Technische Universität München
Zentrum Mathematik
Lehrstuhl für Mathematische Statistik

Efficient Parameter Estimation in the High-Dimensional Inverse Problem of Seismic Tomography

Ran Zhang

Vollständiger Abdruck der von der Fakultät für Mathematik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender:		Univ.-Prof. Fabian Theis
Prüfer der Dissertation:	1.	Univ.-Prof. Claudia Czado, Ph.D.
	2.	Karin Sigloch, Ph.D., habil. Ludwig-Maximilians-Universität München
	3.	Prof. Finn Lindgren, Ph.D. University of Bath, Vereinigtes Königreich

Die Dissertation wurde am 1. Oktober 2013 bei der Technischen Universität München eingereicht und von der Fakultät für Mathematik am 10. Januar 2014 angenommen.

Hiermit erkläre ich, dass ich die Doktorarbeit selbstständig angefertigt und nur die angegebenen Quellen verwendet habe.

München, den 1. Oktober 2013

Abstract

The focus of this dissertation is on efficient parameter estimation and uncertainty quantification in high dimensional seismic tomography within the Bayesian spatial modeling framework. Seismic tomography is an imaging technique in geophysics used to infer the three-dimensional seismic velocity structure of the earth's interior by assimilating data measured at the surface. The research within this dissertation consists of two pillars:

(1) We present a Bayesian hierarchical model to estimate the joint distribution of earth structural and earthquake source correction parameters. We construct an ellipsoidal spatial prior which allows to accommodate the layered nature of the earth's mantle. With our efficient Markov chain Monte Carlo algorithm (MCMC) we sample from the posterior distribution for large-scale linear inverse problems and provide precise uncertainty quantification in terms of parameter distributions and credible intervals given the data.

(2) We develop and implement a spatial dependency model of the earth's three-dimensional velocity structure based on a Gaussian Matérn field approximation using the theory of stochastic partial differential equations (Lindgren et al., 2011). We carry out the uncertainty quantification of the high dimensional parameter space using the integrated nested Laplace approximation (INLA) (Rue et al., 2009).

Both modeling approaches are applied to a full-fledged tomography problem. In particular the inversion for the upper mantle structure under western North America is facilitated. It involves more than 11,000 seismic velocity and source correction parameters using seismological data from the continental-scale USArray experiment. Our results based on the MCMC algorithm reveal major structures of the mantle beneath the western USA with novel uncertainty assessments. We compare both approaches and demonstrate that the INLA algorithm substantially improves previous work based on regular MCMC sampling. The outcome based on the INLA approach confirms the previous results while simultaneously capturing the spatial dependencies caused by the earthquake sources and the

receiver stations. The statistical misfit is reduced by about 40% and the computing time shows a speedup of about 1.5 to 2 times.

Zusammenfassung

Das Thema dieser Dissertation ist die effiziente Parameterschätzung und Unsicherheitsquantifizierung in hochdimensionaler seismischer Tomographie mit Hilfe von Bayesianischen räumlichen Modellierungsmethoden. Seismische Tomographie ist ein Verfahren in der Geophysik, um die drei-dimensionale Geschwindigkeitsstruktur der seismischen Wellenausbreitung im Erdinneren mit Hilfe der an der Oberfläche aufgenommenen Daten zu bestimmen. Die Vorgehensweise und Hauptforschungssergebnisse dieser Dissertation basieren im Wesentlichen auf den folgenden zwei Säulen:

(1) Die Verteilungen der seismischen Geschwindigkeitsstruktur und der Parameter der Erdbebenquellen werden mit Hilfe eines Bayesianischen hierarchischen Modells geschätzt. Wir konstruieren dazu eine ellipsoidische räumliche Priori-Verteilung, die die geschichtete Erdmantelform beschreibt. Mit unserem effizienten Markov Chain Monte Carlo Algorithmus (MCMC) können wir von der Posteriori-Verteilung für das lineare Inverse Problem Stichproben ziehen. Dies erlaubt eine präzise Quantifizierung der Unsicherheit der Parameterschätzung in dem man die Bayesianischen Konfidenzintervalle angibt.

(2) Ein räumliches Abhängigkeitsmodell wird für die drei-dimensionale Struktur der Wellengeschwindigkeiten mit Hilfe einer Approximation des Gaußschen Matérn-Zufallsfeldes entwickelt. Diese Approximation basiert auf der Theorie der stochastischen partiellen Differentialgleichungen (siehe auch Lindgren et al., 2011). Wir führen die Unsicherheitsquantifizierung des hochdimensionalen Parameterraums mit Hilfe der Methode der integrierten geschichteten Laplace Approximation (INLA) (Rue et al., 2009) aus.

Wir wenden beide Modellierungsansätze auf das hochdimensionale Tomographieproblem der Inversion der obereren Mantelstruktur unter dem Westen der USA an. Diese beinhaltet mehr als 11.000 seismische Geschwindigkeits- und Quellenkorrekturparameter und circa

53,000 seismologische Daten aus dem kontinentalen USArray-Experiment. Unsere Ergebnisse aus dem MCMC-Verfahren offenbaren wichtige Strukturen des Erdmantels unter dem Westen der USA mit Unsicherheitseinschätzungen. Ein Vergleich der beiden Ansätze zeigt, dass der INLA-Algorithmus die früheren Ergebnisse, die mit der MCMC-Methode gewonnen wurden, erheblich verbessert. Das INLA-Verfahren liefert die gleichen Ergebnisse wie die aus dem MCMC-Ansatz, gleichzeitig erfasst es räumliche Abhängigkeiten jeweils zwischen den Erdbebenquellen und den Messstationen. Weiterhin verringert sich der statistische Misfit um circa 40% und eine Rechenzeit wird um den Faktor 1,5 bis 2 verbessert.

Acknowledgments

First, I would like to address my sincere thanks to Prof. Dr. Claudia Czado for offering me the opportunity to do this interdisciplinary PhD research. I am very grateful for her supervision and fruitful advice throughout my study. Further, I would like to express my deepest gratitude to Dr. habil. Karin Sigloch for her tremendous support and guidance in the field of seismic tomography. Without them the research work wouldn't be possible.

My PhD subject is a part of the Project B3 "Efficient Inversion Methods for Parameter Identification in the Earth Sciences" of the Munich Center of Advanced Computing (MAC), Technische Universität München. I gratefully acknowledge the organization of the MAC research consortium for the financial support. I would like to acknowledge the Department of Earth and Environmental Sciences at the Ludwig-Maximilians-Universität München and the Leibniz Supercomputing Center for providing their computational facilities.

Finally, I wish to thank my parents and friends for their tremendous support and encouragement throughout my study.

Contents

Chapter 1

Introduction

This dissertation is based on two papers dealing with parameter estimation and uncertainty quantification in high-dimensional inverse problems of seismic tomography using a Bayesian framework (Zhang et al., 2013a,b). In the first article we develop and implement a Bayesian hierarchical linear model using an efficient MCMC algorithm and apply it to the inversion for the upper mantle structure under western North America. This involves more than 11,000 seismic velocity and earthquake source parameters and 53,000 data observations. The second paper extends the Bayesian model of the first paper by incorporating spatial dependency of the receiver and source data. It requires a spatial modeling technique suitable on a three-dimensional space.

Research context: seismic tomography

Seismic tomography is a geophysical imaging method that allows to estimate the three-dimensional structure of the earth's deep interior, using observations of seismic waves made at its surface. Seismic waves generated by moderate or large earthquakes travel through the entire planet, from crust to core, and can be recorded by seismometers anywhere on earth. They are by far the most highly resolving wave type available for exploring the interior at depths to which direct measurement methods will never penetrate (tens to thousands of kilometers). Seismic tomography takes the shape of a large, linear(ized) inverse problem, typically featuring thousands to millions of measurements and similar numbers of parameters to solve for.

To first order, the earth's interior is layered under the overwhelming influence of gravity. Its resulting, spherically symmetric structure had been robustly estimated by the

1980's (Dziewonski and Anderson, 1981; Kennett and Engdahl, 1991), and is characterized by $O(10^2)$ parameters. Since then, seismologists have been mainly concerned with estimating lateral deviations from this spherically symmetric reference model (Nolet, 2008). Though composed of solid rock, the earth's mantle is in constant motion (the mantle extends from roughly 30 km to 2900 km depth and is underlain by the fluid iron core). Rock masses are rising and sinking at velocities of a few centimeters per year, the manifestation of advective heat transfer: the hot interior slowly loses its heat into space. This creates slight lateral variations in material properties, on the order of a few percent, relative to the statically layered reference model. The goal of seismic tomography is to map these three-dimensional variations, which embody the dynamic nature of the planet's interior.

Beneath well-instrumented regions – such as our chosen example, the United States – seismic waves are capable of resolving mantle heterogeneity on scales of a few tens to a few hundreds of kilometers. Parameterizing the three-dimensional earth, or even just a small part of it, into blocks of that size results in the mentioned large number of unknowns, which mandate a linearization of the inverse problem. Fortunately this is workable, thanks to the rather weak lateral material deviations of only a few percent (larger differences cannot arise in the very mobile mantle).

Seismic tomography is almost always treated as an optimization problem. Most often a least squares approach is followed inverting large, sparse and underconstrained matrices used the method of least squares and Tikhonov regularization (Nolet, 1987; Tian et al., 2009; Sigloch, 2011) while adjoint techniques are used when an explicit matrix formulation is computationally too expensive (Tromp et al., 2005; Sieminski et al., 2007; Fichtner et al., 2009).

Quantifying uncertainties in underdetermined, large inverse problems is important, since a single solution is not sufficient for making conclusive judgements. Our research focuses on two types of Bayesian methods for this problem.

Study approaches:

1. Markov chain Monte Carlo method with a spatial conditional autoregressive regressive prior

For exploring high-dimensional parameter space of the linear(ized) problem in seismic tomography we first apply the Markov chain Monte Carlo (MCMC) methods. MCMC meth-

ods in seismic tomography have been given considerable attention by the geophysical (seismological) community, these applications have been restricted to linear or nonlinear problems of much lower dimensionality assuming Gaussian errors (Mosegaard and Tarantola, 1995, 2002; Sambridge and Mosegaard, 2002). For example, Dębski (2010) compares the damped least-squares method (LSQR), a genetic algorithm and the Metropolis-Hastings (MH) algorithm in a low-dimensional linear tomography problem involving copper mining data. He finds that the MCMC sampling technique provides more robust estimates of velocity parameters compared to the other approaches. Bodin and Sambridge (2009) capture the uncertainty of the velocity parameters in a linear model by selecting the representation grid of the corresponding field, using a reversible jump MCMC (RJMCMC) approach. In Bodin et al. (2012) again RJMCMC algorithms are developed to solve certain transdimensional nonlinear tomography problems with Gaussian errors, assuming unknown variances. Khan et al. (2011) and Mosca et al. (2012) study seismic and thermo-chemical structures of the lower mantle and solve a corresponding low-dimensional nonlinear problem using a standard MCMC algorithm.

We approach linearized tomographic problems (physical forward model inexpensive to solve) in a Bayesian framework, for a fully dimensioned, continental-scale study that features ≈53,000 data points and ≈11,000 parameters. To our knowledge, this is by far the highest dimensional application of Monte Carlo sampling to a seismic tomographic problem. Assuming Gaussian distributions for the error and the prior, our MCMC sampling scheme allows for characterization of the posterior distribution of the parameters by incorporating flexible spatial priors using Gaussian Markov random field (GMRF). Spatial priors using GMRF arise in spatial statistics (Pettitt et al., 2002; Congdon, 2003; Rue and Held, 2005), where they are mainly used to model spatial correlation. In our geophysical context we apply a spatial prior to the parameters rather than to the error structure, since the parameters represent velocity anomalies in three-dimensional space. Thanks to the sparsity of the linearized physical forward matrix as well as the spatial prior sampling from the posterior density, a high-dimensional multivariate Gaussian can be achieved by a Cholesky decomposition technique from Wilkinson and Yeung (2002) or Rue and Held (2005). Their technique is improved by using a different permutation algorithm. To demonstrate the method, we estimate a three-dimensional model of mantle structure, that is, variations in seismic wave velocities, beneath the Unites States down to 800 km depth.

Our approach is also applicable to other kinds of travel time tomography, such as cross-borehole tomography or mining-induced seismic tomography (Dębski, 2010). Other

types of tomography, such as X-ray tomography in medical imaging, can also be recast as a linear matrix problem of large size with a very sparse forward matrix. However, the response is measured on pixel areas and, thus, the error structure is governed by a spatial Markov random field, while the regression parameters are modeled non-spatially using for example Laplace priors (Kolehmainen et al., 2007; Mohammad-Djafari, 2012). Some other inverse problems such as image deconvolution and computed tomography (Bardsley, 2012), electromagnetic source problems deriving from electric and magnetic encephalography, cardiography (Hämäläinen and Ilmoniemi, 1994; Uutela et al., 1999; Kaipio and Somersalo, 2007) or convection-diffusion contamination transport problems (Flath et al., 2011) can be also written as linear models. However, the physical forward matrix of those problems is dense in contrast to the situation we consider. For solutions to these problems, matrix-inversion or low-rank approximation to the posterior covariance matrix, as introduced in Flath et al. (2011), are applied to high-dimensional linear problems. In image reconstruction problems Bardsley (2012) demonstrates Gibbs sampling on (1D and 2D-) images using an intrinsic GMRF prior with the preconditioned conjugate gradient method in cases where efficient diagonalization or Cholesky decomposition of the posterior covariance matrix is not available. In other tomography problems, such as electrical capacitance tomography, electrical impedance tomography or optical absorbtion and scattering tomography, the physical forward model cannot be linearized, so that the Bayesian treatment of those problems is limited to low-dimensions (Kaipio and Somersalo, 2007; Watzenig and Fox, 2009).

2. Integrated nested Laplace approximation using the stochastic partial differential equation approach

In this approach we improve on three aspects of previous work using the MCMC method: (1) spatial modeling of 3-D earth's velocity structure, (2) spatial modeling of the data errors and, (3) efficient Bayesian uncertainty analysis.

(1) In this methodology the parameterization of the earth's interior is achieved through a highly irregular tetrahedral mesh of thousands of vertices whose spaced by 60 km to 200 km of kilometers. The resulting, large number of velocity parameters represents an approximation to the continuous seismic velocity field. For modeling the 3-D structure of this velocity field, Zhang et al. (2013a) define a neighborhood within a fixed distance of the velocity parameters, but the number of neighboring vertices within a fixed distance is influenced by the geometry of the triangulation. In this approach we work in a more

general setting, i.e., independent of the geometry of the mesh. We develop a model for the velocity field on a continuous 3-D domain by means of the Gaussian field approximation, based on the theory of stochastic partial differential equations (SPDE) introduced by Lindgren et al. (2011). Gaussian fields (GF) are widely used in spatial statistics to model spatially continuous random effects over a domain of interest. They represent processes that exist independently of whether they are observed in a given location or not.

In the SPDE approach, the dense covariance matrix of a GF from the Matérn class is approximated by the sparse structure of the Gaussian Markov random field (GMRF) using a finite-dimensional basis function representation based on the finite element method (FEM). The sparse precision matrix of the GMRF arising from the SPDE approximation provides a huge computational advantage when dealing with Bayesian inference, since efficient numerical methods, such as fast matrix factorization, can be applied (Rue and Held, 2005). We show an application inverting by 8977 parameters that quantify 3-D velocity structure of the earth's upper mantle down to 800 km depth under the western United States. Our 3-D approximation of a continuous GF with a GMRF opens a new route to efficiently model dependency in many high-dimensional physical and geoscience problems in the physics and geosciences, such as atmosphere/space tomography (Aso et al., 2008), weather and climate forecasting (Möller et al., 2012), and medical imaging (Harrison and Green, 2010).

(2) Modeling of spatial errors is not common in seismic inverse problems. The errors are typically assumed to be Gaussian and no spatial correlation of the data is allowed for by the models. Generally, real seismic data observed at the surface are spatially dependent. The major part of the dependency is eliminated by the physical model modeling the 3-D velocity field. The rest may be caused by unknown spatial errors of the data observed on earth's surface, for example by imperfect knowledge of the characteristics of the wave sources (earthquakes). In the approach in Chapter 4, we take into account and model spatially correlated data errors of both source (or earthquake) and receiver (or station) locations by means of the SPDE approach of Lindgren et al. (2011). The random field over sources is defined on a curved space over the entire earth's surface, and the receiver field is defined on a curved space that covers the western United States. With prediction maps we can identify locations at which the data may not be well explained by the physical model and may show a systematic error at the sources or receivers. As in the Kriging methodology, we produce maps of optimal predictions and associated prediction standard errors from incomplete and noisy spatial data errors (Cressie, 1993). General Kriging methods in large-scale data sets can be found in Furrer et al. (2007) and Banerjee et al.

(2008).

(3) Along with the SPDE approach, the Bayesian inference in our application is carried out by the integrated nested Laplace approximation (INLA) algorithm developed in Rue et al. (2009). INLA is an algorithm tailored to the class of latent Gaussian models. It exploits deterministic nested Laplace approximations and provides a faster and more accurate alternative to stochastic simulations. It is computationally more efficient than MCMC while yielding accurate approximations to the posterior distributions. Incorporating the powerful properties of the SPDE approach, INLA has become very popular in Bayesian modeling of large-scale spatial data over the past years due to its computational advantage. For example, Schrödle and Held (2011) applied INLA in a spatio-temporal disease mapping problem. Simpson et al. (2011) proposed a new formulation of the log-Gaussian Cox processes using SPDE/INLA and conducted inference for a data set on a globe. Cameletti et al. (2012) considered a hierarchical spatio-temporal model for particulate matter concentration in northern Italy. Möller et al. (2012) applied the SPDE/INLA methods to climate forecasting models, jointly using a copula to model the dependency between the variables. Detailed description on the theory of the SPDE and INLA approaches, as well as many code examples can be found in Simpson et al. (2012); Lindgren (2012); Illian et al. (2012), or at the webpage of the R-INLA package (www.r-inla.org). So far, the SPDE approach within the INLA framework has been mainly applied to spatial modeling of large data sets on $\mathbb{R}^2$ or $\mathbb{S}^2$ manifolds. Here we demonstrate a novel application in a full 3-D space and deploy the INLA program for our Bayesian inference of about 13,000 seismic velocity parameters, assimilating over 53,000 observations globally. We show that the INLA algorithm could achieve a speedup of about 1.5 to 2 times compared to the MCMC algorithm.

Achievements

To summarize, this thesis has made several advances in statistical parameter estimation and spatial modeling for high-dimensional seismic tomographic problems. Two Bayesian modeling techniques have been developed. One is a sampling approach based on an efficient MCMC algorithm. The other one is the INLA method based on direct approximation to the posterior distributions. The main achievements of this dissertation are:

- We approach linearized tomographic problems in a Bayesian framework using an efficient MCMC algorithm for sampling over 11,000 seismic velocity and source

correction parameters. To our knowledge, this is by far the highest dimensional application of Monte Carlo sampling to a seismic tomographic problem.

- We developed flexible spatial priors using a Gaussian Markov random field (GMRF) for the seismic velocity anomalies in three-dimensional space.
- We calculate the precision matrix for the 3-D Gaussian Matérn field explicitly based on the theory of the SPDEs introduced by Lindgren et al. (2011). Thereby, we improve on earlier work on GMRF's by modeling the continuous velocity field in 3-D using an appropriate Gaussian field approximation.
- In estimating the velocity parameters in the earth's interior we allow for spatially correlated data errors which depend on both source and receiver locations at its surface.
- We adopt the INLA algorithm by Rue et al. (2009) to improve the computational efficiency of the Bayesian inference. We show that the INLA algorithm achieves a speedup of about 1.5 to 2 times compared to the MCMC algorithm of Zhang et al. (2013a).
- Both approaches developed in this thesis are applied to estimate a three-dimensional model of mantle structure, that is, variations in seismic wave velocities, beneath the Unites States down to 800 km depth. This continental-scale study uses approximately 53,000 seismological data observations from the continental-scale USArray experiment and reveals major structures of the mantle beneath the western USA with uncertainty assessments on over 11,000 parameters.

Thesis organization

The remainder of this thesis is organized as follows: Chapter 2 describes the general setting of the linear inverse problem of seismic tomography along with the geophysical forward model and the seismic travel time data. Chapter 3 discusses the efficient Metropolis-Gibbs sampling algorithm developed and implemented for estimating the high-dimensional parameter vectors of seismic velocities and source corrections. These results were published in the *Annals of Applied Statistics*. Chapter 4 describes the spatial modeling technique using the SPDE approach. It applies the INLA technique to estimate parameters in 2-D and 3-D spaces. Results of this chapter are submitted to the *Royal Journal of Statistical*

Society: Series C (Applied Statistics). Both articles are co-authored with Prof. Dr. Claudia Czado and Prof. Dr. Karin Sigloch.

Chapter 2

The linear inverse problem of seismic tomography

In this chapter, based on Zhang et al. (2013a,b), we introduce the physics and the structure of the seismological data. We also discuss the well-established modeling techniques in seismic tomography. In Chapter 3 and 4 we present our statistical models tailored to this type of tomography problem.

Every larger earthquake generates seismic waves of sufficient strength to be recorded by seismic stations around the globe. Such seismograms are time series at discrete surface locations, that is, spatially sparse point samples of a continuous wavefield that exists everywhere inside the earth and at its surface. Figure 2.1 illustrates the spatial distribution of sources (large earthquakes, blue) and receivers (seismic broadband stations, red) that generated our data. The data consist of traveltime anomalies which indirectly reflect wave velocity variations inside the mantle. Traveltime anomalies are derived from seismograms by cross-correlating the observed waveform (in a suitable time window containing P-waves) with its forward-predicted waveform computed in a (spherically symmetric) reference model. Time lags indicate that the wave sampled seismically slow material than assumed by the reference model, whereas traveltime advances indicate anomalously fast seismic structure somewhere on the wave path. Seismic velocity variations as a function of location in 3-D space are the parameters to solve for. The measure of misfit is the sum of the squared traveltime anomalies.

Each datum y_i measures the difference between an observed arrival time y_i^{obs} of a

seismic wave i (source-receiver combination) and its predicted arrival time y_i^{pred}:

$$y_i = y_i^{\text{obs}} - y_i^{\text{pred}}.$$

y_i^{pred} is evaluated using the spherically symmetric reference model IASP91 by Kennett and Engdahl (1991). For the teleseismic P waves used in our application, this difference y_i would typically be on the order of one second, whereas y_i^{obs} and y_i^{pred} are on the order 600-1000 seconds. y_i can be explained by slightly decreasing the modeled velocity in certain subvolumes of the mantle.

