinnik, L., C. Reigber, I. Aleshin, G. Kosarev, M. Kaban, S. Oreshin and S. Roecker (2004). Receiver function tomography of the central Tien Shan. *Earth and Planetary Science Letters 225*(1-2), 131–146.

Virieux, J. and V. Farra (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.

Weaver, R., B. Froment and M. Campillo (2009). Correlation of nonisotropically distributed ballistic scalar diffuse waves in two dimensions. *The Journal of the Acoustical Society of America 125*, 2536.

Yang, Y., M. Ritzwoller, A. Levshin and N. Shapiro (2006). Ambient noise Rayleigh wave tomography across Europe. *Geophysical Journal International 168*(1), 259.

Yao, H. and R. Van der Hilst (2009). Analysis of ambient noise energy distribution and pha se velocity bias in ambient noise tomography, with application to SE Tibet. *Geophys. J. Int., doi 10*.

Yao, H., R. van der Hilst and M. de Hoop (2006). Surface-wave array tomography in SE Tibet from ambient seismic noise and two-station analysis-I. Phase velocity maps. *Geophysical Journal International 166*(2), 732–744.

Yoo, H., R. Herrmann, K. Cho and K. Lee (2007). Imaging the three-dimensional crust of the Korean Peninsula by joint inversion of surface-wave dispersion and teleseismic receiver functions. *Bulletin of the Seismological Society of America 97*(3), 1002.

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

Young, M., N. Rawlinson, P. Arroucau and H. Tkalcic (2010). Ambiant noise tomography of Tasmania. In *Seismics conference, Cairns*.

Yuan, X., S. Sobolev and R. Kind (2002). Moho topography in the central Andes and its geodynamic implications. *Earth and Planetary Science Letters 199*(3-4), 389–402.

Zhang, H. and C. Thurber (2005). Adaptive mesh seismic tomography based on tetrahedral and Voronoi diagrams: Application to Parkfield, California. *J. Geophys. Res 110*.

Zhao, L., M. Sen, P. Stoffa and C. Frohlich (1996). Application of very fast simulated annealing to the determination of the crustal structure beneath Tibet. *Geophysical Journal International 125*(2), 355–370.

Zhu, L. and H. Kanamori (2000). Moho depth variation in southern California from teleseismic receiver functions. *Journal of Geophysical Research 105*, 2969–2980.

Zielhuis, A. and R. Hilst (2007). Upper-mantle shear velocity beneath eastern Australia from inversion of waveforms from SKIPPY portable arrays. *Geophysical Journal International 127*(1), 1–16.