We adopt the parametrization and a subset of the data measured by Sigloch et al. (2008). The earth is meshed as a sphere of irregular tetrahedra with 92,175 mesh nodes. At each mesh mode, the parameters of interest are the relative velocity variation of the mantle with respect to the reference velocity of spherically-symmetric model IASP91 (Kennett and Engdahl, 1991). The parameter vector is denoted as $\boldsymbol{\beta} := (\beta(\boldsymbol{r}), \boldsymbol{r} \in M_{\text{Earth}}) \in \mathbb{R}^{92{,}175}$, where the set of mesh node M_{Earth} fills the entire interior of the earth. The finite element discretization on a tetrahedral mesh with basis function $b_j(\boldsymbol{r})$ at location $\boldsymbol{r}$ is given by

$$\beta(\boldsymbol{r}) = \sum_{j=1}^{p} b_j(\boldsymbol{r})\beta_j, \quad b_j(\boldsymbol{r}_i) := \begin{cases} 1 & : \quad i = j \\ 0 & : \quad i \neq j. \end{cases} \tag{2.1}$$

at the vertices $\boldsymbol{r}_i$'s and linearly interpolated for other locations using the tetrahedron containing $\boldsymbol{r}$ (Sambridge and Gudmundsson, 1998). Since the magnitude of $\beta(\boldsymbol{r})$ is only on the order of few percent, the wave equation can be linearized around the spherical symmetric reference earth model using finite-frequency theory (Dahlen et al., 2000):

$$y_i = \iiint_{Earth} x_i(\boldsymbol{r})\beta(\boldsymbol{r}) d^3\boldsymbol{r}, \tag{2.2}$$

where $x_i(\boldsymbol{r}) \in \mathbb{R}$ represents the Fréchet sensitivity kernel of the ith wavepath, that is, the partial derivatives of the chosen misfit measure or data y_i with respect to the parameters $\beta(\boldsymbol{r})$. Taking the discrete representation of the velocity field in (2.1) into account, (2.2) takes the form

$$\begin{aligned} y_i &= \iiint_{Earth} x_i(\boldsymbol{r}) \sum_{j=1}^{p} b_j(\boldsymbol{r})\beta_j d^3\boldsymbol{r} = \sum_{j=1}^{p} \left[\iiint_{Earth} x_i(\boldsymbol{r}) b_j(\boldsymbol{r}) d^3\boldsymbol{r} \right] \beta_j = \sum_{j=1}^{p} x_{ij}\beta_j \\ &= \boldsymbol{x}_i'\boldsymbol{\beta}, \end{aligned} \tag{2.3}$$

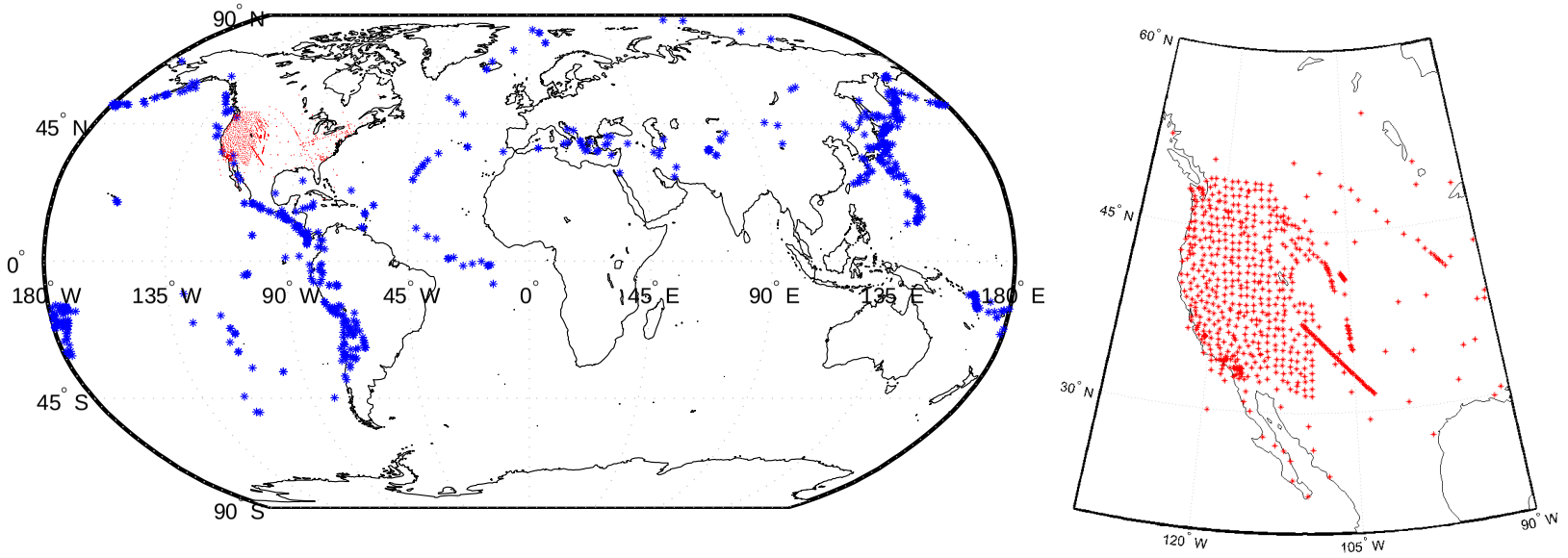

Figure 2.1: Left: Distribution of the seismic wave sources (large earthquakes, blue) and receivers (seismic broadband stations, red) that generated our data. Right: the target region of tomographic inversion, including a more detailed view of receiver distribution. This is a regional tomography study that includes only data recorded in North America. In the mantle under this region, down to a few hundreds of kilometers depth, paths of incoming waves cross densely and from many directions, yielding good resolution for a three-dimensional imaging study.

Geometrically speaking, row vector $\boldsymbol{x}_i'$ maps out the mantle subvolume that would influence the travel time y_i if some velocity anomaly $\beta(\boldsymbol{r})$ were located within it. This sensitivity region between an earthquake and a station essentially has ray-like character (Figure 2.2), though in physically more sophisticated approximations, the ray widens into a banana shape (Dahlen et al., 2000). Over the past decade, intense research effort has gone into the computability of sensitivity kernels under more and more realistic approximations (Dahlen et al., 2000; Tromp et al., 2005; Tian et al., 2007; Nolet, 2008). Since this issue is only tangential to our focus, we chose to keep the sensitivity calculations as simple as possible by modeling them as rays (the $\boldsymbol{x}_i'$ are computed only once and stored). We note that the dependence of x_i on $\boldsymbol{\beta}$ can be neglected, as is common practice. This is justified by two facts: (i) velocity anomalies $\boldsymbol{\beta}$ deviate from those of the (spherically symmetric) reference model by only a few percent, since the very mobile mantle does not support larger disequilibria, and (ii), even though the ray path in the true earth differs (slightly) from that in the reference model, this variation affects the travel time observable only to second order, according to Fermat's principle (and analogous arguments for true finite-frequency sensitivities, Dahlen et al. (2000); Nolet (2008); Mercerat and Nolet (2013)). Whatever the exact modeling is, it is very sparse, since every ray or banana visits only a small subvolume of the entire mantle – this sparsity is important for the computational efficiency of the MCMC sampling or the INLA method.

Gathering all N observations, (2.3) can be rewritten as $\boldsymbol{y} = X\boldsymbol{\beta}$, where sparse matrix $X \in \mathbb{R}^{N\times p}$ contains in its rows the N sensitivity kernels. The left panel of Figure 2.2 illustrates the sensitivity kernels between one station and several earthquakes (i.e., several matrix rows). In practice, the problem never attains full rank, so that regularization must be added to remove the remaining non-uniqueness. The linear system $\boldsymbol{y} = X\boldsymbol{\beta}$ is usually solved by some sparse matrix solver – a popular choice is the Sparse Equations and Least Squares (LSQR) algorithm by Paige and Saunders (1982), which minimizes $\|X\boldsymbol{\beta} - y\|^2 + \lambda^2\|y\|^2$, where λ is a regularization parameter. The effect is to remove non-uniqueness from the system, essentially by adding a multiple of the identify matrix onto the normal equations (Nolet, 1987; Tian et al., 2009; Sigloch, 2011).

In summary, we have formulated the seismic tomography problem as it is overwhelmingly practiced by the geophysical community today. We use travel time differences y_i as the misfit criterion, that is, as input data to the inverse problem, and seek to estimate the three-dimensional distribution of seismic velocity deviations $\boldsymbol{\beta}$ that have caused these travel time anomalies. The sensitivity kernels $\boldsymbol{x}_i'$ are modeled using ray theory, a high-frequency approximation to the full wave equation. In the conventional optimization

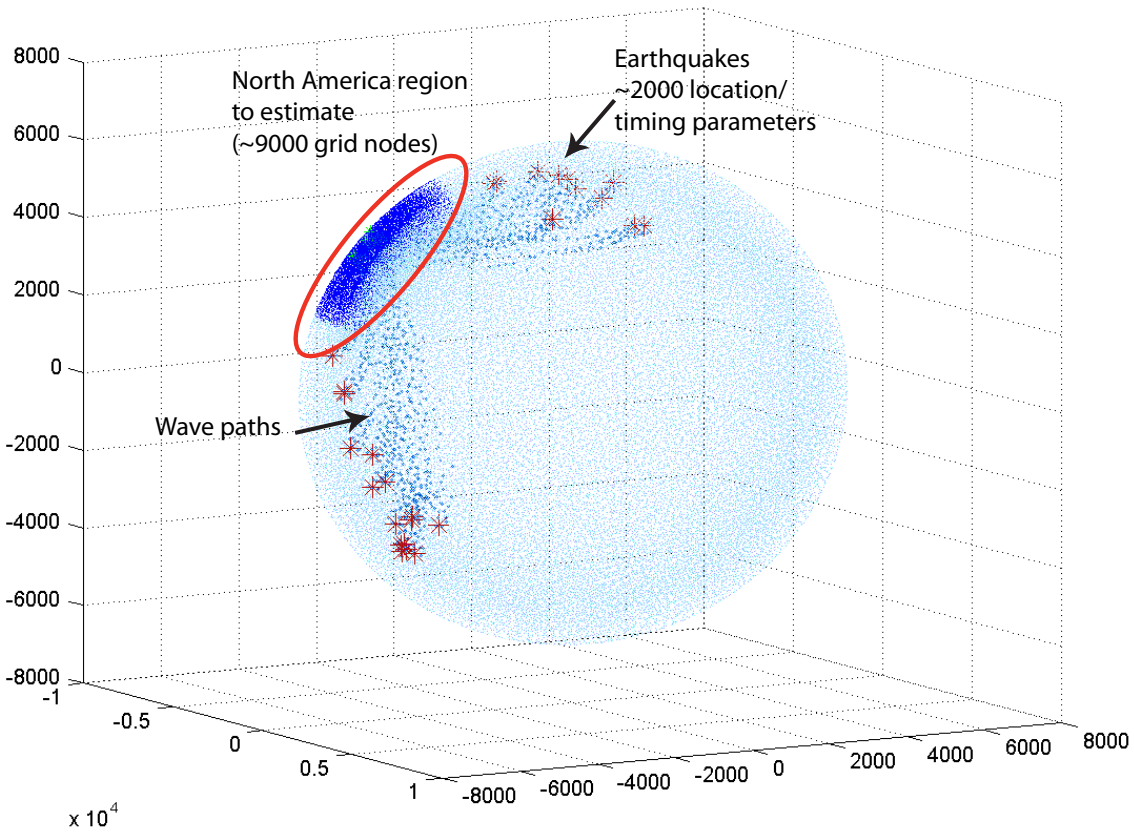

Figure 2.2: Physical setup and forward modeling of the seismic tomography problem. Parametrization of the spherical earth. Grid nodes are shown as blue dots. The goal is to estimate seismic velocity deviations $\boldsymbol{\beta}$ at $\sim$ 9000 grid nodes under North America, inside the subvolume marked by the red ellipse. Red stars mark a few of the earthquake sources shown in Figure 1. The densified point clouds, between the sources and a few stations in North America, map out the sensitivity kernels of the selected wave paths. Each sensitivity kernel fills one row of matrix X.

approach, a regularization term is added, and the inverse problem is solved by minimizing the L2 norm misfit.

Chapter 3

A Bayesian linear model for the high-dimensional inverse problem of seismic tomography

3.1 Introduction

In this chapter, based on Zhang et al. (2013a), we develop and implement a linear Bayesian model to seismic tomography. This involves a high-dimensional inverse problem in geophysics. The objective is to estimate the three-dimensional structure of the earth's interior from data measured at its surface. Since this typically involves estimating thousands of unknowns or more, it has always been treated as a linear(ized) optimization problem. Here we present a Bayesian hierarchical model to estimate the joint distribution of earth structural and earthquake source parameters. An ellipsoidal spatial prior allows to accommodate the layered nature of the earth's mantle. With our efficient algorithm we sample the posterior distributions for large-scale linear inverse problems, and provide precise uncertainty quantification in terms of the posterior distributions of the parameters. This allows to construct credible intervals given the data. We apply the method to a full-fledged tomography problem, an inversion for upper-mantle structure under western North America that involves more than 11,000 parameters. In studies on simulated and real data, we show that our approach retrieves the major structures of the earth's interior similarly well as classical least-squares minimization, while additionally providing uncertainty assessments.

3.2 Setup of the statistical model

As shown in Chapter 2 the earth is parameterized as a sphere of irregular tetrahedra containing 92,175 tetrahedral nodes which represent the velocity deviation parameters. Since all 92,175 velocity deviation parameters of the entire earth are currently not manageable for MCMC sampling, we regard as free parameters only 8977 of those parameters which are located beneath the western U.S., that is between latitudes $20°N$ to $60°N$, longitudes $90°W$ to $130°W$, and 0-800km depth. Tetrahedra nodes are spaced by 60-150km. We denote this subset of velocity parameters as $\boldsymbol{\beta}_{\text{usa}}$.

Besides velocity parameters, we also consider the uncertainty in the location and the origin time of each earthquake source, which contribute to the travel time measurement. Government and research institutions routinely publish location estimates for every larger earthquake, but any event may easily be mistimed by a few seconds, and mislocated by ten or more kilometers (corresponding to a travel duration of 1 s or more). This is a problem, since the structural heterogeneities themselves only generate travel time delays on the order of a few seconds. Hence the exact locations and timings of the earthquakes – or rather: their deviations from the published catalogue values – need to be treated as additional free parameters, to be estimated jointly with the structural parameters. These so-called "source corrections" are captured by three-dimensional shift corrections of the hypocenter ($\boldsymbol{\beta}_{\text{hyp}}$) and time corrections ($\boldsymbol{\beta}_{\text{time}}$) per earthquake.

Using the LSQR method, Sigloch et al. (2008) jointly estimate all 92,175 parameters together with these "source corrections". Using those LSQR solutions we have two modeling alternatives for the earth structural inversion with N travel delay time observations:

$$\textbf{Model 1:} \quad \boldsymbol{y}_{\text{usa}} = X_{\text{usa}}\boldsymbol{\beta}_{\text{usa}} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}_N(\mathbf{0},\, \tfrac{1}{\phi} I_N), \tag{3.1}$$

where $X_{\text{usa}} \in \mathbb{R}^{N\times 8977}$ denotes the ensemble of sensitivity kernels of the western USA. $\mathcal{N}_N(\boldsymbol{\mu},\, \Sigma_y)$ denotes the N-dimensional multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance Σ, and the N-dimensional unity matrix is denoted by I_N. In Model 1, we only estimate the velocity parameters $\boldsymbol{\beta}_{\text{usa}}$ using the travel delay time $\boldsymbol{y}_{\text{usa}} \in \mathbb{R}^N$ (the path DC) and keep the part of the travel delay time for the corrections parameters (path AB in right panel of Figure 3.1) fixed at the LSQR solutions of $\boldsymbol{\beta}_{\text{hyp}}$ and $\boldsymbol{\beta}_{\text{time}}$ estimated by Sigloch et al. (2008). The extended model with joint estimation of source corrections is

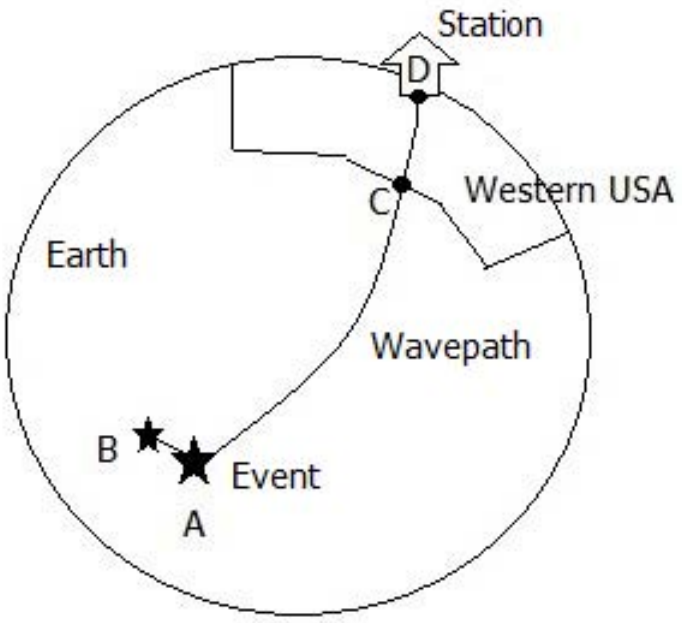

Figure 3.1: Schematic illustration of the components of an individual wave path.

given by

$$\textbf{Model 2:} \quad \boldsymbol{y}_{\text{cr}} = X_{\text{usa}}\boldsymbol{\beta}_{\text{usa}} + X_{\text{hyp}}\boldsymbol{\beta}_{\text{hyp}} + X_{\text{time}}\boldsymbol{\beta}_{\text{time}} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}_N(\mathbf{0}, \tfrac{1}{\phi} I_N), \tag{3.2}$$

Here we apply the travel delay time $\boldsymbol{y}_{\text{cr}}$ assuming that the part of the travel time running through path AC is given. This given part of the travel times is again based on the LSQR solution estimated by Sigloch et al. (2008).

The number of travel time data from source-receiver pairs is $N = 53,270$, collected from 760 stations and 529 events. The number of hypocenter correction parameters is 1587 (529 earthquakes $\times$ 3) and there are 529 time correction parameters. Sigloch (2008) found that in the uppermost mantle, between 0 km to 100 km depth, the velocity can deviate by more than $\pm 5\%$ from the spherically symmetric reference model. As depth increases, the mantle becomes more homogeneous, and the velocity deviates less from the reference model.

3.3 Estimation method

3.3.1 Modeling the spatial structure of the velocity parameters

In both models (3.1) and (3.2) we have the spatial parameter $\boldsymbol{\beta}_{\text{usa}}$, which we denote generically as $\boldsymbol{\beta}$ in this section. In the Bayesian approach we need a proper prior distribution for this high-dimensional parameter vector $\boldsymbol{\beta}$. To account for their spatially correlated structure, we apply the conditional autoregressive model (CAR) and assume a Markov random field structure for $\boldsymbol{\beta}$. This assumption says that the conditional distribution of the local characteristics β_i, given all other parameters β_j, $j \neq i$, only depends on the neighbors, that is,

$$P(\beta_i \mid \boldsymbol{\beta}_{-i}) = P(\beta_i \mid \beta_j, j \sim i),$$

where $\boldsymbol{\beta}_{-i} := (\beta_1, ..., \beta_{i-1}, \beta_{i+1}, ..., \beta_d)'$ and '$\sim i$' denotes the set of neighbors of site i. The CAR model and its application have been investigated in many studies, such as Pettitt et al. (2002) or Rue and Held (2005). Since the earth is heterogeneous and layered, lateral correlation length scales are larger than over depths, and so we propose an ellipsoidal neighborhood structure for the velocity parameters. Let $(x_j, y_j, z_j)' \in \mathbb{R}^3$ be the positions of the ith and the jth nodes in Cartesian coordinates. The jth node is a neighbor of node i if the ellipsoid equation is satisfied, that is,

$$\left(\frac{x_i - x_j}{D_x}\right)^2 + \left(\frac{y_i - y_j}{D_y}\right)^2 + \left(\frac{z_i - z_j}{D_z}\right)^2 \leqslant 1.$$

To add a rotation of the ellipsoid to an arbitrary direction in the space we could simply modify the vector $(x_i - x_j, y_i - y_j, z_i - z_j)'$ to $R(x_i - x_j, y_i - y_j, z_i - z_j)'$ with a rotation matrix

$$R := R_x R_y R_z,$$

for given rotation matrices R_x, R_y and R_z in the x, y and z directions, respectively. The spherical neighborhood structure is a special case of the ellipsoidal structure with $D_x = D_y = D_z$. Let D be the maximum distance of D_x, D_y and D_z.

For weighting the neighbors we adopt either the exponential $w_e(\cdot)$ or reciprocal weight functions $w_r(\cdot)$, that is,

$$w_e(d_{ij}) := \exp\{-\tfrac{3d_{ij}^2}{D^2}\} \quad \text{and} \quad w_r(d_{ij}) := \tfrac{D}{d_{ij}} - 1, \tag{3.3}$$

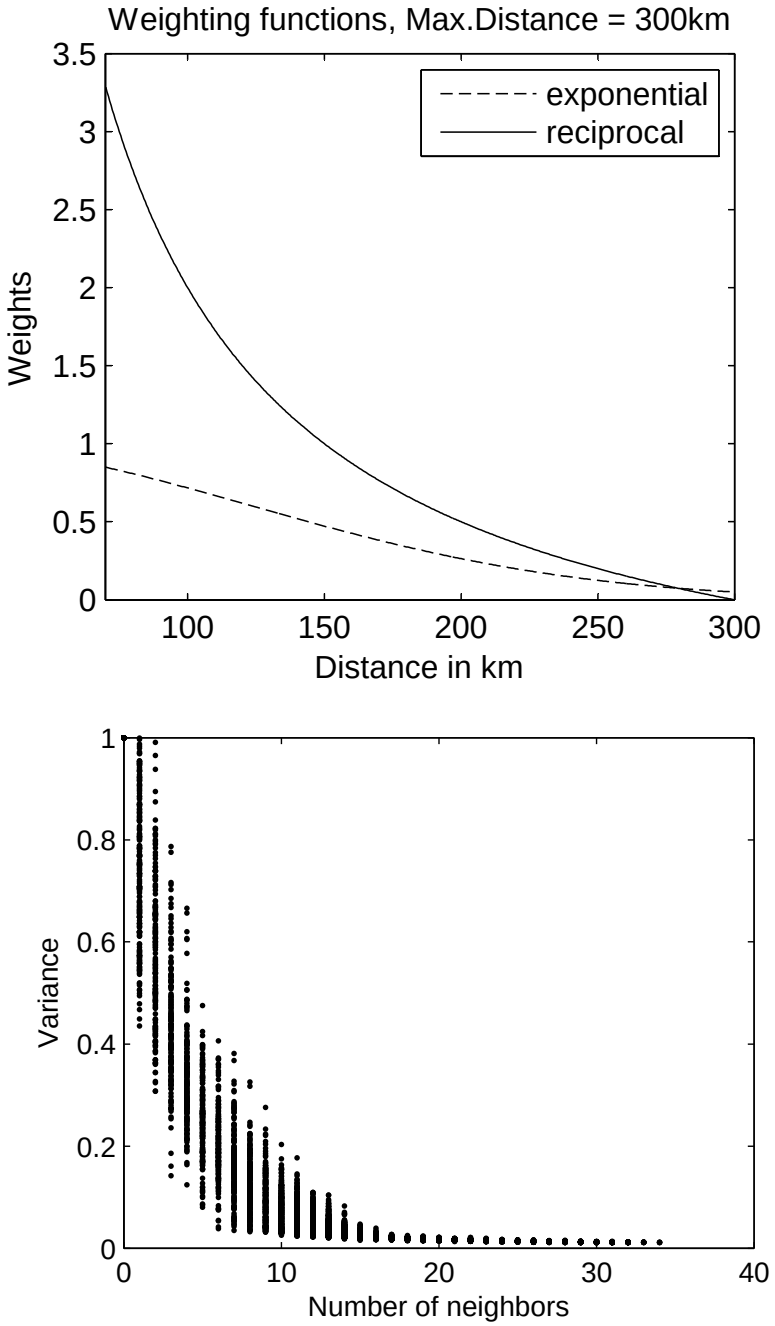

Figure 3.2: Top: Exponential and reciprocal weight functions for the spatial prior, for $D = 150$ km and $D = 300$ km. Bottom: the trade-off relationship between numbers of neighbors and the prior variance $\text{diag}(Q^{-1}(\psi))$, $\psi = 10$, $D = 150\ km$, $w =$ reciprocal weights.

where d_{ij} is the Euclidean distance between node i and node j. The exponential weight function is bounded while the reciprocal weight function is unbounded. Those weighting functions have been studied by Pettitt et al. (2002) or Congdon (2003). The top panel of Figure 3.2 illustrates the weight functions for $D = 300$ km.

Let $\omega(d_{ij})$ be either $w_e(\cdot)$ or $w_r(\cdot)$ in (3.3). To model the spatial structure of $\boldsymbol{\beta}_{\text{usa}}$ in (3.1) and (3.2), a CAR model is used. Following Pettitt et al. (2002) let $\boldsymbol{\beta}_{\text{usa}} \sim \mathcal{N}_{p_{\text{usa}}}\left(\mathbf{0}, \frac{1}{\eta_{\text{usa}}} Q^{-1}(\psi)\right)$ with precision matrix

$$Q_{ij}(\psi) := \begin{cases} 1 + |\psi| \sum_{i:j\sim i} \omega(d_{ij}) & : \quad i = j \\ -\psi\omega(d_{ij}) & : \quad i \neq j, i \sim j \text{ for } \psi \in \mathbb{R}. \end{cases} \tag{3.4}$$

They showed that Q is symmetric and positive definite, and that conditional correlations can be explicitly determined. For $\psi \to 0$, the precision matrix Q converges to the identity matrix, that is, $\psi = 0$ corresponds to independent elements of $\boldsymbol{\beta}_{\text{usa}}$. The precision matrix in (3.4) for both elliptical and spherical cases indicates anisotropic covariance structure and depends on the distance between nodes, the number of neighbors of each node, and the weighting functions. The elliptical precision matrix additionally depends on the orientation. The bottom panel of Figure 3.2 shows the trade-off between numbers of neighbors and prior variance, which indicates that the more neighbors the ith node has, the smaller is its prior variance $(Q^{-1}(\psi))_{ii}$. Posterior distribution of velocity parameters from regions with less neighborhood information can be rough, since they are not highly regularized due to the large prior covariances. This may produce sharp edges in the tomographic image. However, this is a realistic modeling method since one is more sure about the optimization solution if a velocity parameter has more neighbors. Moreover, this prior specification is adapted to the construction of the tetrahedral mesh: regions with many nodes have better ray coverage than regions with less nodes. In summary, the prior incorporates diverse spatial knowledge about the velocity parameters. Since a precision matrix is defined, which is sparse and positive definite, it provides a computational advantage in sampling from a high-dimensional Gaussian distribution as required in our algorithm (shown in the following sections).

3.3.2 A Gibbs-Metropolis sampler for parameter estimation in high dimensions

To quantify uncertainty, we adopt a Bayesian approach. Posterior inference for the model parameters is facilitated by a Metropolis within Gibbs sampler (Brooks et al., 2011). Recall the linear model in (3.2),

$$\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}_N(\mathbf{0}, \tfrac{1}{\phi} I_N),$$

where $\boldsymbol{\beta} := (\boldsymbol{\beta}_{\text{usa}}, \boldsymbol{\beta}_{\text{hyp}}, \boldsymbol{\beta}_{\text{time}})'$ and $X := (X_{\text{usa}}, X_{\text{hyp}}, X_{\text{time}})$. We now specify the prior distribution of $\boldsymbol{\beta}$ as

$$\boldsymbol{\beta} \sim \mathcal{N}_p(\boldsymbol{\beta}_0, \Sigma_\beta) \quad \text{with } \boldsymbol{\beta}_0 := (\boldsymbol{\beta}_{0,\text{usa}}, \boldsymbol{\beta}_{0,\text{hyp}}, \boldsymbol{\beta}_{0,\text{time}})'.$$

Here, p is given by $p := p_{\text{usa}} + p_{\text{hyp}} + p_{\text{time}} = 8977 + 1597 + 529 = 11103$. The prior covariance matrix Σ_β is chosen as

$$\Sigma_\beta := \begin{pmatrix} \frac{1}{\eta_{\text{usa}}} Q^{-1}(\psi) & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \frac{1}{\eta_{\text{hyp}}} I_{p_{\text{hyp}}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \frac{1}{\eta_{\text{time}}} I_{p_{\text{time}}} \end{pmatrix}. \tag{3.5}$$

Since we are interested in modeling positive spatial dependence, we impose that the spatial dependence parameter ψ is the truncated normal distribution a priori, that is,

$$\psi \sim \mathcal{N}(\mu_\psi, \sigma_\psi^2) \mathbb{1}(\psi > 0).$$

The priors for the precision scale parameters η_{usa}, η_{hyp}, η_{time} and ϕ are specified in terms of a Gamma distribution $\Gamma(a, b)$ with density $g(x;\, a, b) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp\{-bx\}$, $x > 0$. The corresponding first two moments are $\frac{a}{b}$ and $\frac{a}{b^2}$, respectively.

The MCMC procedure is derived as follows: The full conditionals of $\boldsymbol{\beta}$ are

$$\begin{aligned} \boldsymbol{\beta} \mid \boldsymbol{y}, \psi, \boldsymbol{\eta} &\sim \mathcal{N}_p(\Omega_\beta^{-1} \boldsymbol{\xi}_\beta,\, \Omega_\beta^{-1}), \\ &\text{with } \Omega_\beta := \Sigma_\beta^{-1} + \phi X'X,\, \boldsymbol{\xi}_\beta := \Sigma_\beta^{-1} \boldsymbol{\beta}_0 + \phi X' \boldsymbol{y}, \end{aligned} \tag{3.6}$$

and $\boldsymbol{\eta} := (\eta_{\text{usa}}, \eta_{\text{hyp}}, \eta_{\text{time}})$. For η_{usa}, η_{hyp}, η_{time}, and ϕ, the full conditionals are again Gamma distributed. The estimation of ψ requires a Metropolis-Hastings (MH) step. The logarithm

of the full conditional of ψ is proportional to

$$\log \pi(\psi \mid \boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{\eta}) \propto \frac{1}{2} \log |Q(\psi)| - \frac{\eta_{\text{usa}}}{2} (\boldsymbol{\beta}_{\text{usa}} - \boldsymbol{\beta}_{0,\text{usa}})' Q(\psi) (\boldsymbol{\beta}_{\text{usa}} - \boldsymbol{\beta}_{0,\text{usa}}) - \frac{(\psi - \mu_\psi)^2}{2\sigma_\psi^2}.$$

For the MH step, we choose a truncated normal random walk proposal for ψ to obtain a new sample, that is, $\mathcal{N}(\psi^{\text{old}}, \bar{\sigma}_\psi^2)\mathbb{1}(\psi > 0)$. We use a Cholesky decomposition with permutation to obtain a sample of $\boldsymbol{\beta}$ in (3.6) (Section 3.4). The method by Pettitt et al. (2002), solving a sparse matrix equation, is not useful. Here, computing the determinant of the Cholesky factor of $Q(\psi)$ is much more efficient than calculating its eigenvalues, due to the size and sparseness of $Q(\psi)$.

3.3.3 Relationship to ridge regression

To show the relationship between our approach and ridge regression (also called Tikhonov regularization), we consider only Model 1. For simplicity we neglect the notation "usa" in (3.1). The analysis is also applicable to Model 2.

Let $\hat{\boldsymbol{\beta}}^{\text{ridge}}(\lambda) := (X'X + \lambda I_p)^{-1} X'\boldsymbol{y}$ be the corresponding ordinary ridge regression (ORR) estimate with shrinkage parameter λ (Hoerl and Kennard, 1970; Swindel, 1976; Dębski, 2010). For given hyperparameters η, ϕ and ψ, the full conditional of $\boldsymbol{\beta}$ is $\boldsymbol{\beta} \mid \boldsymbol{y}, \eta, \phi, \psi \sim \mathcal{N}_p(\Omega_\beta^{-1}\boldsymbol{\xi}_\beta, \Omega_\beta^{-1})$ with $\Omega_\beta := \eta Q(\psi) + \phi X'X$ and $\boldsymbol{\xi}_\beta := \eta Q(\psi)\boldsymbol{\beta}_0 + \phi X'\boldsymbol{y}$. The corresponding full conditional mean can therefore be expressed as

$$E[\boldsymbol{\beta}|\boldsymbol{y}, \eta, \psi, \eta] = \left(X'X + \tfrac{\eta}{\phi} Q(\psi)\right)^{-1} \left(X'y + \tfrac{\eta}{\phi} Q(\psi)\boldsymbol{\beta}_0\right).$$

This is close to the modified ridge regression estimator

$$\hat{\boldsymbol{\beta}}^{\text{ridge}}(\lambda, \boldsymbol{\beta}_0) := (X'X + \lambda I_p)^{-1}(X'\boldsymbol{y} + \lambda\boldsymbol{\beta}_0),$$

(Swindel, 1976). We can see that if $\psi \to 0$, then $\frac{\eta}{\phi} Q(\psi) \to \frac{\eta}{\phi}$, which is the equivalent to λ in the modified ridge regression. This shows that the prior precision matrix $\eta Q(\psi)$ is a regularization matrix with parameter ψ controlling the prior covariance. As discussed in Section 3.1, the prior covariance $\frac{1}{\eta} Q^{-1}(\psi)$ also varies with the specified weights in (3.3) with maximum distance D and with number of neighboring nodes. For large ψ or large weights function values, as well as large number of neighbors, the prior variances are small, which well reflects the prior knowledge about the data coverage and parameter uncertainty. Thus, the full conditional mean is close to the prior mean in this case.

3.3.4 Computational issues

Since the size of the travel time data requires high-dimensional parameters to be estimated, the traditional method of sampling the parameter vector $\boldsymbol{\beta}$ from $\mathcal{N}_p(\Omega_\beta^{-1}\boldsymbol{\xi}_\beta, \Omega_\beta^{-1})$ directly, as defined in (3.6), is not efficient with respect to computing time. We instead use a Cholesky decomposition of Ω_β. Since the sensitivity kernel X is sparse, and the prior covariance matrix is sparse and positive definite, the matrix Ω_β remains sparse and symmetric positive definite. Therefore, we can reduce the cost of the Cholesky decompositions. For this we apply an approximate minimum degree ordering algorithm (AMD algorithm) to find a permutation P of Ω_β so that the number of nonzeros in its Cholesky factor is reduced (Amestoy et al., 1996). In our case, the number of nonzeros of the full conditional precision matrix Ω_β in (3.6) is about 5% of all elements. After this permutation the nonzeros of the Cholesky factor are reduced by 50% compared to the original number of non-zeros.

To sample a multivariate normal distributed vector after permutation, we follow Rue and Held (2005). Given the permutation matrix P of Ω_β, we sample a vector $\boldsymbol{v} := P\boldsymbol{\beta}$ with

$$\boldsymbol{v} = (L_p')^{-1}((L_p^{-1})P\boldsymbol{\xi}_\beta + \boldsymbol{Z})$$

where L_p is a lower triangular matrix resulting from the Cholesky decomposition of $P\Omega_\beta$, and $\boldsymbol{Z}$ is a standard normal distributed vector, that is, $\boldsymbol{Z} \sim \mathcal{N}_p(\mathbf{0}, I_p)$. The original parameter vector of interest $\boldsymbol{\beta}$ can be obtained after permuting vector $\boldsymbol{v}$ again. Rue and Held (2005) suggested finding a permutation such that the matrix is banded. However, we found that in our case the AMD algorithm is more efficient with regard to computing time. Using MATLAB built-in functions, the Cholesky decomposition with an approximate minimum degree ordering takes 8 seconds on a Linux-Cluster 8-way Opteron with 32 cores, while the Cholesky decomposition based on a banded matrix takes 15 seconds. The traditional method without permutation requires 118.5 seconds.

3.4 Simulation study

3.4.1 Simulation setups

In this section we examine the performance of our approach for Model 1. We want to investigate whether the method works correctly under the correct model assumptions, and how much influence the prior has on the posterior estimation. We consider five different

prior neighborhood structures of $\boldsymbol{\beta}_{\text{usa}}$:

(0) Independent model of $\boldsymbol{\beta}_{\text{usa}}$, $\psi = 0$ fixed, that is, $\boldsymbol{\beta}_{\text{usa}} \sim \mathcal{N}_{p_{\text{usa}}}(\boldsymbol{\beta}_0, \frac{1}{\eta_{\text{usa}}} I_{p_{\text{usa}}})$,

(1) Spherical neighborhood structure with reciprocal weight function,

(2) Ellipsoidal neighborhood structure with reciprocal weight function,

(3) Spherical neighborhood structure with exponential weight function,

(4) Ellipsoidal neighborhood structure with exponential weight function.

Note that the independent model of $\boldsymbol{\beta}_{\text{usa}}$ corresponds to the Bayesian ridge estimator as described in Section 3.3. For the weight functions in (3.3), we set $D_x = D_y = 300$ km and $D_z = 150$ km for modeling ellipsoidal neighborhood structures, and $D = 150$ km for the spherical neighborhood distance.

Setup I: Assume the solution by Sigloch et al. (2008), denoted as $\hat{\boldsymbol{\beta}}_{\text{usa}}^{\text{LSQR}}$, represents true mantle structure beneath North America. We use the forward model $X_{\text{usa}}\hat{\boldsymbol{\beta}}_{\text{usa}}^{\text{LSQR}}$ to compute noise-free, synthetic data. Then, we generate two types of noisy data, that is, $\boldsymbol{Y} = X_{\text{usa}}\hat{\boldsymbol{\beta}}_{\text{usa}}^{\text{LSQR}} + \boldsymbol{\epsilon}$ with

(A) Gaussian noise ($\boldsymbol{\epsilon} \sim \mathcal{N}_N(\mathbf{0}, \frac{1}{\phi_{tr}} I_N)$, $\phi_{tr} = 0.4$),

(B) t-noise ($\boldsymbol{\epsilon} \sim t_N(\mathbf{0}, I_N, \nu_{tr})$, $\nu_{tr} = 3$, corresponds to $\phi_{tr} = 0.333$).

Although we add t-noise to our synthetic earth model $\hat{\boldsymbol{\beta}}_{\text{usa}}^{\text{LSQR}}$, our posterior calculation is based on Gaussian errors. Additionally, we compare two priors for $\boldsymbol{\beta}_{\text{usa}} \sim \mathcal{N}_{p_{\text{usa}}}(\boldsymbol{\beta}_0, \frac{1}{\eta_{\text{usa}}} Q^{-1}(\psi))$ to examine the sensitivity of the posterior estimates to the prior choices:

(a) $\boldsymbol{\beta}_0 \sim \mathcal{N}_{p_{\text{usa}}}(\hat{\boldsymbol{\beta}}_{\text{usa}}^{\text{LSQR}}, 0.32^2 I_d)$,

(b) $\boldsymbol{\beta}_0 = \mathbf{0}$ (spherically symmetric reference model).

The priors for the hyperparameters are set as follows: $\psi \sim \mathcal{N}(10, 0.2^2)$, $\phi \sim \Gamma(1, 0.1)$ resulting in expectation and standard deviation of 10, $\eta_{\text{usa}} \sim \Gamma(10, 2)$ resulting in expectation of 5 and standard deviation of 1.6.

Setup II: In this case we examine the performance under known prior neighborhood structures. We construct a synthetic true mantle model with two types of known prior neighborhood structures $\boldsymbol{\beta}_{\text{usa,tr}} \sim \mathcal{N}_{p_{\text{usa}}}(\hat{\boldsymbol{\beta}}_{\text{usa}}^{\text{LSQR}}, \frac{1}{\eta_{\text{usa,tr}}} Q^{-1}(\psi_{\text{tr}}))$ with $\eta_{\text{usa,tr}} = 0.18$ and $\psi_{\text{tr}} = 10$ using

(a) a spherical neighborhood structure for $\boldsymbol{\beta}_{\text{usa,tr}}$ with reciprocal weights,

(b) an ellipsoidal neighborhood structure for $\boldsymbol{\beta}_{\text{usa,tr}}$ with reciprocal weights.

Again, Gaussian noise is added to the forward model, that is, $\boldsymbol{Y} = X_{\text{usa}}\hat{\boldsymbol{\beta}}_{\text{usa}}^{\text{LSQR}} + \boldsymbol{\epsilon}$, $\boldsymbol{\epsilon} \sim \mathcal{N}_N(0, \frac{1}{\phi_{tr}} I_N)$, $\phi_{\text{tr}} = 0.4$. Posterior estimation is carried out assuming the five different prior structures.

The number of MCMC iterations for scenarios in Setup I and Setup II is 3000, thinning is 15, and burn-in after thinning is 100. For convergence diagnostics we compute the trace, autocorrelation and estimated density plots as well as the effective sample size (ESS) using `coda` package in R for those samples. According to Brooks et al. (2011), the ESS is defined by

$$ESS := \frac{n}{1 + 2\sum_{k=1}^{\infty} \rho_k},$$

with the original sample size n and autocorrelation $\rho_k < 0.05$ at lag k. The infinite sum can be truncated at lag k when ρ_k becomes smaller than 0.05 (Kass et al., 1998; Liu, 2008).

3.4.2 Performance evaluation measures

To evaluate the results we use the standardized Euclidean norm for both data and model misfits, $\|\cdot\|_{\Sigma_y}$ and $\|\cdot\|_{\Sigma_\beta}$, respectively. The function $\|\boldsymbol{x}\|_\Sigma$ of a vector $\boldsymbol{x}$ of mean $\boldsymbol{\mu}$ and covariance Σ is called the *Mahalanobis distance*, defined by

$$\|\boldsymbol{x}\|_\Sigma := \sqrt{(\boldsymbol{x} - \boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu})}.$$

To include model complexity we calculate the deviance information criterion (DIC) (Spiegelhalter et al., 2002). Let $\boldsymbol{\theta}$ denote the parameter vector to be estimated. Furthermore, the likelihood of the model is denoted by $\ell(\boldsymbol{y}|\bar{\boldsymbol{\theta}})$, where $\bar{\boldsymbol{\theta}}$ is the estimated posterior mean of $\boldsymbol{\theta}$, estimated by $\frac{1}{R}\sum_{r=1}^{R} \boldsymbol{\theta}^r$ with R number of independent MCMC samples. According to Spiegelhalter et al. (2002) and Celeux et al. (2006), the deviance is defined as

$$D(\boldsymbol{\theta}) := -2\log(\ell(\boldsymbol{y}|\bar{\boldsymbol{\theta}})) + 2\log h(\boldsymbol{y}).$$

The term $h(\boldsymbol{y})$ is a standardizing term which is a function of the data alone and does not need to be known. Thus, for model comparison we take $D(\boldsymbol{\theta}) = -2\log(\ell(\boldsymbol{y}|\bar{\boldsymbol{\theta}}))$. The effective number of parameters in the model, denoted by p_D, is defined by

$$p_D := E_\theta[D(\boldsymbol{\theta})] - D(\bar{\boldsymbol{\theta}}).$$

The term $E_\theta[D(\boldsymbol{\theta})]$ is the posterior mean deviance and is estimated by $\frac{1}{R}\sum_{r=1}^{R} D(\boldsymbol{\theta}^r)$. This term can be regarded as a Bayesian measure of fit. In summary, the DIC is defined

as

$$DIC := E_\theta[D(\boldsymbol{\theta})] + p_D = D(\bar{\boldsymbol{\theta}}) + 2p_D.$$

The model with the smallest DIC is the preferred model under the trade-off of model fit and model complexity.

3.4.3 Results and interpretations

Table 3.1 illustrates posterior estimation results for Setup I. It shows that the estimation method with ellipsoidal prior structures (2) and (4) turn out to be the most adequate, according to the DIC criterion. The standardized data misfit criteria $\|\cdot\|_{\Sigma_y}$ given the estimated posterior mode $\hat{\boldsymbol{\beta}}_{\text{usa}}$ show similar results in all scenarios. However this measure ignores the uncertainty of $\boldsymbol{\beta}_{\text{usa}}$. The criteria $\|\boldsymbol{y} - X\hat{\boldsymbol{\beta}}_L\|_{\Sigma_y}$ and $\|\boldsymbol{y} - X\hat{\boldsymbol{\beta}}_U\|_{\Sigma_y}$ show the data misfit given the 90% credible interval with lower and upper quantile posterior estimates $\hat{\boldsymbol{\beta}}_L$ and $\hat{\boldsymbol{\beta}}_U$, respectively. These estimates give a range of the data misfit for all possible posterior solutions of $\boldsymbol{\beta}_{\text{usa}}$ and show that methods with independent prior generally yield larger ranges of misfit values than the ones with spatial structures. This indicates that the credible intervals of methods with spatial priors can fit the data better.

Further, methods with spatial priors in Setup I (b) show smaller model misfit under $\|\cdot\|_{\Sigma_\beta}$ than ones with independent prior, while in Setup I (a) results with independent priors are better. Generally, estimated posterior modes of η_{usa} vary considerably due to the different prior assumptions. Models with ellipsoidal neighborhood structures have a stronger prior (in the sense of a smaller prior variance) than models with spherical neighborhood structure. Similarly, models with reciprocal weights have a stronger regularization toward the prior mean than models with exponential weights. This means that the posterior estimates of η_{usa} adapt to different prior settings. Moreover, we notice that the estimate of the spatial dependence parameter ψ depends strongly on its prior, as the prior mean is close to the posterior estimates of ψ in all scenarios. Table 3.2 illustrates results from Setup II assuming known spatial structure including hyperparameters. The DIC values indicate that our approach correctly detects the underlying prior structures (in (a) it is prior structure (1), in (b) it is prior structure (2)). We can also observe that our approach estimates the hyperparameters correctly. Estimated posterior modes of the parameters from the identified model are close to their true values.

Generally, tomographic images illustrate velocity parameters as deviation of the solution from the spherically symmetric reference model (in %). Blue colors represent zones that have faster seismic velocities than the reference earth model while red colors denote

Setup I

(a) $\boldsymbol{\beta} \sim \mathcal{N}_{p_{\text{usa}}}(\boldsymbol{\beta}_0, \frac{1}{\eta_{usa}} Q^{-1}(\psi))$, $\boldsymbol{\beta}_0 \sim \mathcal{N}_{p_{\text{usa}}}(\hat{\boldsymbol{\beta}}_{usa}^{LSQR}, 0.32^2 I_{p_{\text{usa}}})$

Noises	Prior struct	$\|\boldsymbol{y} - X\hat{\boldsymbol{\beta}}\|_{\Sigma_y}$	$\|\boldsymbol{y} - X\hat{\boldsymbol{\beta}}_L\|_{\Sigma_y}$	$\|\boldsymbol{y} - X\hat{\boldsymbol{\beta}}_U\|_{\Sigma_y}$	$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_{tr}\|_{\Sigma_\beta}$	DIC	mode $\hat{\eta}_{usa}$
	(0)	232.28	312.82	308.27	91.30	103,748	9.28
(A) Gaussian noise	(1)	231.71	258.32	255.74	349.70	103,467	2.89
$\epsilon_i \sim N(0, 1/\phi_{tr})$	(2)	231.67	264.81	261.59	200.34	103,442	0.11
$\phi_{tr} = 0.4$,	(3)	231.77	263.89	260.96	256.52	103,478	2.70
η_{usa}, ψ unknown	(4)	231.70	267.81	264.44	185.04	103,456	0.20
(B) t-noises	(0)	227.74	602.35	604.21	46.83	112,436	0.57
$\epsilon_i \sim t(0, 1, \nu_{tr})$	(1)	228.58	443.00	443.02	69.50	112,226	0.09
$\phi_{tr} = 0.33$,	(2)	228.57	437.58	437.80	57.83	112,118	0.01
$\nu_{tr} = 3$,	(3)	228.51	450.67	450.27	62.03	112,175	0.11
η_{usa}, ψ unknown	(4)	228.52	445.22	445.42	55.87	112,126	0.01

(b) $\boldsymbol{\beta} \sim \mathcal{N}_{p_{\text{usa}}}(\mathbf{0}, \frac{1}{\eta_{usa}} Q^{-1}(\psi))$

Noises	Prior struct	$\|\boldsymbol{y} - X\hat{\boldsymbol{\beta}}\|_{\Sigma_y}$	$\|\boldsymbol{y} - X\hat{\boldsymbol{\beta}}_L\|_{\Sigma_y}$	$\|\boldsymbol{y} - X\hat{\boldsymbol{\beta}}_U\|_{\Sigma_y}$	$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_{tr}\|_{\Sigma_\beta}$	DIC	mode $\hat{\eta}_{usa}$
	(0)	234.33	635.99	632.19	50.53	106,563	0.59
(A) Gaussian noise	(1)	233.46	458.55	454.42	40.53	105,365	0.09
$\epsilon_i \sim N(0, 1/\phi_{tr})$	(2)	233.36	449.35	444.06	42.45	105,200	0.01
$\phi_{tr} = 0.4$,	(3)	233.52	466.36	462.04	38.55	105,357	0.12
η_{usa}, ψ unknown	(4)	233.40	458.05	452.60	42.71	105,256	0.01
(B) t-noises	(0)	226.53	831.85	832.84	40.10	113,023	0.19
$\epsilon_i \sim t(0, 1/\phi_{tr}, \nu_{tr})$	(1)	227.60	599.53	598.99	33.50	112,575	0.03
$\phi_{tr} = 0.33$,	(2)	227.56	596.04	595.78	33.62	112,512	0.00
$\nu_{tr} = 3$,	(3)	227.61	606.60	605.77	33.48	112,541	0.03
η_{usa}, ψ unknown	(4)	227.55	607.03	606.69	34.03	112,536	0.00

Table 3.1: Posterior estimation results of the simulation study under Setup I using synthetic earth models. The posterior mode of the velocity parameters is denoted as $\hat{\boldsymbol{\beta}}$ in the table instead of $\hat{\boldsymbol{\beta}}_{\text{usa}}$ for simplicity. $\boldsymbol{\beta}_0$ is denoted as the prior mean. The quantities $\hat{\boldsymbol{\beta}}_L$ and $\hat{\boldsymbol{\beta}}_U$ are lower and upper quantiles of the 90% credible interval of the MCMC estimates, respectively.

Setup II							
(a) $\boldsymbol{\beta} \sim \mathcal{N}_{p_{\text{usa}}}(\boldsymbol{\beta}_0, \frac{1}{\eta_{usa}} Q^{-1}(\psi))$ with a spherical neighborhood structure for Q							
Noises	Prior struct	$\|\boldsymbol{y} - X\hat{\boldsymbol{\beta}}\|_{\Sigma_y}$	$\|\boldsymbol{y} - X\hat{\boldsymbol{\beta}}_L\|_{\Sigma_y}$	$\|\boldsymbol{y} - X\hat{\boldsymbol{\beta}}_U\|_{\Sigma_y}$	$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_{tr}\|_{\Sigma_\beta}$	DIC	mode $\hat{\eta}_{usa}$
Gaussian noise	(0)	279.38	575.55	520.63	40.06	129,433	1.09
$\epsilon_i \sim N(0, 1/\phi_{tr})$	(1)	279.82	439.32	389.22	35.83	128,882	0.20
$\phi_{tr} = 0.4$,	(2)	279.78	475.57	422.49	36.95	129,034	0.01
$\eta_{usa} = 0.18$,	(3)	279.96	455.81	404.30	36.23	128,986	0.22
$\psi = 10$	(4)	279.79	481.30	427.90	37.10	129,066	0.02
(b) $\boldsymbol{\beta} \sim \mathcal{N}_{p_{\text{usa}}}(\boldsymbol{\beta}_0, \frac{1}{\eta_{usa}} Q^{-1}(\psi))$ with an ellipsoidal neighborhood structure for Q							
Noises	Prior struct	$\|\boldsymbol{y} - X\hat{\boldsymbol{\beta}}\|_{\Sigma_y}$	$\|\boldsymbol{y} - X\hat{\boldsymbol{\beta}}_L\|_{\Sigma_y}$	$\|\boldsymbol{y} - X\hat{\boldsymbol{\beta}}_U\|_{\Sigma_y}$	$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_{tr}\|_{\Sigma_\beta}$	DIC	mode $\hat{\eta}_{usa}$
Gaussian noise	(0)	234.71	305.10	292.47	27.71	104,662	11.84
$\epsilon_i \sim N(0, 1/\phi_{tr})$	(1)	234.37	257.34	249.46	26.10	104,262	4.19
$\phi_{tr} = 0.4$,	(2)	234.27	251.55	244.20	24.59	104,152	0.30
$\eta_{usa} = 0.18$,	(3)	234.41	260.03	251.59	25.72	104,284	4.22
$\psi = 10$	(4)	234.30	253.50	245.76	24.50	104,173	0.53

Table 3.2: Posterior estimation results of the simulation study under Setup II, using synthetic earth models. The notations here are the same as in Table 3.1.

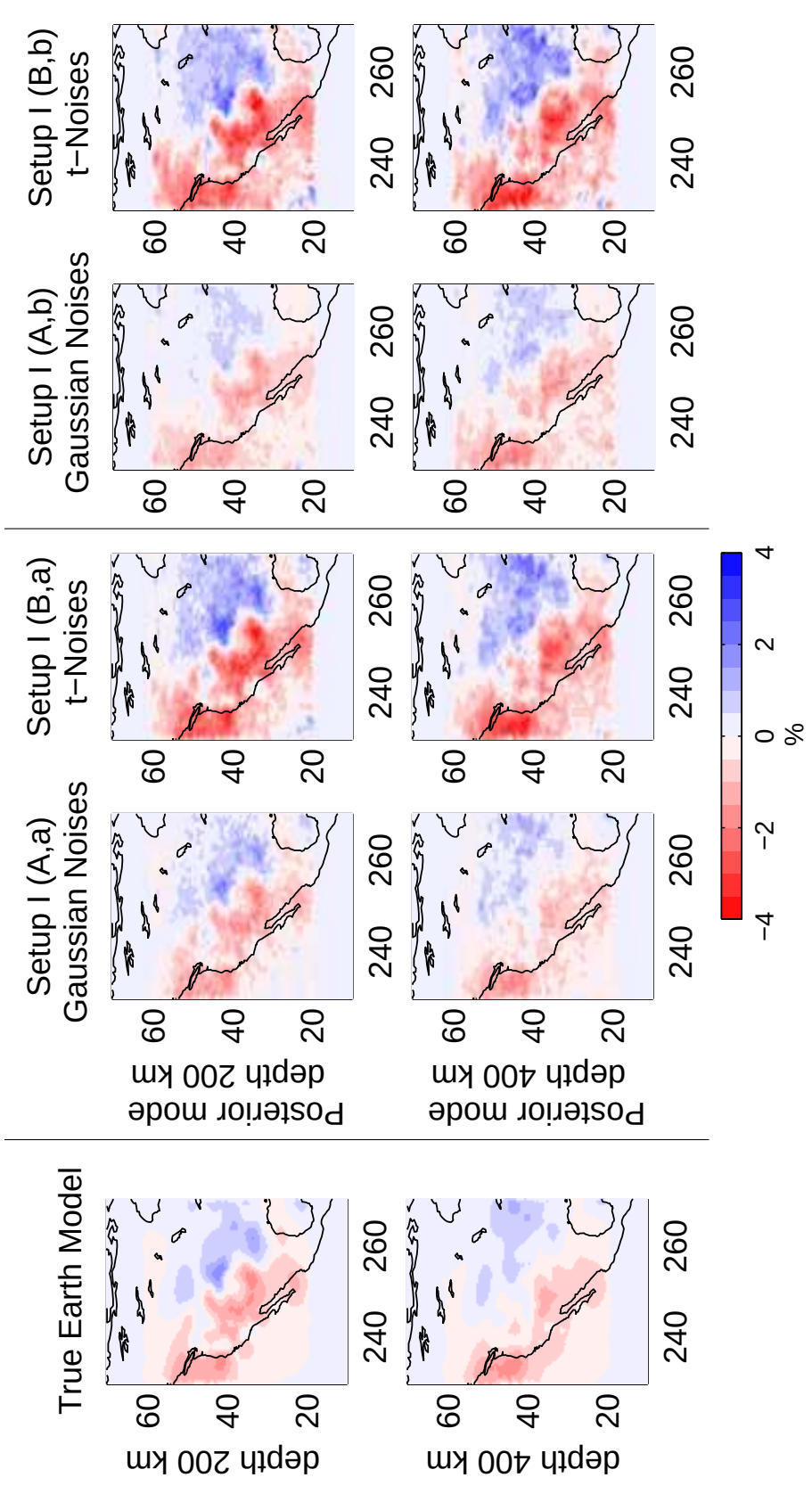

Figure 3.3: Mantle models resulting from the simulation study. Left column shows the "true" model, used to generate the synthetic data. The unit on the color bar is velocity deviation $\boldsymbol{\beta}_{\text{usa}}$ in % from the spherically symmetric reference model. All other columns show the posterior mode of velocity deviation $\boldsymbol{\beta}_{\text{usa}}$, estimated using ellipsoidal prior structure with reciprocal weights. Middle columns show results for Setup I(a), which uses the prior mean $\boldsymbol{\beta}_0 \neq \mathbf{0}$. Right columns show results for Setup I(b) assuming prior mean $\boldsymbol{\beta}_0 = \mathbf{0}$.

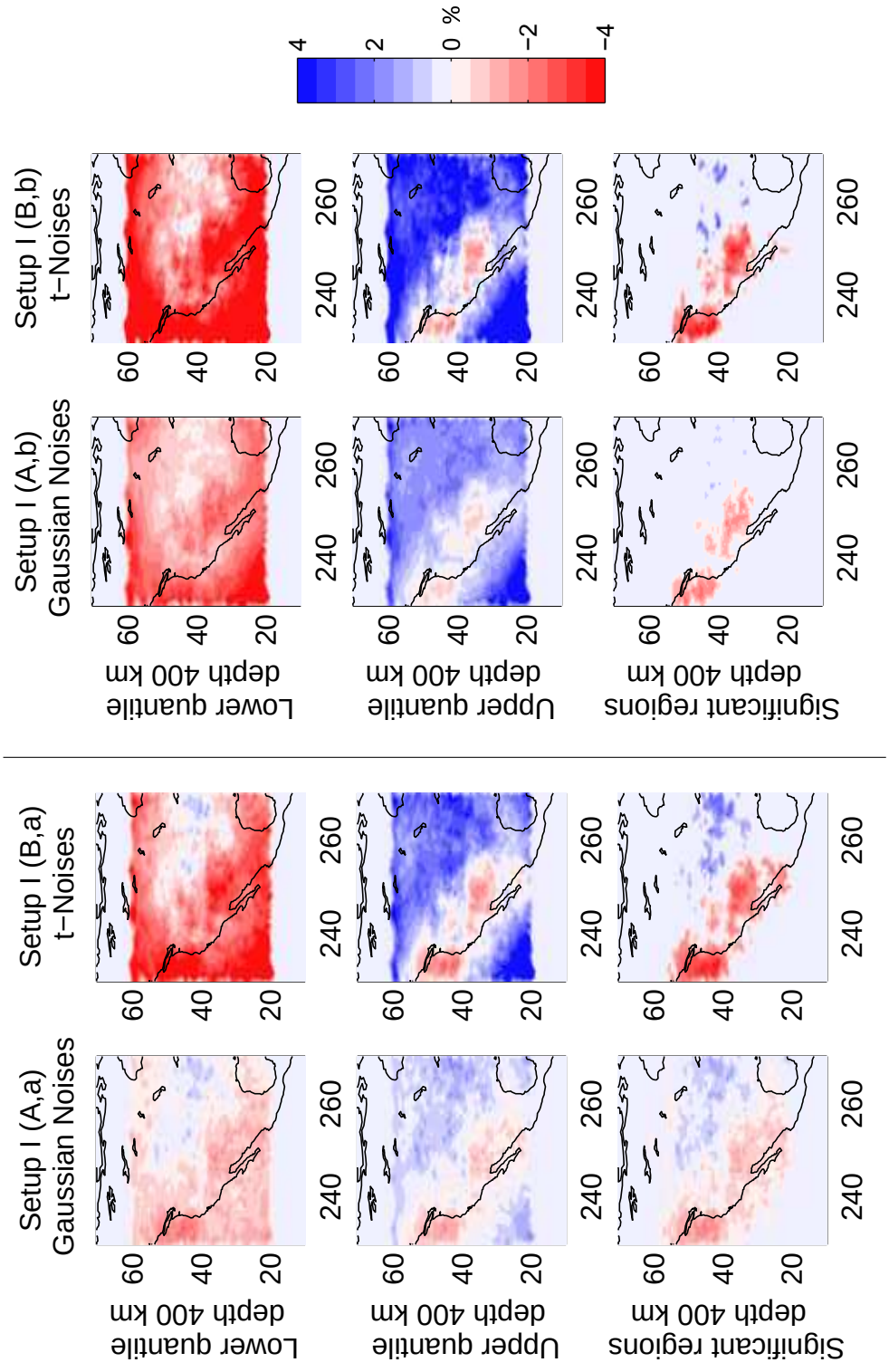

Figure 3.4: Continuation of Figure 3.3. The maps show velocity deviation in % from the reference Earth model. Left half shows the results under Setup I(a), which uses the prior mean $\boldsymbol{\beta}_0 \neq \mathbf{0}$; right half describes Setup I(b), which uses prior mean $\boldsymbol{\beta}_0 = \mathbf{0}$. First and second rows map out the lower and upper quantiles of the 90% confidence interval. Third row shows the posterior mode of velocity structure $\boldsymbol{\beta}$, but rendered only in regions that differ significantly from the reference model, according to the 90% confidence interval.

slower velocities. Physically, blue colors usually imply that those regions are colder than the default expectation for the corresponding mantle depth, while red regions are hotter than expected. In our simulation study, we assumed the true earth to be represented by the solution of Sigloch et al. (2008), shown in the left column of Figure 3.3. The middle and right columns of Figure 3.3 illustrate the estimated posterior modes $\hat{\boldsymbol{\beta}}_{\text{usa}}$ from Setup I with ellipsoidal neighborhood structure and reciprocal weight for both Gaussian and t-noises, respectively. They show that the parameter estimates from Gaussian noises are close to the true solution while the solution from the t-noises tends to overestimate the parameters. The magnitude of mantle anomalies is overestimated but major structures are correctly recovered. The same effect can be seen in the last column of Figure 3.3 which displays the estimated posterior modes of the tomographic solutions in Setup I (b). We have overestimation since the noise is not adequate to the Gaussian model assumption. Moreover, we also observe that tomographic solutions with the prior mean $\boldsymbol{\beta}_0 \neq \mathbf{0}$ are smoother than the ones with the prior mean $\boldsymbol{\beta}_0 = \mathbf{0}$.

Figure 3.4 shows estimated credible intervals for the solutions of Figure 3.3. Credible intervals for solutions with t-noises are larger than those for the Gaussian noises, as indicated by the darker shades of blue/red colors, which denote higher/lower quantile estimates. This implies that parameter uncertainty is greater if noise does not fit the model assumption. The same effect can be seen for results with the prior mean $\boldsymbol{\beta}_0 = \mathbf{0}$. The bottom row of Figure 3.4 maps out how the regions differ from the reference model with 90% posterior probability. For model-conform Gaussian distributed noises and informative prior mean, more regions differ from the reference model with 90% posterior probability, than if we added t-noise or used the less informative prior. In the case of an informative prior and/or correctly modeled noise, we achieve more certainty about the velocity deviations from the reference earth model.

3.5 Application to real seismic travel time data

In this section we apply our MCMC approach to actually measured travel time data. The measurements are a subset of those generated by Sigloch et al. (2008). We use the same wave paths, but only measurements made on the broadband waveforms, whereas they further bandpassed the data for finite-frequency measurements, and also included amplitude data (Sigloch and Nolet, 2006). Most stations are located in the western U.S., as part of the largest-ever seismological experiment (USArray), which is still in the process

of rolling across the continent from west to east. Numerous tomographic studies have incorporated USArray data – the ones most similar to ours are Burdick et al. (2008), Sigloch et al. (2008), Tian et al. (2009), and Schmandt and Humphreys (2010). All prior studies obtained their solutions through least-squares minimization, which yields no uncertainty estimates. Here we use 53,270 broadband travel time observations to estimate velocity structure under western North America (over 11,000 parameters), plus source corrections for 529 events (2116 parameters). We conduct our Bayesian inversion following two different scenarios:

Model 1: We only invert for earth structural parameters using data $\boldsymbol{y}_{\text{usa}}$ as stated in (3.1). For the velocity parameters we assume $\boldsymbol{\beta}_{\text{usa}} \sim \mathcal{N}_{p_{\text{usa}}}(\hat{\boldsymbol{\beta}}_{\text{usa}}^{\text{LSQR}}, \frac{1}{\eta_{\text{usa}}} Q^{-1}(\psi))$ as in (3.1) with $\psi \sim \mathcal{N}(10, 0.5^2)\mathbb{1}(\psi > 0)$, $\phi \sim \Gamma(1, 0.1)$ and $\eta_{\text{usa}} \sim \Gamma(10, 2)$.
Model 2: We invert for both earth structural parameters and the source corrections using data $\boldsymbol{y}_{\text{cr}}$ as given in (3.2). The prior distributions are set to $\boldsymbol{\beta} \sim \mathcal{N}_p(\hat{\boldsymbol{\beta}}^{\text{LSQR}}, \Sigma_\beta)$ as in (3.2) and Σ_β as defined in (3.5). For ψ, ϕ and η_{usa}, we adopt the same distribution as in Model 1. For the parameters of the source corrections we adopt $\eta_{\text{hyp}} \sim \Gamma(1, 5)$ and $\eta_{\text{time}} \sim \Gamma(10, 2)$.

We use the same five prior structures (0)-(4) as in the simulation study and run the MCMC algorithm for 10,000 iterations. The high-dimensional $\boldsymbol{\beta}$ vector can be sampled efficiently in terms of ESS with low burn-in and thinning rates thanks to the efficient Gibbs sampling scheme in (3.6). However, the hyperparameters, for example, ψ_β, are more difficult to sample. To achieve a good mixing we applied a burn-in of 200 and a thinning rate of 25 (393 samples for each parameter) in our analysis. On average, the effective sample size ESS values for $\boldsymbol{\beta}_{\text{usa}}$, $\boldsymbol{\beta}_{\text{hyp}}$ and $\boldsymbol{\beta}_{\text{time}}$ are about 393, 393 and 327, respectively, which indicate very low autocorrelations for most of the parameters. The ESS of both η_{usa} and ψ_β is about 103, while both η_{hyp} and ϕ have good mixing characteristics with ESS values equal to the sample size, and η_{time} has ESS value equal to 165. Figure 3.5 shows as examples the parameters $\beta_{\text{usa},955}$ at node 955, η_{usa} and ψ_β. The computing cost of our algorithm is about $O(n^4)$. Sampling Model 1 with about 9000 parameters, our algorithm needs 12 hours in 10,000 runs on a 32-core cluster, while under the same condition it needs 38 hours for Model 2.

Table 3.3 shows the results from Model 1 (estimation of earth structure) and Model 2 (earth structure plus source corrections). For both models, results from the independent prior structures, corresponding to the Bayesian ridge estimator, provide the best fit according to the DIC criterion. We also run the Model 1 with prior mean $\boldsymbol{\beta}_0 = \mathbf{0}$ (the

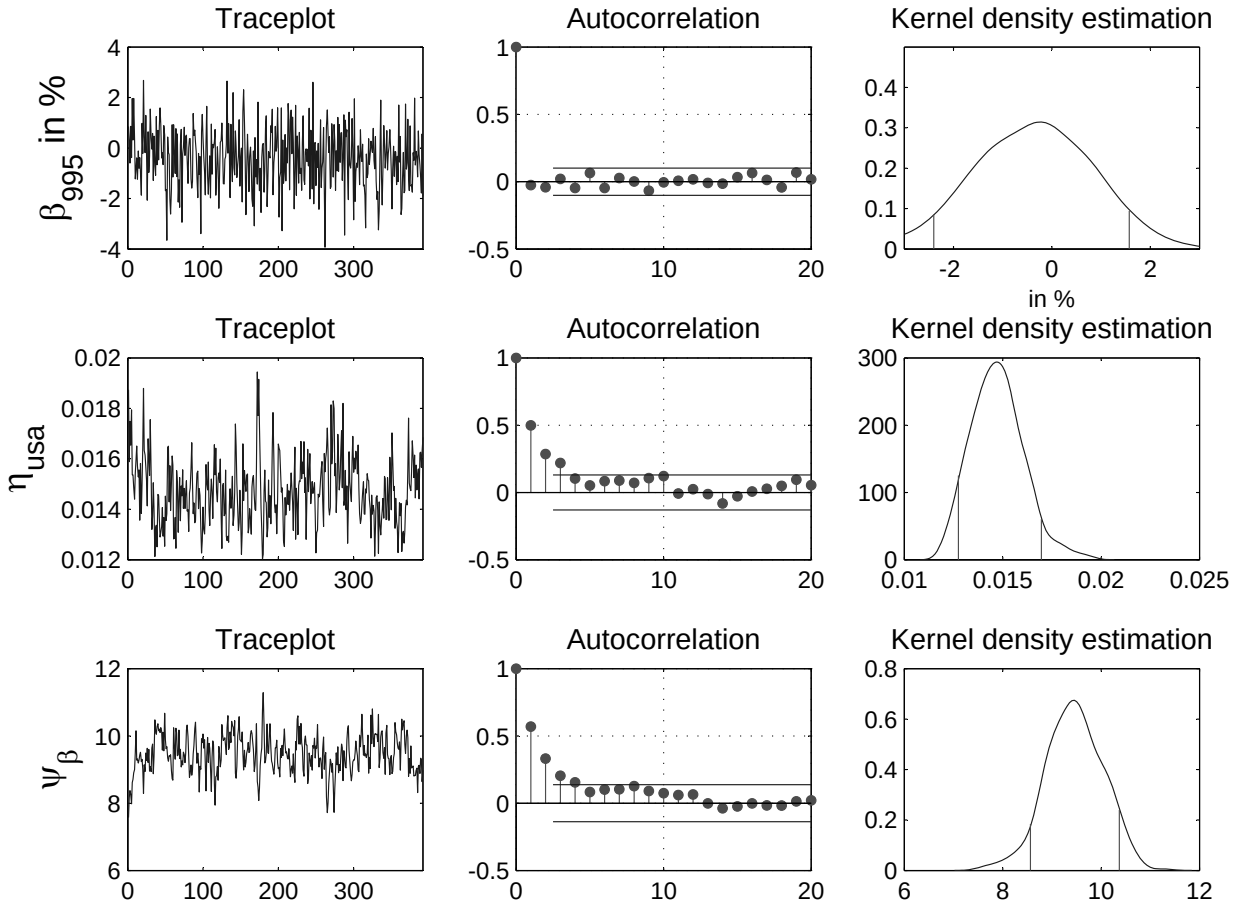

Figure 3.5: Convergence diagnostics: trace plot, autocorrelation and kernel density estimation of the parameters $\boldsymbol{\beta}_{\text{usa}}$ at node 955, η_{usa} and ψ_{β}. For 10,000 MCMC iterations the samples shown in plots are based on a burn-in of 200 and a thinning rate of 25.

Model 1									
Prior struct	$\|\boldsymbol{y} - X\hat{\boldsymbol{\beta}}\|_{\Sigma_y}$	$\|\boldsymbol{y} - X\hat{\boldsymbol{\beta}}_L\|_{\Sigma_y}$	$\|\boldsymbol{y} - X\hat{\boldsymbol{\beta}}_U\|_{\Sigma_y}$	DIC	mode $\hat{\phi}$	mode $\hat{\eta}_{\text{usa}}$	mode $\hat{\psi}$		
(0)	228.46	490.92	490.46	102,928	0.40	1.40	–		
(1)	229.14	389.73	390.21	104,096	0.39	0.20	9.63		
(2)	228.72	464.73	465.67	103,466	0.40	0.01	9.98		
(3)	228.90	430.78	431.34	103,749	0.39	0.17	9.63		
(4)	228.74	471.50	472.37	103,408	0.40	0.01	9.98		
Model 2									
Prior struct	$\|\boldsymbol{y} - X\hat{\boldsymbol{\beta}}\|_{\Sigma_y}$	$\|\boldsymbol{y} - X\hat{\boldsymbol{\beta}}_L\|_{\Sigma_y}$	$\|\boldsymbol{y} - X\hat{\boldsymbol{\beta}}_U\|_{\Sigma_y}$	DIC	mode $\hat{\phi}$	mode $\hat{\eta}_{\text{usa}}$	mode $\hat{\psi}$	mode $\hat{\eta}_{\text{hyp}}$	mode $\hat{\eta}_{\text{time}}$
(0)	225.40	483.96	488.35	93,788	0.49	1.15	–	0.01	5.01
(1)	225.76	515.29	524.61	94,993	0.48	0.10	9.63	0.01	4.53
(2)	225.48	498.96	501.45	94,374	0.48	0.00	9.55	0.01	4.70
(3)	225.61	503.20	512.01	94,669	0.48	0.11	9.63	0.01	4.53
(4)	225.44	496.69	498.97	94,312	0.49	0.01	10.00	0.01	4.70

Table 3.3: Posterior estimation results for the inversion using real data, under Model 1 and Model 2 specifications. For both models, results from the independent prior structures, corresponding to the Bayesian ridge estimator, provide the best fit according to the DIC criterion.

spherically symmetric reference model) and different covariance structures (0)-(2). The DIC results for priors (0), (1) and (2) are 103,100, 103,700 and 103,370, respectively. Two reasons may explain the selection of prior (0): (1) The data has generally more correlation structure than the i.i.d Gaussian assumption, which can not be solely explained by the spatial prior structure of the $\boldsymbol{\beta}$-fields. However, in our simulation study where different prior structures and the corrected data error are applied (Table 1), the DIC was able to identify the correct models; (2) Since the data are noisy, fitting could be difficult without a shrinkage prior. The prior in (0) can be compared to shrinkage in the ridge regression, which is the limiting case of priors in (1) to (4). Priors in (1) to (4) do not shrink the solutions of $\boldsymbol{\beta}$-fields as much as prior (0). They better reflect the uncertainty since the prior covariances in (1)-(4) are larger than variances in prior (0) in regions that have no data (no neighboring nodes), and smaller in regions with lots of data (lots of neighboring nodes).

Furthermore, the standardized data misfit criteria $\|\cdot\|_{\Sigma_y}$ do not show much difference between models with different prior specifications. According to the estimated 90% credible interval, estimates using spherical prior structure show a smaller range of data misfit in Model 1, whereas in Model 2, the independence prior shows a better result. Since our method assumes i.i.d. Gaussian errors, the resulting residuals might not be optimally fitted as expected. With regard to computational time, the independent prior model has a definite advantage over other priors in both Model 1 and Model 2. The general advantage of our Bayesian method is that the independent model yields an estimate given as the ratio between the variance of the data and the variance of the priors corresponding to ridge estimates with *automatically chosen shrinkage* described in Section 3.3, whereas in Aster et al. (2005), Sigloch et al. (2008), Bodin et al. (2012) and all other prior work, the shrinkage parameter (strength of regularization) had to be chosen by the user a priori.

Figure 3.6 shows the estimated posterior and prior densities of parameters in Model 2, at four different locations of varying depth. We see that parameters at locations with good ray coverage, for example, node 5400 and node 3188, have smaller credible intervals than parameters at locations with no ray coverage, for example, node 5564 and node 995 beneath the uninstrumented oceans. Geologically, the regions between node 5400 and node 3188 are well known to represent the hot upper mantle, where seismic waves travel slower than the reference velocity. This is consistent with our results in Figure 3.6: the fact that $\boldsymbol{\beta} = \mathbf{0}$ does not fall inside the 90% credible intervals indicates a velocity deviation from the spherically symmetric reference model with high posterior probability. Figure 3.6 shows that the posterior is more diffuse than the prior. As mentioned in Section 3.1, the

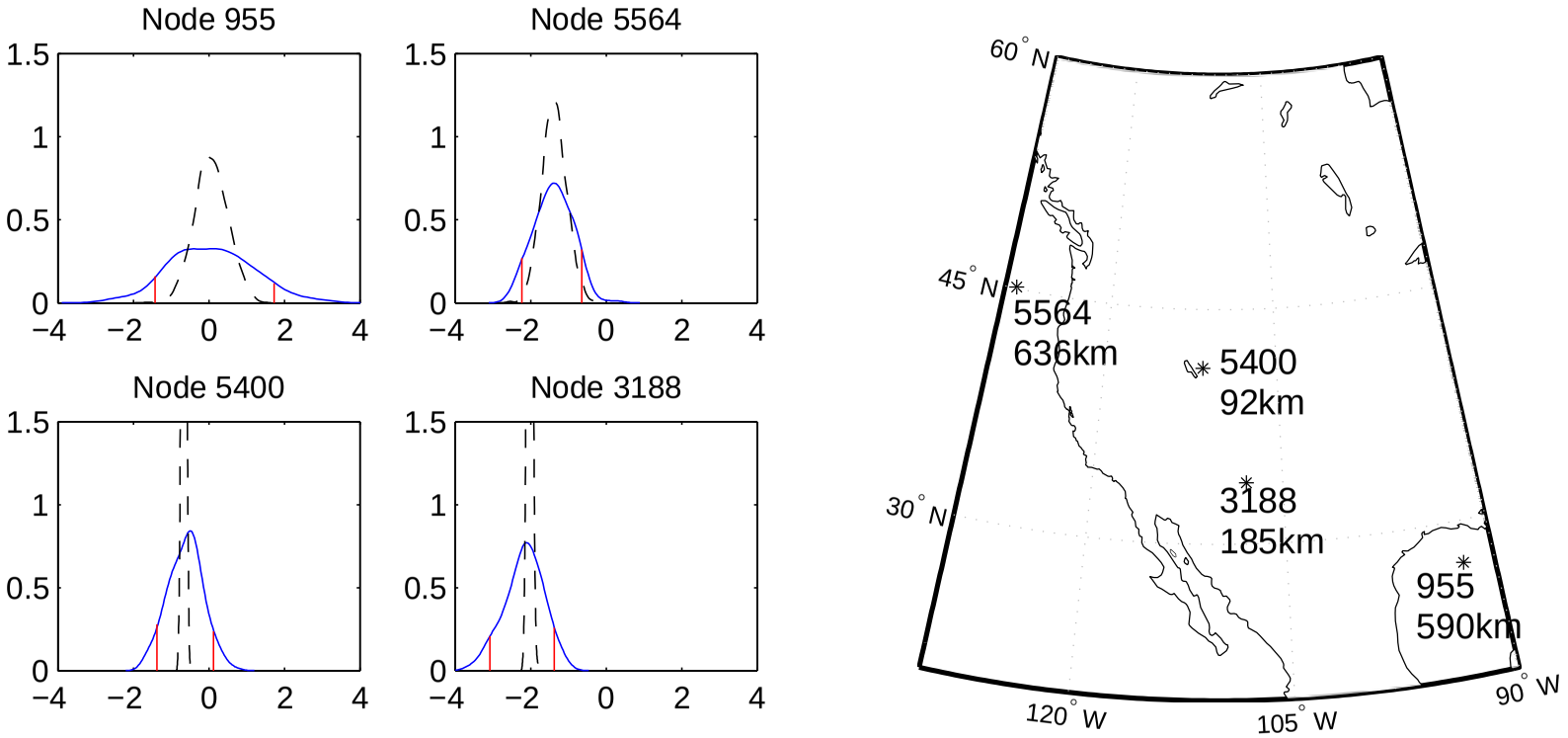

Figure 3.6: Results of the Bayesian tomography using real travel time observations. Left: Estimated posterior density of $\boldsymbol{\beta}_{\text{usa}}$ at a few selected model nodes, whose locations and depths are indicated on the map. Unit on the x-axes is velocity deviation in %. Dashed lines: prior density, the prior variance can be very small if number of neighbors is large. Solid lines: posterior density with 90% credible intervals.

spatial prior for $\boldsymbol{\beta}$ depends on distance of neighboring nodes, number of neighbors and orientation. The variance can be very small if the number of neighbors is very large, as shown in Figure 3.2. Incorporating data, the information about $\boldsymbol{\beta}$ is updated and thus may yield more diffuse posteriors than the priors, as we see here. The left half of Figure 3.7 shows the estimated posterior modes of mantle structure obtained by Model 2, for independent and for ellipsoidal priors with reciprocal weights. The right half of Figure 3.7 extracts only those regions that differ from the reference model according to the 90% credible interval. The ellipsoidal prior results in higher certainty of velocity deviations at a depth of 200 km, compared to the independence prior. At a depth of 400 km, the credible regions resemble each other more strongly. This confirms geological arguments that deeper regions of the mantle are more homogeneous and do not differ as much from the spherically symmetric reference model as shallower regions.

Many lines of geoscientific investigations provide independent confirmation of the significantly anomalous regions of Figure 3.7. The red areas map out the hot upper mantle under the volcanic, extensional Basin and Range province and Yellowstone; the blue anomalies map out the western edge of the old and cool North American craton.

The overall comparison of our solutions to earlier least-squares inversions, for example, the model by Sigloch (2008) shown in the left column of Figure 3.3, confirms that Bayesian inversion successfully retrieves the major features of mantle structure. The images are similar, but the major advantage and novelty of our approach is that it also quantifies uncertainties in the solution (which we have chosen to visualize as credible intervals here).

3.6 Discussion and outlook

Uncertainty quantification in underdetermined, large inverse problems is important, since a single solution is not sufficient for making conclusive judgements. Two central difficulties for MCMC methods have always been the dimensionality of the problem (number of parameters to sample) or the evaluation of the complex physical forward model (nonlinear problems) in each MCMC iteration (Tarantola, 2004; Bui-Thanh et al., 2011; Martin et al., 2012).

Consider the model $\boldsymbol{Y} = f(\boldsymbol{\beta}) + \boldsymbol{\epsilon}$ with the physical forward model $f(\cdot)$, high-dimensional parameter $\boldsymbol{\beta}$ and error $\boldsymbol{\epsilon}$. In general, if the physical problem is linear ($f(\boldsymbol{\beta}) = X\boldsymbol{\beta}$) and the full conditional of $\boldsymbol{\beta}$ is Gaussian, efficient sampling from the high-dimensional Gaussian conditional distribution is essential for the exploration of model space. In this

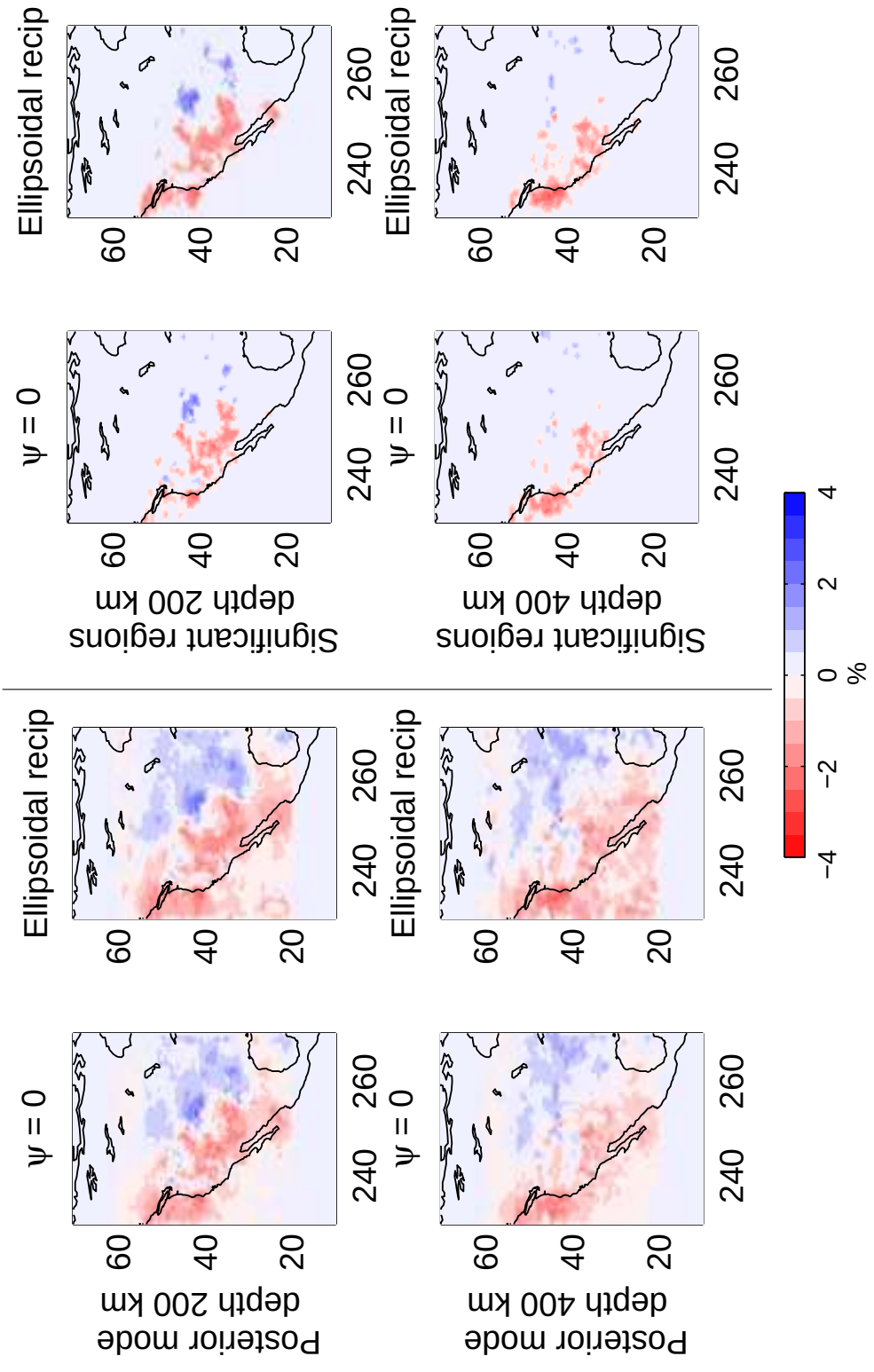

Figure 3.7: Results of the Bayesian tomography using real travel time data. All maps show estimated velocity deviation from the reference earth model IASP91 (in %) Left columns: Estimated posterior mode of velocity deviation, for the scenario of Model 2. Right columns: Same posterior mode, but rendered are only regions that differ from the reference model with 90% posterior probability.

case the error $\boldsymbol{\epsilon}$ need not necessarily be Gaussian, but may be t or skewed-t distributed (Sahu et al., 2003; Frühwirth-Schnatter and Pyne, 2010), or a Gaussian error with a spatial correlation such as considered in Banerjee et al. (2003). Given a sparse posterior precision matrix (e.g. (3.6)) efficient sampling from a multivariate normal can be carried out by Cholesky decomposition of a permuted precision matrix as discussed in Wilkinson and Yeung (2002) or Rue and Held (2005), by using an approximate minimum-degree ordering algorithm. A further improvement to the current sampling approach might be to apply the Krylov subspace method from Simpson et al. (2008). This would require substantial implementation efforts and is the subject of further research. If the forward matrix or the prior precision matrix is not sparse, a dense posterior precision matrix for $\boldsymbol{\beta}$ will result. In this case our sampling scheme is inefficient, but the model-space reduction method developed by Flath et al. (2011) might be used instead. They exploit the low-rank structure of the preconditioned Hessian matrix of the data misfit, involving eigenvalue calculations. However, this approximation quantifies uncertainty of large-scale linear inverse problems only for known hyperparameters, thus ignoring uncertainty in those parameters. Eigenvalue calculation in each MCMC step can be time consuming and prohibitive for hierarchical models with unknown hyperparameters when the posterior covariance matrix in every MCMC step changes. Here additional research is needed.

If the full conditionals cannot be written as Gaussian (this case includes the cases of a nonlinear $f(\cdot)$, a non-Gaussian prior of $\boldsymbol{\beta}$ or non-Gaussian, non-elliptical distributed errors), using the standard MH algorithm to sample from the high-dimensional posterior distribution is often computationally infeasible. Constructing proposal density that provides a good approximation of the stationary distribution while keeping the high-dimensional forward model $f(\cdot)$ inexpensive to evaluate has been the focus of the research over the past years: Lieberman et al. (2010) have drawn samples from an approximate posterior density on a reduced parameter space using a projection-based reduced-order model. In the adaptive rejection sampling technique by Cui et al. (2011), the exact posterior density is evaluated only if its approximation is accepted. The stochastic Newton approach proposed by Martin et al. (2012) approximates the posterior density by local Hessian information, thus resulting in an improvement of the Langevin MCMC by Stramer and Tweedie (1999). Other random-walk-free, optimization-based MCMC techniques for improving the proposal and reducing correlation between parameters have been developed, such as Hamiltonian Monte Carlo (HMC) (Neal, 2010), Adaptive Monte Carlo (AM) (Haario et al., 2001; Andrieu and Thoms, 2008) and several variations, for example, delay rejection AM (DRAM) (Haario et al., 2006), differential evolution MC

(DEMC) (Ter Braak, 2006), differential evolution adaptive Metropolis (DREAM) (Vrugt et al., 2009), just to mention a few. However, MCMC sampling of high-dimensional problems still requires a massive amount of computing time and resources. For example, the quasi three-dimensional nonlinear model of Herbei et al. (2008) contains about 9000 parameters on a 37x19 grid. We expect a long computing time since they use standard MCMC sampling methods. The example by Cui et al. (2011) shows that their algorithm achieves a significant improvement in both computing time and efficiency of parameter space sampling for a large nonlinear system of PDEs that includes about 10,000 parameters. However, their algorithm gives 11,200 iterations in about 40 days, while our problem requires only 38 hours (on a 32-core cluster) for the same number of iterations for about 11,000 parameters.

While the future may be in effective uncertainty quantification of nonlinear physical problems using model reduction and optimization techniques, the computing time and resources at the moment are too demanding to explore the large model space. This thesis demonstrates effective Bayesian analysis tailored to a realistically large seismic tomographic problem, featuring over 11,000 structural and source parameters. We deliver a precise uncertainty quantification of tomographic models in terms of posterior distribution and credible intervals using the MCMC samples, which allows us to detect regions that differ from the reference earth model with high posterior probability. Our approach is the first to solve seismic tomographic problems in such high dimensions on a fine grid, and thus provides ground work in this important research area.

Chapter 4

Bayesian spatial modeling for high-dimensional seismic inverse problems using the SPDE and INLA approaches

4.1 Introduction

In this chapter, based on Zhang et al. (2013b), we study high-dimensional linearized inverse problems in seismic tomography by means of Bayesian spatial modeling. The same as in Chapter 3, our goal is to estimate volumetric anomalies in seismic wave velocities, i.e., the shapes and magnitudes of regions in the earth's interior where seismic waves travel faster or slower than predicted by a simple, layered earth reference model. The resulting, 3-D earth model represents the primary output of seismic tomography. In practice things are complicated by the fact that traveltime (or other) measurements are also influenced by near-source and near-receiver effects not accounted for by 3-D mantle anomalies. Earthquake characteristics are unknown a priori. We extract information about their timing and locations from routinely assembled catalogues, but such catalogue estimates are usually not accurate enough for our purposes. Earthquake mislocations cause traveltime anomalies, which if unchecked would propagate into the solution for 3-D mantle structure. Hence tomographers commonly introduce "source corrections", typically four free parameters per earthquake (latitude, longitude, depth, and time), which are allowed to adjust optimally during inversion. In this chapter we take a different approach and introduce a 2-D random

field of traveltime anomalies, the "source field", to estimate traveltime anomalies due to source mislocations. We introduce a second 2-D random field, the "receiver field" to correct for traveltime anomalies near each receiver. Note that for simplicity we ignore the depth of the sources and the altitude of the receivers and assume they are on the surface of the earth here. These systematic anomalies are mostly caused by poorly known crustal structure beneath the seismic station, although misfunctioning of the station (e.g., clock errors) might also contribute. In contrast to conventional corrections, which are associated with discrete source and receiver locations, our source and receiver fields are continuous over (parts of) the earth's surface.

To model the earth's three-dimensional velocity structure as well as the source and receiver fields, we develop and implement a spatial dependency model based on a Gaussian Matérn field approximation using the theory of stochastic partial differential equations (SPDEs) from Lindgren et al. (2011). This allows for modeling spatial dependency of the data errors caused by sources and receivers individually or jointly. Subsequently we carry out the uncertainty quantification of the high-dimensional parameter space by using the integrated nested Laplace approximation (INLA) (Rue et al., 2009). We provide an application using seismological data from the continental-scale USArray experiment. Thereby, we reveal major structures of the mantle beneath the western USA with uncertainty assessments, and provide additionally correlation estimates between the parameters.

4.2 Setup of the statistical spatial model

An n-sphere of radius r centered at the origin, denoted by $\mathbb{S}^n$, is embedded in the Euclidean space $\mathbb{R}^{n+1}$ and defined by $\mathbb{S}^n := \{\boldsymbol{x} \in \mathbb{R}^{n+1} : \|\boldsymbol{x}\| = r\}$. Our target region of estimation is the mantle beneath North America, a sub-mesh of the global 3-D domain ($\mathbb{R}^3$), which contains 8977 free velocity parameters located between latitudes $20° - 60°N$, longitudes $90° - 130°W$, and $0 - 800$ km depth. Tetrahedral nodes of the sub-mesh are spaced by $60 - 150$ km. From here on, we call this subset of velocity parameters "velocity field" and denote it as $\boldsymbol{\beta}_{\text{usa}}$ (as in Chapter 3). To model the spatial dependency of seismic traveltime delay data errors, we consider the following fields. For the data from the sources (the source field), we have a mesh covering the entire earth on a curved space in $\mathbb{S}^2$. For the data from the receivers (the receiver field), the mesh is generated over the North America region again in $\mathbb{S}^2$ space with vertices close to the locations of the sources and receivers as shown in Figure 4.1.

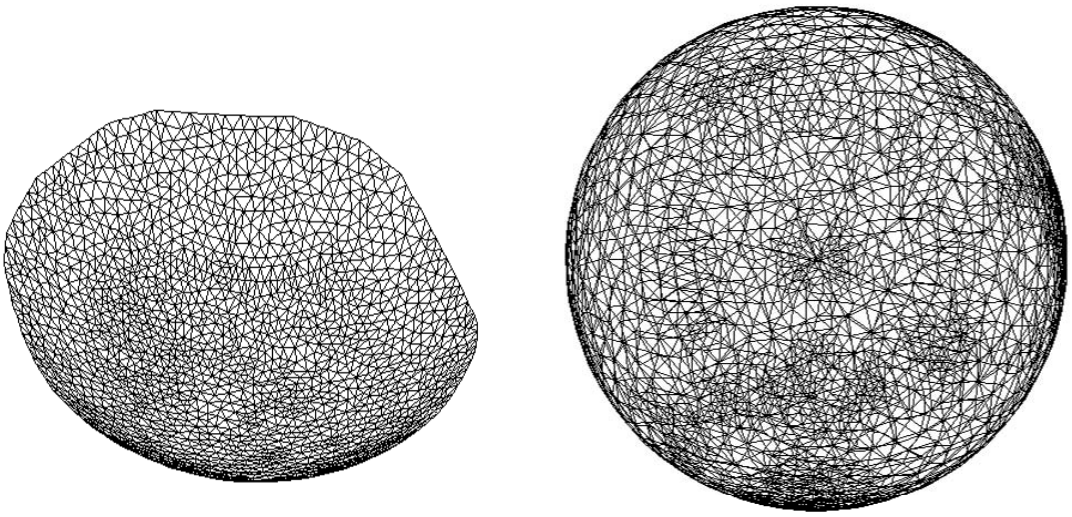

Figure 4.1: Triangular meshes on a $\mathbb{S}^2$ space with vertices located close to the locations of sources (right) and receivers (left). Left: receivers were deployed mainly in western North America, the dark region of dense nodes. Right: earthquakes (sources) occur over the entire globe.

Let y_i denote the ith traveltime delay datum for $i = 1, ..., N$ as described in Chapter 2. Further let $r(i)$ be the receiver that recorded the ith observation y_i, $r = 1, ..., R$. Similarly, let $s(i)$ be the source that caused the ith observation, $s = 1, ..., S$. Let p_s and p_r denote the numbers of mesh vertices of the source and receiver fields on $\mathbb{S}^2$, respectively. Define the mapping matrices between data and mesh vertices as

$$B_s := Z^s A^s \in \mathbb{R}^{N \times p_s} \text{ and } B_r := Z^r A^r \in \mathbb{R}^{N \times p_r}.$$

Here, the matrices $A^s \in \mathbb{R}^{S \times p_s}$ and $A^r \in \mathbb{R}^{R \times p_r}$ map from the vertex locations to the positions of the sources and receivers, respectively. The matrices $Z^s \in \mathbb{R}^{N \times S}$ and $Z^r \in \mathbb{R}^{N \times R}$ are incident matrices defined as

$$Z^s_{is} := \begin{cases} 1 & : \quad s = s(i) \\ 0 & : \quad \text{otherwise.} \end{cases} \quad \text{and} \quad Z^r_{ir} := \begin{cases} 1 & : \quad r = r(i) \\ 0 & : \quad \text{otherwise.} \end{cases} \tag{4.1}$$

Our data are ordered by the stations. To illustrate this consider Figure 4.2 as a toy example, where the data $\boldsymbol{y}$ arise from (source, receiver): (1,1), (1,3), (2,1), (2,2), (3,2), (4,1) and (4,3), respectively. The corresponding Z^s and Z^r for the 7 observations are

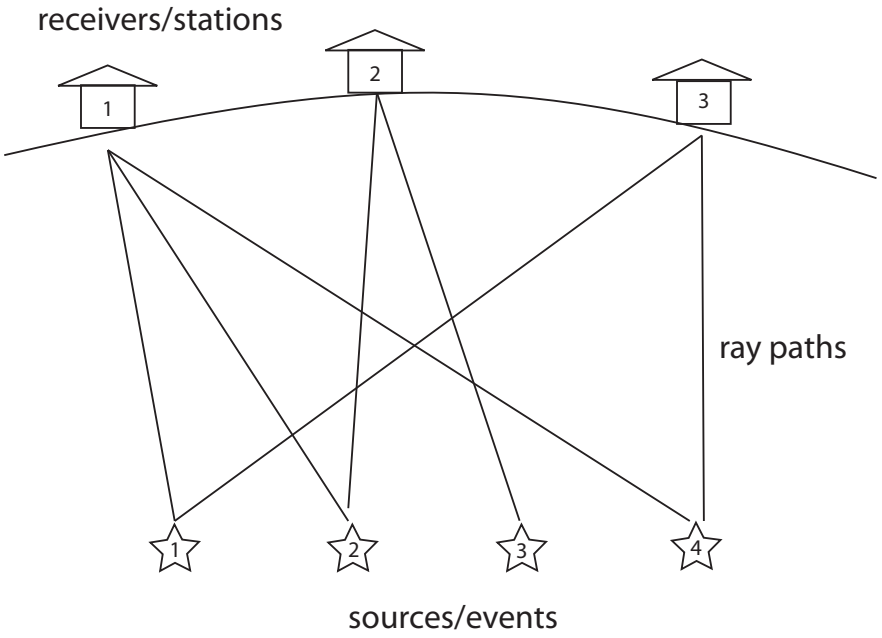

Figure 4.2: One traveltime observation caused by a large earthquake can be recorded by several stations and each station can record earthquakes occur at different locations.

therefore specified as

$$Z^s = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad Z^r = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Let $\boldsymbol{\gamma} \in \mathbb{R}^{p_s}$ and $\boldsymbol{\xi} \in \mathbb{R}^{p_r}$ be the Gaussian processes on those mesh vertices including the sources and receivers, where p_s and p_r are the number of vertices on the meshes for the sources and receivers. Let p_{usa} be number of parameters of the velocity field $\boldsymbol{\beta}_{\text{usa}}$. The

full statistical model defined on *mesh vertices* of receivers and sources can be written as

$$Y_i = \sum_{j=1}^{p_{\text{usa}}} x_{ij}^t \beta_j + \sum_{k=1}^{p_s} B_{s,ik}\gamma_k + \sum_{l=1}^{p_r} B_{r,il}\xi_l + \epsilon_i$$

or in matrix form:

$$\textbf{Full Model (VSR):} \quad \boldsymbol{Y}_{\text{usa}} = X_{\text{usa}}\boldsymbol{\beta}_{\text{usa}} + B_s\boldsymbol{\gamma} + B_r\boldsymbol{\xi} + \boldsymbol{\epsilon}. \tag{4.2}$$

The term "VSR" stands for the combination of velocity, source and receiver fields. Note that the term x_{ij} is the sensitivity kernel of the ith wavepath at node j as given in (2.3). The matrix X_{usa} is the sensitivity kernel for the velocity field $\boldsymbol{\beta}_{\text{usa}}$ as given in Chapter 2 and the same as the matrix X_{usa} in (3.1) in Chapter 3. Since the parameters are linear we can gather all parameters and covariates together into

$$\boldsymbol{Y}_{\text{usa}} = H\boldsymbol{m} + \boldsymbol{\epsilon}, \tag{4.3}$$

with $H := [X_{\text{usa}}, B_s, B_r] \in \mathbb{R}^{N\times D}$ and the random vector $\boldsymbol{m} := [\boldsymbol{\beta}'_{\text{usa}}, \boldsymbol{\gamma}', \boldsymbol{\xi}']' \in \mathbb{R}^D$ with $D = p_{\text{usa}} + p_s + p_r$.

Source field $\boldsymbol{\gamma}$ and receiver field $\boldsymbol{\xi}$ may either be used to estimate the traveltime effects of mislocated earthquakes and uncertain near-receiver structure from scratch, or to assess the quality of source and receiver corrections from earlier, deterministic solutions (including their spatial correlations). We take the first approach in our synthetic study of Section 4.4, which tests the proper functioning of our method, and the latter approach in our study of real data (Section 4.5). There we evaluate and improve on the source and receiver corrections obtained by Sigloch (2008), which allows for direct comparison to the data study by Zhang et al. (2013a) (see Chapter 3), who also used the corrections of Sigloch (2008).

Thus, we have constructed a statistical model that takes into account and estimates the seismic velocity field inside the earth, jointly with traveltime corrections both at the earthquake sources (source field $\boldsymbol{\gamma}$) and at the seismic stations (receiver field $\boldsymbol{\xi}$) on the earth's surface. In the following Section we present the estimation methods developed for this statistical model. We adopt the stochastic partial differential equation formulation from Lindgren et al. (2011) for the approximation of the Gaussian random field to specify our structural parameters $\boldsymbol{\beta}_{\text{usa}}$ as well as the Gaussian process parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\xi}$ on tetrahedral meshes in the $\mathbb{R}^3$ and $\mathbb{S}^2$ spaces, respectively.

4.3 Estimation methods

4.3.1 An approximation to Gaussian fields with continuous Markovian structures on manifolds

Suppose $U \subseteq \mathbb{R}^d$, a subset of $\mathbb{R}^d$. For $d = 2$, $\boldsymbol{u} := (u_x, u_y) \in U$ is characterized by fixed geographic coordinates, and if $d = 3$, we have $\boldsymbol{u} := (u_x, u_y, u_z) \in U$ with depth u_z. Let $\{Z(\boldsymbol{u}) : \boldsymbol{u} \in U \text{ and } U \subseteq \mathbb{R}^d\}$ a stochastic process or a spatial process for $d > 1$ (Banerjee et al., 2003). In practice, the observations will be a partial realization of that spatial process measured at a finite set of locations $\{\boldsymbol{u}_1, ..., \boldsymbol{u}_p\}$. The process is said to be Gaussian if, for any $p \geqslant 1$ and any set of sites $\{\boldsymbol{u}_1, ..., \boldsymbol{u}_p\}$, $\boldsymbol{Z} := (Z(\boldsymbol{u}_1), ..., Z(\boldsymbol{u}_p))'$ has a multivariate normal distribution (Banerjee et al., 2003). A spatial process $\boldsymbol{Z}$ is *weakly stationary* if the process has a constant mean, and $Cov(Z(\mathbf{0}), Z(\boldsymbol{u})) = c(\boldsymbol{u})$ for all $\boldsymbol{u} \in U$ (Cressie, 1993). If the covariance function $c(\cdot)$ only depends on the length $\|\boldsymbol{u}\|$ of the vector $\boldsymbol{u}$, then we say that the process is *isotropic*. In this chapter our spatial processes are set to be weakly stationary and isotropic with mean zero. We refer to the Gaussian distributed spatial process as Gaussian random field or Gaussian field (GF). There are a number of parametric covariance functions in the literature. Here, we consider a particular class of the GFs, called the Matérn random field (MF) (Note that we will later only consider a special case of the MF). Traditionally, the Matérn random fields are zero mean Gaussian stationary, isotropic random fields. The covariance function at location $\boldsymbol{u} \in \mathbb{R}^d$ is given by

$$Cov(Z(\mathbf{0}), Z(\boldsymbol{u})) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} (\kappa \|\boldsymbol{u}\|)^\nu K_\nu(\kappa \|\boldsymbol{u}\|). \tag{4.4}$$

(Cressie, 1993; Stein, 1999). The function $K_\nu(\cdot)$ is the modified Bessel function of the second kind with the smoothing parameter $\nu > 0$ and $\Gamma(\cdot)$ denotes the Gamma function. ν controls the smoothness of the realized random field. For all $\boldsymbol{u}$ the marginal variance $\sigma^2 := Var[Z(\boldsymbol{u})] > 0$ is defined as

$$\sigma^2 := \frac{\Gamma(\nu)}{\tau^2 \Gamma(\nu + d/2)(4\pi)^{\frac{d}{2}} \kappa^{2\nu}}. \tag{4.5}$$

Here, τ is with a scaling parameter proportional to $1/\sigma^2$ if other parameters are fixed. Furthermore, we have $\nu > 0$ and d is the dimensionality of the random field which is

either $d = 2$ for source/receiver fields or $d = 3$ for the velocity field. The parameter κ controls the degree of spatial dependency of the random field. For more straightforward interpretability, κ can be also specified through a function which depends on the range parameter ρ, i.e.

$$\kappa := \frac{\sqrt{8\nu}}{\rho},$$

which is derived empirically by Lindgren et al. (2011). The parameter ρ means that for all ν the spatial correlation is about 0.1 at a distance ρ (Lindgren et al., 2011). Given a fixed ν, the larger the value of ρ is, the stronger the spatial correlation. Since the smoothing parameter ν is difficult to estimate, and also due to the Gaussian field property which will be introduced in the later section, we will fix ν at $\nu + d/2 = 2$ in this chapter depending on the dimensionality d of the random field. Given d we can obtain the value of ν, and it results in a simplified formula for variance σ^2: For $d = 2$, the smoothness is fixed at $\nu = 1$ and the variance is

$$\sigma^2 := 1/(4\pi\kappa^2\tau^2),$$

whereas for $d = 3$, we have $\nu = 1/2$ and

$$\sigma^2 = 1/(8\pi\kappa\tau^2).$$

The latter case yields the exponential covariance function, which is given by

$$Cov(Z(\mathbf{0}), Z(\boldsymbol{u})) = \sigma^2 \exp(-\kappa\|\boldsymbol{u}\|).$$

Gaussian fields (GFs) are the most important model class in spatial statistics with well studied, interpretable statistical properties. However, if the dimension of the random vector $\boldsymbol{Z} := (Z(\boldsymbol{u}_1), ..., Z(\boldsymbol{u}_p))'$ is high, its covariance function yields a very large, dense covariance matrix, which is computationally not feasible for inference. On the other hand, there is a special case of Gaussian fields called the Gaussian Markov random fields (GMRFs) (Rue and Held, 2005; Simpson et al., 2012). A GMRF is a GF which satisfies the conditional independence assumptions, also called the Markovian property. The Markovian property is related to the definition of a neighborhood structure. This assumption says that the conditional distribution of a random variable, say Z_i, given all other variables Z_j, $j \neq i$, only depends on the neighbors. That is,

$$P(Z_i|\boldsymbol{Z}_{-i}) = P(Z_i|\boldsymbol{Z}_{\delta_i}) \text{ with } \boldsymbol{Z}_{-i} := (Z_1, ..., Z_{i-1}, Z_{i+1}, ..., Z_p)'$$

and more general $\boldsymbol{Z}_{-S} := Z_i, i \notin S$. The notation $\boldsymbol{Z}_{\delta_i}$ denotes the set of random variables in the neighborhood of Z_i. Following the notations of Rue and Held (2005) and Cameletti et al. (2012), the conditional independence condition can be written as

$$Z_i \perp \boldsymbol{Z}_{-i,\delta_i}, \mid \boldsymbol{Z}_{\delta_i}, \text{ for } i = 1, ..., p.$$

This condition is closely related the precision matrix Q, in particular,

$$Z_i \perp Z_j \mid \boldsymbol{Z}_{-i,j} \Leftrightarrow Q_{ij} = 0.$$

The nonzero elements of Q are given by the neighborhood structure of the random field. Therefore, the precision matrix Q of a GMRF is sparse which allows for fast computations. In classical spatial statistics, GMRF models are mainly used in modeling areal data. In order to benefit from the fast computational property of the GMRFs while still modeling the continuously indexed GFs, Lindgren et al. (2011) developed the link between GFs and GMRFs through a type of stochastic partial differential equations (SPDE). We briefly review their results.

Let $(Z_1, ..., Z_p) := (Z(\boldsymbol{u}_1), ..., Z(\boldsymbol{u}_p))'$ be multivariate Gaussian at random locations $\{\boldsymbol{u}_1, ..., \boldsymbol{u}_p\}$. Lindgren et al. (2011) proposed to use a mesh for parameterizing the GF (the same mesh as adopted for the physical model). They suggested a finite dimensional basis function expansion for a continuous specification of the GF $\boldsymbol{Z}$ given by,

$$Z(\boldsymbol{u}) \approx \sum_{j=1}^{p} \varphi_j(\boldsymbol{u}) Z_j, \quad \text{with } \varphi_j(\boldsymbol{u}_i) := \begin{cases} 1 & : \quad i = j \\ 0 & : \quad i \neq j \end{cases}. \tag{4.6}$$

The $\varphi_j(\boldsymbol{u}_i)$ are deterministic piecewise linear basis functions. Linear interpolation is applied to the basis function $\varphi_j(\boldsymbol{u})$ at points $\boldsymbol{u} \neq \boldsymbol{u}_i$. Lindgren et al. (2011) showed that a subset of Matérn fields gives a piecewise linear representation of the GF $Z(\boldsymbol{u})$. Define

$$\alpha := \nu + d/2.$$

In particular, when α in (4.4) is an integer, the Matérn field has the piecewise linear representation as in (4.6). In this case, the piecewise linear representation of $Z(\boldsymbol{u})$ in (4.6) is the stationary solution to the stochastic partial differential equation (SPDE),

$$(\kappa^2 - \Delta)^{\frac{\alpha}{2}} (\tau Z(\boldsymbol{u})) = \mathcal{W}(\boldsymbol{u}), \quad \alpha := \nu + d/2 \tag{4.7}$$

where Δ is the Laplacian defined as $\Delta := \sum_{i=1}^{d} \frac{\partial^2}{\partial u_i^2}$. Here, $\mathcal{W}(\boldsymbol{u})$ denotes a Gaussian white noise process. Whittle (1954, 1963) showed that Matérn fields $Z(\boldsymbol{u})$ on $\mathbb{R}^d$ (with the covariance in (4.4)) are the only stationary solutions to the SPDE in (4.7). Lindgren et al. (2011) further showed that for $\alpha = 2$ the random vector $\boldsymbol{Z} := (Z_1, ..., Z_p)'$ in (4.6) is a Gaussian Markov random field (GMRF) with mean zero, and its precision matrix Q is given by

$$Q := \tau^2(\kappa^4 \tilde{C} + 2\kappa^2 G + G\tilde{C}^{-1}G), \tag{4.8}$$

with elements of $\tilde{C}$ and G given by

$$\tilde{C}_{ii} := \int_U \varphi_i(\boldsymbol{u}) d\boldsymbol{u} \quad \text{and} \quad G_{ij} := \int_U \nabla\varphi_i(\boldsymbol{u})^t \nabla\varphi_j(\boldsymbol{u}) d\boldsymbol{u}, \text{ respectively.}$$

Here, $\tilde{C}$ is a diagonal matrix. We integrate over the observation domain U. For the velocity field, U is the earth's interior. For the source field, U is the earth's surface, and for the receiver field, U is the curved space over the western US region (Figure 4.1). In subsequent sections we only consider the model with $\nu + \frac{d}{2} =: \alpha = 2$ due to computational simplicity and software limitations. As mentioned in the previous section, for source/receiver fields ($d = 2$) we have $\nu = 1$, and for the velocity field ($d = 3$), we have $\nu = 1/2$ which yields an exponential covariance function. For theoretical results for $\alpha > 2$ and $d = 2$ we refer to Lindgren et al. (2011).

The technique for calculating Q is based on the finite element method (FEM). For the theory of the FEM we refer to Hughes (1987) and Gockenbach (2006). In the context of our model the Z_j in (4.6) correspond to the velocity field given by the β_j in (2.1). For modeling $\boldsymbol{\beta}$ we derive the explicit form of the precision matrix Q on the $\mathbb{R}^3$ manifold. The calculation technique of the 3-D field is similar to the method used in the 2-D case as described in Lindgren et al. (2011). The difference is that each element of a 3-D mesh is a tetrahedron instead of a triangle in a 2-D mesh and the vertices of the tetrahedra are in $\mathbb{R}^3$. In this case, the volume and surface of each tetrahedron are needed for the calculation of the precision matrix. Explicit expressions of $\tilde{C}$ and G matrices for one building block of Q using tetrahedral meshes are provided in the Appendix.

4.3.2 Model specification using the Integrated Nested Laplace Approximations (INLA)

Bayesian inference of the model (4.2) in our study is facilitated by the MCMC-free method – the integrated nested Laplace approximations (INLA) developed by Rue et al. (2009). The INLA method performs deterministic inference for latent Gaussian models, where the latent field is Gaussian, parameterized by a few hyper-parameters, and the response can be Gaussian or non-Gaussian observations. For Gaussian observations, the accuracy of such deterministic inference is limited only by the error in the numerical integration, whereas for non-Gaussian observations, the approximation can be accurately carried out by a set of Laplace approximations (Simpson et al., 2012). The INLA method has been successfully applied to a large number of spatial problems, for example, Schrödle and Held (2011), Cameletti et al. (2012), Illian et al. (2012). For more details of the theory of INLA and related models we refer to Rue and Martino (2007) and Rue et al. (2009). All calculations in our study are performed in R with the INLA software packages (www.r-inla.org).
Recall the linear model in (4.3),

$$\boldsymbol{Y} = H\boldsymbol{m} + \boldsymbol{\epsilon} \quad \text{with } \boldsymbol{m} := (\boldsymbol{\beta}'_{\text{usa}}, \boldsymbol{\gamma}', \boldsymbol{\xi}')' =: (m_1, ..., m_D)'.$$

For simplicity we neglect the notation "usa" for the data $\boldsymbol{y}$ in this section. Assuming stationary Gaussian fields for the parameters $\boldsymbol{m}$, we denote $\boldsymbol{\theta} := (\tau_\beta, \kappa_\beta, \tau_\gamma, \kappa_\gamma, \tau_\xi, \kappa_\xi, \phi)$ to be a set of hyper-parameters. Then, the posterior distribution is specified by

$$\pi(\boldsymbol{m}, \boldsymbol{\theta}|\boldsymbol{y}) \propto \pi(\boldsymbol{\theta})\pi(\boldsymbol{m}|\boldsymbol{\theta})\pi(\boldsymbol{y}|\boldsymbol{m}, \boldsymbol{\theta}).$$

Assuming the hierarchical model

$$\begin{aligned} \boldsymbol{Y} \mid \boldsymbol{m}, \boldsymbol{\theta} &\sim \mathcal{N}_N(H\boldsymbol{m}, \frac{1}{\phi} I_N), \\ \boldsymbol{m} \mid \boldsymbol{\theta} &\sim \mathcal{N}_D(\boldsymbol{m}_0, \boldsymbol{Q}(\boldsymbol{\theta})^{-1}), \\ \boldsymbol{\theta} &\sim \pi(\boldsymbol{\theta}), \end{aligned} \tag{4.9}$$

The prior precision matrix $\boldsymbol{Q}(\boldsymbol{\theta})$ is given by

$$\boldsymbol{Q}(\boldsymbol{\theta}) := \begin{pmatrix} Q_\beta(\kappa_\beta, \tau_\beta) & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & Q_\gamma(\kappa_\gamma, \tau_\gamma) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & Q_\xi(\kappa_\xi, \tau_\xi) \end{pmatrix}. \tag{4.10}$$

Here, $Q_\beta(\kappa_\beta, \tau_\beta)$ is the precision matrix for $\boldsymbol{\beta}_{\text{usa}}$ as defined in (4.8) with the corresponding $\tilde{C}$ and G for dimensionality $d = 3$, whereas $Q_\gamma(\kappa_\gamma, \tau_\gamma)$ and $Q_\xi(\kappa_\xi, \tau_\xi)$ are the precision matrices for $\boldsymbol{\gamma}$ and $\boldsymbol{\xi}$ for $d = 2$, respectively. The prior mean is denoted as $\boldsymbol{m}_0 := (\boldsymbol{\beta}_0, \mathbf{0}, \mathbf{0})'$, where for $\boldsymbol{\beta}_0$ we apply the LSQR solution $\boldsymbol{\beta}_{\text{usa}}^{\text{LSQR}}$ obtained by minimizing the objective function $\|\boldsymbol{y} - X_{\text{usa}}\boldsymbol{\beta}_{\text{usa}}\|^2 + \lambda\|\boldsymbol{\beta}_{\text{usa}}\|^2$ using the LSQR algorithm (Sigloch, 2008). The full conditional of $\boldsymbol{m}$ can be expressed as,

$$\boldsymbol{m} \mid \boldsymbol{y}, \boldsymbol{\theta} \sim \mathcal{N}_D(\Omega^{-1}\boldsymbol{\zeta}, \Omega^{-1}), \quad \text{with } \Omega := \boldsymbol{Q}(\boldsymbol{\theta}) + \phi H^t H,\ \boldsymbol{\zeta} := \boldsymbol{Q}(\boldsymbol{\theta})\boldsymbol{m}_0 + \phi H^t \boldsymbol{y}. \tag{4.11}$$

Traditionally, the marginal posterior distributions $\pi(\boldsymbol{m}|\boldsymbol{y})$ and $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ are obtained by sampling, using MCMC methods. For example, (4.11) can be sampled using the efficient Cholesky decomposition method if the posterior precision matrix Ω is sparse. The sparsity of Ω is guaranteed by the sparsity of the matrix H (Rue and Held, 2005; Zhang et al., 2013a). This is important in the INLA context. Since the sparsity of H ensures the Markovian property of the latent field $\boldsymbol{m}$, fast approximation using the INLA scheme is valid and possible for this model (Simpson et al., 2012).

Following the description in Rue et al. (2009) and Simpson et al. (2012), the main goal in INLA is to compute the marginal posterior distribution $\pi(m_i|\boldsymbol{y})$ of the latent Gaussian field $\boldsymbol{m}$. In particular,

$$\pi(m_i|\boldsymbol{y}) = \int \pi(m_i|\boldsymbol{\theta}, \boldsymbol{y})\pi(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}, \quad i = 1, ..., D. \tag{4.12}$$

Rue et al. (2009) presented three approximations, denoted as $\tilde{\pi}(m_i|\boldsymbol{\theta}, \boldsymbol{y})$, for approximating the true posterior distributions $\pi(m_i|\boldsymbol{\theta}, \boldsymbol{y})$, i.e., the Gaussian, the Laplace, and a simplified Laplace approximation. Further, the posterior distribution of the hyper priors $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ is approximated again by the Laplace approximation given by (Tierney and Kadane, 1986)

$$\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y}) \propto \frac{\pi(\boldsymbol{\theta})\pi(\boldsymbol{m}|\boldsymbol{\theta})\pi(\boldsymbol{y}|\boldsymbol{m}, \boldsymbol{\theta})}{\pi(\boldsymbol{m}|\boldsymbol{\theta}, \boldsymbol{y})} \approx \frac{\pi(\boldsymbol{\theta})\pi(\boldsymbol{m}|\boldsymbol{\theta})\pi(\boldsymbol{y}|\boldsymbol{m}, \boldsymbol{\theta})}{\tilde{\pi}_G(\boldsymbol{m}|\boldsymbol{\theta}, \boldsymbol{y})}\Big|_{\boldsymbol{m}=\boldsymbol{m}^*(\boldsymbol{\theta})}.$$

Here $\tilde{\pi}_G(\boldsymbol{m}|\boldsymbol{\theta}, \boldsymbol{y})$ is the Gaussian approximation to $\pi(\boldsymbol{m}|\boldsymbol{\theta}, \boldsymbol{y})$. Using the integration points $\boldsymbol{\theta}_k$, $\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$ is calculated numerically. The approximation of the marginal posterior in (4.12), denoted as $\tilde{\pi}(m_i|\boldsymbol{y})$, is calculated using numerical integration, i.e.

$$\tilde{\pi}(m_i|\boldsymbol{y}) = \int \tilde{\pi}(m_i|\boldsymbol{\theta}, \boldsymbol{y})\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta} \approx \sum_{k=1}^{K} \tilde{\pi}(m_i|\boldsymbol{\theta}_k, \boldsymbol{y})\tilde{\pi}(\boldsymbol{\theta}_k|\boldsymbol{y})\Delta_k$$

Δ_k are integration weights and $\boldsymbol{\theta}_k$ are the integration points.

To wrap up, we develop the Gaussian Matérn field approximation for a $\mathbb{R}^3$ manifold by means of the SPDE approach using finite dimensional linear basis functions. Thus, spatial dependency of the velocity field as well as the source and receiver fields are specified using the SPDE representations. The INLA approach is then applied to the GMRF to facilitate the Bayesian inference for the high-dimensional parameter estimations.

4.4 Simulation study

4.4.1 Simulation setup

In this section we investigate the performance of our approach estimating synthetic data. We assume that the errors $\boldsymbol{\epsilon}$ are independently identically normal distributed i.e. $\boldsymbol{\epsilon} \sim \mathcal{N}(0, 1)$ with a variance $1/\phi = 1.0$ s. We choose a synthetic variance of 1.0 s since it is similar to the variance we expect from the real data. We consider the full model in (4.2) and three reduced models as follows:

$$\begin{aligned}
\text{(VSR)} &: \boldsymbol{y} = X_{\text{usa}}\boldsymbol{\beta}_{\text{usa}} + B_s\boldsymbol{\gamma} + B_r\boldsymbol{\xi} + \boldsymbol{\epsilon} \\
\text{(VS)} &: \boldsymbol{y} = X_{\text{usa}}\boldsymbol{\beta}_{\text{usa}} + B_s\boldsymbol{\gamma} + \boldsymbol{\epsilon} \\
\text{(VR)} &: \boldsymbol{y} = X_{\text{usa}}\boldsymbol{\beta}_{\text{usa}} + B_r\boldsymbol{\xi} + \boldsymbol{\epsilon} \\
\text{(V)} &: \boldsymbol{y} = X_{\text{usa}}\boldsymbol{\beta}_{\text{usa}} + \boldsymbol{\epsilon}
\end{aligned}$$

The synthetic traveltime data $\boldsymbol{y}$ are generated using the full model (VSR: velocity, source and receiver fields) and estimated using the (VSR), (VS: velocity and source fields) and (VR: velocity and receiver fields) models, respectively, since we want to examine the performance of these models. We simulate the synthetic random fields $\boldsymbol{\beta}_{\text{usa}}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\xi}$ using (4.9) for both high and low spatial correlations ρ. The true hyper parameters are specified as follows.

- *Simulated velocity field $\boldsymbol{\beta}_{usa}$ in $\mathbb{R}^3$:*
  For dimensionality $d = 3$ and variance of the random field $\sigma_\beta = 1$, we have $\kappa_\beta := 2/\rho$ and $\tau_\beta := 1/\sqrt{8\pi\kappa_\beta}$. We want to investigate the performance of the models when the correlation range ρ_β is high, i.e. $\rho_\beta = 0.8$. Recall that in the Matérn covariance function ρ_β means the distance between vertices at which the spatial correlation is about 0.1. It follows that the true values for κ_β and τ_β are fixed at $\kappa_\beta = 2.5$ and $\tau_\beta = 0.126$, respectively.

- *Simulated source and receiver fields $\boldsymbol{\gamma}$ and $\boldsymbol{\xi}$ in $\mathbb{S}^2$:*
  For $d = 2$ and $\sigma_\gamma = 1$ the hyper parameters of source field parameters are given as $\kappa_\gamma = 2\sqrt{2}/\rho_\gamma$ and $\tau_\gamma = 1/\sqrt{4\pi\kappa_\gamma^2}$ (the same equations are applied for the receiver field parameters). We want to investigate the performance of the models when correlation ranges of source and receiver fields are arbitrarily fixed at high and low values, respectively.

  - *Source field $\boldsymbol{\gamma}$ with a high spatial correlation:*
    We fix the true value $\rho_\gamma = 0.8$. It follows that the values for τ_γ and κ_γ are given by $\tau_\gamma = 0.080$ and $\kappa_\gamma = 3.536$, respectively.
  - *Receiver field $\boldsymbol{\xi}$ with a low spatial correlation:*
    Here, the true value is $\rho_\xi = 0.3$. It follows that the true values of τ_ξ and κ_ξ are $\tau_\xi = 0.030$ and $\kappa_\xi = 9.428$, respectively.

Although the true values of the hyper parameters can be chosen arbitrarily, the range of parameter ρ should not exceed the range of the entire domain (the maximum distance between two arbitrary mesh vertices) as mentioned in Simpson et al. (2012). We note that the sphere with filling interior used for the velocity field $\boldsymbol{\beta}_{\text{usa}}$ is re-scaled from a radius of 6420 km to a radius of 20 km to avoid numerical problems (6420 km is slightly larger than earth radius to ensure a convex mesh). However, the estimated values of the hyper parameters κ and τ are adapted to the radius of the sphere. The parameters could be re-scaled to obtain a model for any radius.

4.4.2 Performance evaluation measures

To assess the performance of the models, we evaluate the simulated data using the Mahalanobis distance, defined by $\|\boldsymbol{x}\|_\Sigma := \sqrt{(\boldsymbol{x}-\boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}$ with the covariance matrix Σ. The deviance information criterion (DIC) and the effective number of parameters (denoted as p_D) are computed using the INLA program. The model with the smallest DIC

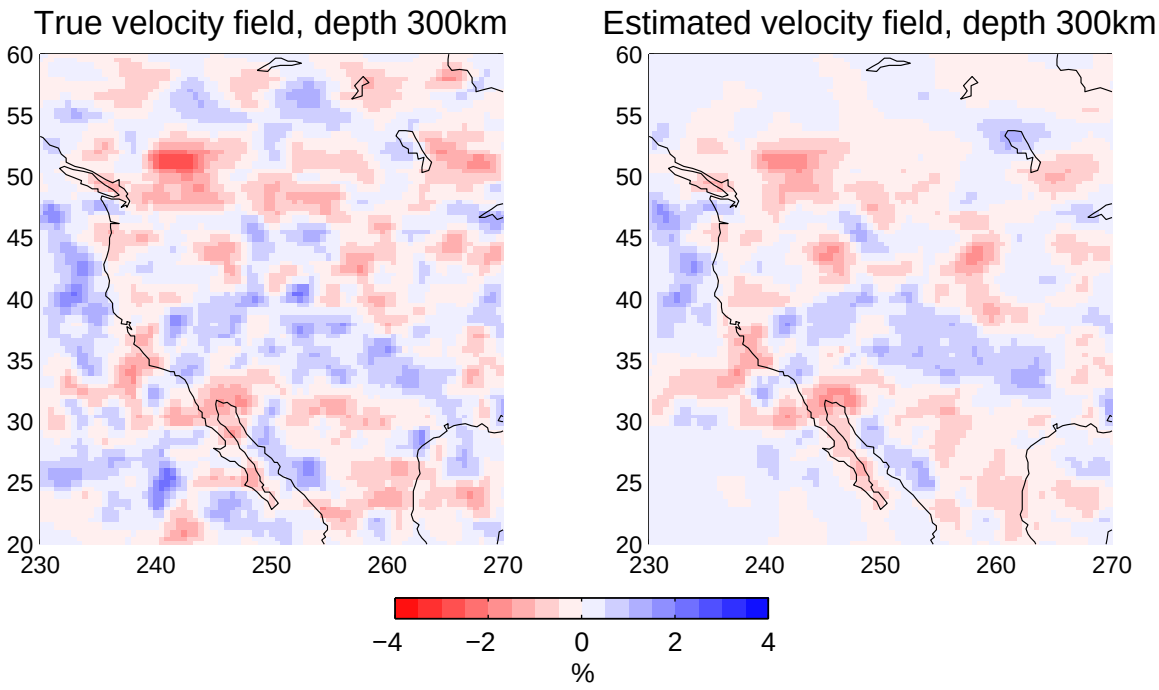

Figure 4.3: Synthetic true velocity field $\boldsymbol{\beta}_{\text{usa}}$ and its estimated posterior mean in the model (V: single velocity) with prior $\boldsymbol{\beta}_0 = \mathbf{0}$. Unit on the color bar is %, indicating the velocity deviation $\boldsymbol{\beta}$ from the spherically symmetric reference earth model. Blue colors represent zones where seismic velocities are faster than in the reference earth model while red colors denote relatively slower velocities.

value is preferred with regard to both model fit and model complexity (Spiegelhalter et al., 2002). For p_D, we obtain $p_D \approx D - tr(Q(\boldsymbol{\theta})Q^{\star}(\boldsymbol{\theta})^{-1})$, where $Q(\boldsymbol{\theta})$ is the prior precision matrix and $Q^{\star}(\boldsymbol{\theta})$ is the posterior covariance matrix and $D := -2\log(\ell(\boldsymbol{y}|\bar{\boldsymbol{\theta}}))$. The term p_D can be seen as a Bayesian measure of fit. As pointed out in Rue et al. (2009), the smaller the value p_D is compared to the size of the data, the better the model fits the data.

4.4.3 Simulation results and interpretation

Recall that the synthetic data are generated using the (VSR) model and estimated using (VSR), (VS) and (VR) models. We show an exemplary 2-D section through the 3-D velocity field $\boldsymbol{\beta}_{\text{usa}}$ in Figure 4.3, comparing the true synthetic $\boldsymbol{\beta}$-field from a depth of 300 km and its estimated posterior mean from the (VSR) model. The images show that

our model can reconstruct the major structural features, especially the portion of the seismic velocity field located beneath the densely instrumented western United States.

Figure 4.4 illustrates the true random fields for source and receivers (top row), as well as their posterior mean estimates from the (VSR) model (bottom row). It shows that the (VSR) model can recover the basic structure of the true source and receiver fields. However, for areas with less data or less dense mesh points, structure cannot be reconstructed as well (e.g. regions between the Antarctic, the Indian Ocean, and the Pacific region). This is explained by the lack of sources in these regions.

The summary statistics resulting from the (VSR), (VS) and (VR) models in Table 4.1 show that the true hyper parameters can be recovered within the 95% credible intervals from the estimates of the (VSR) and (VS) models. Estimates of (VR) can only correctly identify the receiver field while the velocity estimate is polluted by the attempt to fit traveltimes originating from the unmodeled source field. We also examine the performance of the (VR) model if the synthetic receiver field is simulated with a high spatial correlation parameter (not shown in Table 4.1), i.e., $\rho_\xi = 0.8$, instead of the listed value of $\rho_\xi = 0.3$. In this case the true hyper parameters of the velocity field can be recovered within 95% credible intervals, however the estimated mean cannot identify the true value.

The DIC and p_D values measuring Bayesian complexity and data fit, as well as the maximum likelihood value MaxLik in Table 4.2 provide the same conclusion: (VSR) and (VS) perform about equally well according to their DIC, p_D and MaxLik values, whereas performance of (VR) is worse on all accounts: DIC and p_D are higher than for (VSR) and (VS), whereas MaxLik is lower. Although (VR) achieves lower misfit values with respect to the measure $\|\boldsymbol{y} - H\hat{\boldsymbol{m}}\|_{\Sigma_y}$, the misfits within the estimated credible intervals ($\|\boldsymbol{y} - H\hat{\boldsymbol{m}}_L\|_{\Sigma_y}$ and $\|\boldsymbol{y} - H\hat{\boldsymbol{m}}_U\|_{\Sigma_y}$) indicate that the uncertainty is smaller for both the (VSR) and (VS) models.

4.5 Application to the seismic traveltime data

In this section we apply the INLA algorithm to infer the velocity field as well as the source and receiver fields using real traveltime delay data measured by Sigloch (2008) or as described in (3.1). Here, we use only measurements on the broadband waveforms whereas the original data set contained additional bandpassed measurements for the same source-receiver paths. Similar data sets for the western United States were studied by other researchers, e.g., Burdick et al. (2008), Sigloch et al. (2008), Schmandt and Humphreys

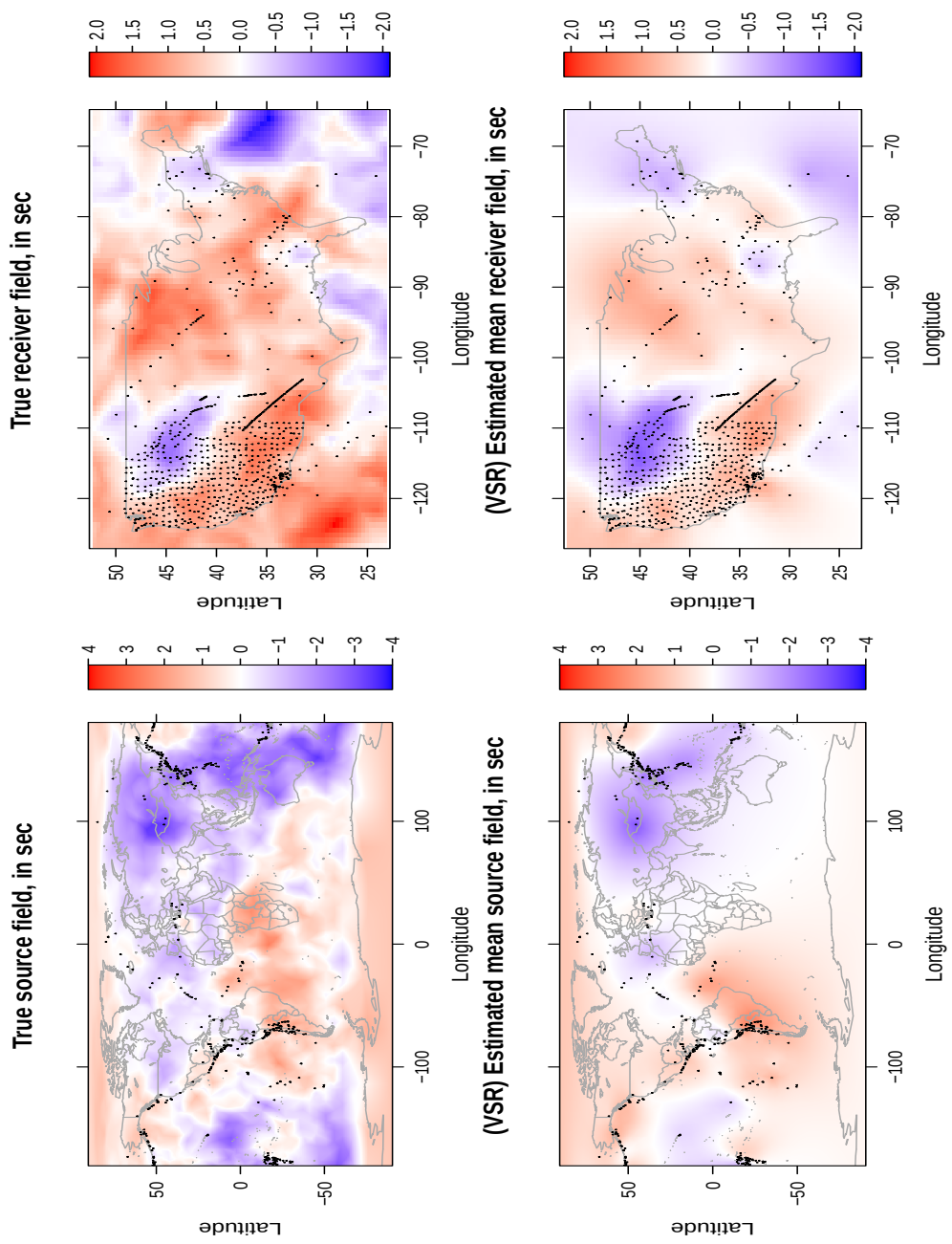

Figure 4.4: Synthetic source and receiver fields and their posterior mean estimation from the (VSR) model. Unit on the color bar is in seconds, measuring traveltime delays. The blue color means that the observed arrival time y_i^{obs} are advanced relative to the predicted arrival time y_i^{pred} whereas the red color means a traveltime lag. The black dots on the maps indicate the source and receiver locations.

Summary statistics								
	Parameters	$\hat{\phi}$	$\hat{\tau}_\beta$	$\hat{\kappa}_\beta$	$\hat{\tau}_\gamma$	$\hat{\kappa}_\gamma$	$\hat{\tau}_\xi$	$\hat{\kappa}_\xi$
Model	True	1.000	0.126	2.5	0.080	3.536	0.030	9.428
(VSR)	mean	1.001	0.122	2.343	0.074	3.581	0.031	15.720
	2.5%	0.989	0.113	1.984	0.065	1.774	0.024	10.714
	97.5%	1.015	0.130	2.876	0.088	5.347	0.039	23.226
(VS)	mean	0.981	0.122	2.207	0.080	2.995		
	2.5%	0.969	0.115	1.896	0.070	2.002	–	–
	97.5%	0.998	0.132	2.513	0.095	4.340		
(VR)	mean	0.669	0.064	0.564			0.031	13.066
	2.5%	0.661	0.061	0.280	–	–	0.024	8.626
	97.5%	0.680	0.067	0.977			0.039	19.320

Table 4.1: Summary statistics of the posterior estimates of the hyper parameters using simulated data. The true parameter values are well covered by the estimates of (VSR) and (VS) models whereas the (VR) model cannot estimate the hyper parameters of the 3-D velocity field correctly, as indicated by the estimated mean and credible intervals compared to the true values.

Performance measures						
Model	$\|\boldsymbol{y} - H\hat{\boldsymbol{m}}\|_{\Sigma_y}$	$\|\boldsymbol{y} - H\hat{\boldsymbol{m}}_L\|_{\Sigma_y}$	$\|\boldsymbol{y} - H\hat{\boldsymbol{m}}_U\|_{\Sigma_y}$	DIC	p_D	MaxLik
(VSR)	226.99	633.09	634.22	152,917	1813	-428,397
(VS)	232.39	750.28	748.27	153,929	1709	-428,852
(VR)	184.93	891.80	891.98	174,707	2162	-439,986

Table 4.2: Performance measures of the posterior estimation results using simulated data show that (VSR) and (VS) perform better than the (VR) model with respect to DIC, p_D and MaxLik.

(2010), and Zhang et al. (2013a). Traditionally, the velocity field is estimated by minimizing the L2-norm using the LSQR algorithm (Nolet, 1987; Sigloch et al., 2008; Tian et al., 2009), whereas Zhang et al. (2013a) applied the MCMC algorithm to provide estimates with uncertainty assessment (see Chapter 2 and 3). Here, we use broadband traveltime observations from $N = 53,270$ unique source-receiver paths to estimate over $D = p_{\text{usa}} + p_s + p_r \approx 13,000$ parameters consisting of $p_{\text{usa}} = 8977$ tetrahedral nodes, $p_s = 2200$ source field parameters and $p_r = 1800$ receiver field parameters. With Gaussian error assumption we fit the data using the models (VSR), (VS), (VR) or (V) (see Section 4.4) within the INLA framework.

As prior information, we used the solution obtained by Sigloch (2008) using iterative least-squares optimization (LSQR). This solution included a velocity field (our prior $\boldsymbol{\beta}_0$), but also estimates for source corrections. We converted these corrections (for earthquake location and time) into traveltime anomalies, and applied them to our data $\boldsymbol{Y}_{\text{usa}}$. Hence our source field $\boldsymbol{\gamma}$ effectively evaluates the quality of their corrections, rather than correcting for earthquake mislocations from scratch. We made this choice in the interest of direct comparison with the results of Zhang et al. (2013a), who also accepted the corrections obtained by Sigloch (2008).

Receiver corrections were not treated as free parameters by Sigloch (2008). Instead they used prior models of the earth's crustal structure, topography, and ellipticity to compute receiver corrections. We apply these same traveltime corrections to our data $\boldsymbol{Y}_{\text{usa}}$, so that effectively our receiver field $\boldsymbol{\xi}$ assesses prior information about surface structure around the receiver, especially crustal structure, which is rather uncertain.

4.5.1 Posterior results and interpretation of the velocity, source and receiver fields

Table 4.3 provides an overview of the performance measures of the models. Results from (VSR) and (VS) models yield similar estimates, which are considerably better than the results from the (VR) and (V) models with respect to model complexity (DIC) and the maximum likelihood estimate (MaxLik). Due to the high number of parameters, the computing time of (VSR) is significantly higher than for the models with only one or two random fields. In summary, (VSR) or (VS) models would estimate the data better from the perspective of DIC. With regards to computing time, the reduced models (VS) and (V) are 3 times and 8 times less expensive than the (VSR) model, and may therefore be preferable.

Performance measure							
Model	$\|\boldsymbol{y} - H\hat{\boldsymbol{m}}\|_{\Sigma_y}$	$\|\boldsymbol{y} - H\hat{\boldsymbol{m}}_L\|_{\Sigma_y}$	$\|\boldsymbol{y} - H\hat{\boldsymbol{m}}_U\|_{\Sigma_y}$	DIC	p_D	MaxLik	Time in hours
(VSR)	147.60	288.44	288.20	198,666	1,282	-450,623	33.8
(VS)	146.61	319.78	319.53	198,817	1,253	-450,677	10.5
(VR)	144.96	277.70	277.73	200,607	967	-451,374	8.7
(V)	143.21	333.47	333.56	200,633	1,035	-451,362	4.18

Table 4.3: Performance measures of posterior estimation results using real seismic data show that (VSR) and (VS) models achieve better estimates than (VR) or (V) models.

In Figure 4.5 the posterior mean of the 3-D velocity field at a depth of 400 km below the earth's surface from the (VSR) and (V) models display strong similarities, but their respective credible intervals differ in some regions. In particular, the credible intervals from the MCMC approach in Zhang et al. (2013a) are smaller for the north-eastern region than the credible intervals from this study. Colored regions on the maps of significant regions indicate that $\boldsymbol{\beta}_{\text{usa}} = \mathbf{0}$ does not fall inside the 95% credible intervals, and hence that velocity deviation significantly differs from the spherically symmetric reference model with high posterior probability. The maps of significant regions provide independent confirmation of geological interpretation of the regions: the red areas map out hotter upper mantle regions under the volcanic provinces and plate boundary regions of western North America (Juan de Fuca spreading ridge, Basin and Range province, Yellowstone), where seismic waves travel slower than predicted by the globally averaged IASP91 earth model; the blue anomalies map out the western edge of the old and cool North American craton (Zhang et al., 2013a).

Figure 4.6 illustrates the significant regions of the estimated posterior mean of the source and receiver fields from (VSR) including nonsignificant receiver positions. Significant regions map out the areas that differ from a zero mean field with 95% posterior probability. The colored regions indicate that seismic traveltime delay patterns exist with high probability at the surface and/or at the earthquake source which influence the traveltime, which cannot be solely explained by the seismic velocity deviations, i.e. by the $\boldsymbol{\beta}$-field (which is already corrected for the deterministic source and receiver effects estimated by Sigloch (2008)). Since the estimated mean source field is very similar to its significant regions, we only show the map of significant regions here. The obtained source field looks rather "white", indicating no strong traveltime effects from mislocated sources. This is a plausible outcome because it indicates that the deterministic source corrections estimated by Sigloch (2008) were adequate. (In effect we chose to estimate required deviations from their corrections, rather than the corrections themselves. We return to this point below.)

In the following we compare estimated posterior results from INLA with results of the MCMC method by Zhang et al. (2013a) or in Chapter 3. The statistical model in Chapter 3 is formulated as $\boldsymbol{y}_{cr} = X_{\text{usa}}\boldsymbol{\beta}_{\text{usa}} + X_{\text{hyp}}\boldsymbol{\beta}_{\text{hyp}} + X_{\text{time}}\boldsymbol{\beta}_{\text{time}} + \boldsymbol{\epsilon}$, where X_{hyp} and X_{time} are sensitivity kernels for the hypocenter and origin time corrections at earthquake sources. The quantities $\boldsymbol{\beta}_{\text{hyp}}$ and $\boldsymbol{\beta}_{\text{time}}$ are the source parameters to be estimated. The model error $\boldsymbol{\epsilon}$ is assumed to be i.i.d. Gaussian distributed with zero mean and a constant variance. In Chapter 3 an ellipsoidal spatial prior on $\boldsymbol{\beta}_{\text{usa}}$ is constructed, which follows

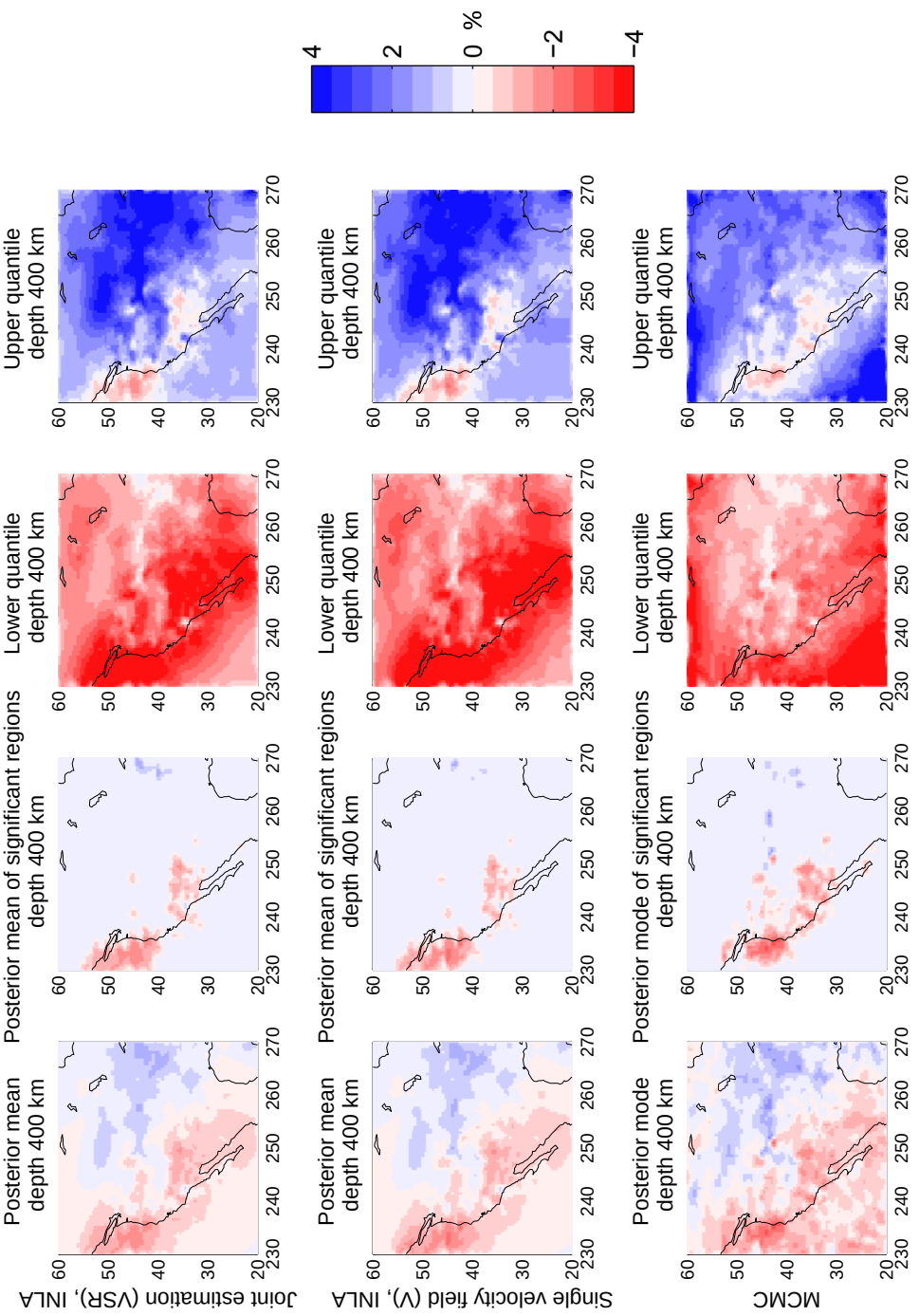

Figure 4.5: Comparison of Bayesian tomographic images with posterior mean/mode, lower and upper quantiles, and posterior mean/mode of significant regions from the (VSR) model (top row), the (V) model (mid row) and the MCMC estimates (bottom row) from Zhang et al. (2013a). Blue colored regions exhibit positive velocities of the seismic waves with respect to the reference earth IASP91 model, while the red colored areas show slower seismic velocities than the reference model. Physically, the blue color usually indicates a colder region than average, whereas red (slow) means hotter.

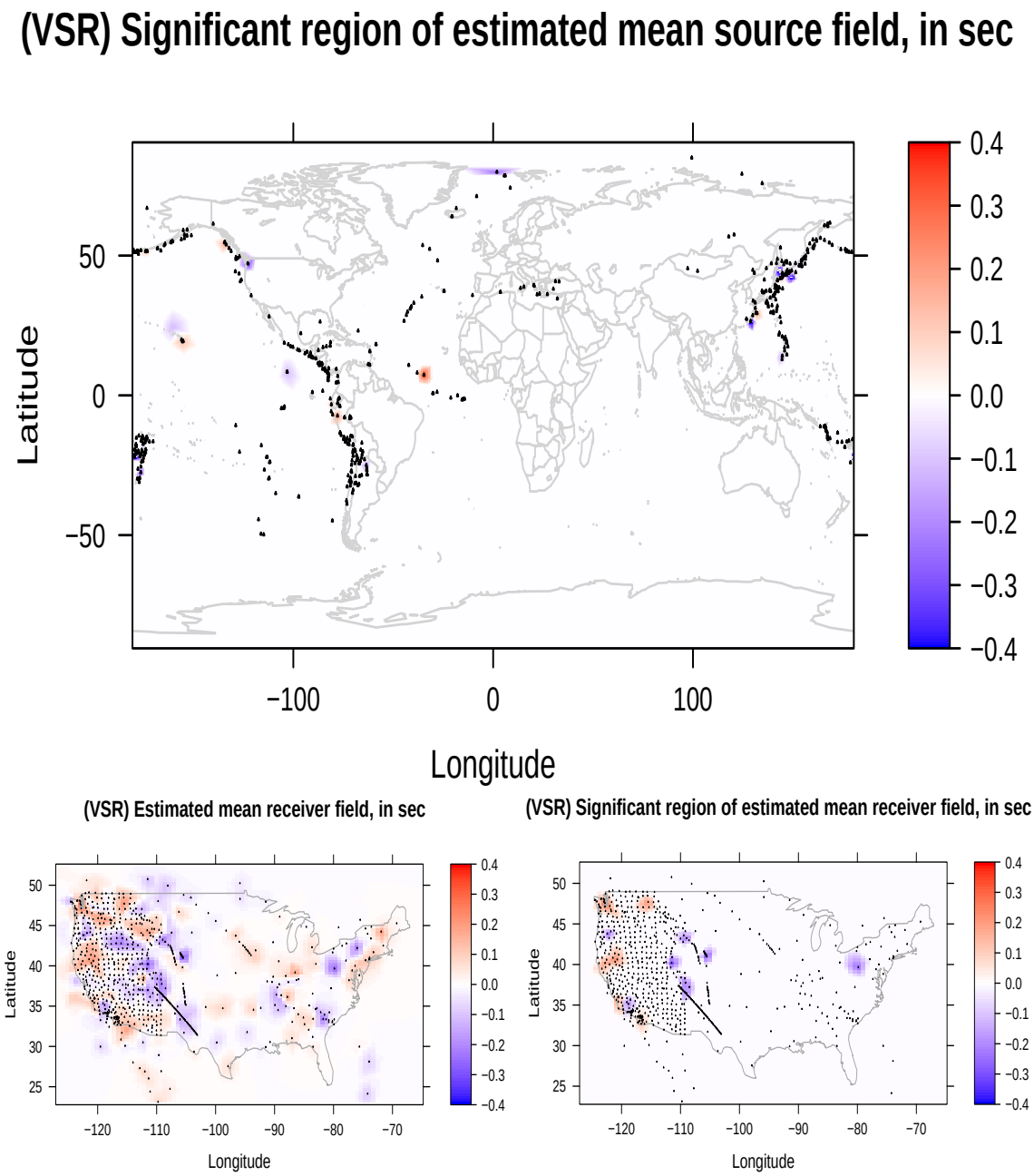

Figure 4.6: Inversion of real data: significant regions of the estimated posterior mean of source (top panel) and receiver fields (bottom panels including the estimated mean field) using the (VSR) model. Red color means faster observed arrival time than predicted while blue color implies the opposite. Black dots indicate source and receiver locations. While only few locations on the receiver fields are significantly different from a zero mean field with 95% posterior probability, most parts of the estimated mean source field are significant (not shown separately here).

a conditional autoregressive model as defined Pettitt et al. (2002). This model is non-isotropic and non-stationary. However, the same prior mean - the LSQR solution of $\boldsymbol{\beta}_0$ is applied in both previous and this study. A Gibbs-sampling with a MH-step is adopted for estimation. Their parameter dimension is about 11,000, which is less than the number of model parameters of this study ($\approx 13,000$).

$\boldsymbol{\beta}_{\text{hyp}}$ (three free parameters per source) and $\boldsymbol{\beta}_{\text{time}}$ (one parameter per source) have the physical meaning of source location and origin time, respectively, whereas $X_{\text{hyp}}\boldsymbol{\beta}_{\text{hyp}}$ and $X_{\text{time}}\boldsymbol{\beta}_{\text{time}}$ are traveltime anomalies. Both Sigloch (2008) and Zhang et al. (2013a) estimated $\boldsymbol{\beta}_{\text{hyp}}$ and $\boldsymbol{\beta}_{\text{time}}$. In principle, the INLA method is equally capable of handling such spatially discrete corrections, but we did not attempt this due to the large number of additional parameters and considerations of computing time. Instead, we chose the approach of estimating the continuous source field $\boldsymbol{\gamma}$ of traveltime anomalies, and in addition we *a priori* corrected our traveltime observations $\boldsymbol{Y}_{\text{usa}}$ using the deterministic source and receiver corrections obtained by Sigloch (2008). These fields $\boldsymbol{\gamma}$ and $\boldsymbol{\xi}$ exist at all vertices of the source and receiver field meshes, and their spatial correlations are modeled, whereas $\boldsymbol{\beta}_{\text{hyp}}$ and $\boldsymbol{\beta}_{\text{time}}$ assumes no spatial correlations and exist only at source and receiver locations. As a practical note on this rather technical subject, different tomographers choose to treat source and receiver corrections differently. Our main purpose here was not (yet) to quantitatively compare to existing practices, but to introduce an alternative concept for how these corrections could be done, namely by continuous and spatially correlated 2-D random fields, and to deliver a proof of concept.

The comparison of the INLA results to earlier MCMC and LSQR inversions, e.g. the models by Zhang et al. (2013a) and Sigloch (2008), confirms that Bayesian inference using INLA successfully retrieves the first-order structures of earth's upper mantle under the densely instrumented and thus well illuminated western United States. Values at the boundaries of the images from INLA are smoother than the results from MCMC. Maps of significant regions resulting from (VSR), (V) and MCMC are similar. Further, the statistical misfit $\|\boldsymbol{y} - H\hat{\boldsymbol{m}}\|_{\Sigma_y}$ of model (V) in Table 4.3 shows an improvement of about 40% over the corresponding misfit measure in Table 3.2 in Chapter 3. The MCMC algorithm runs for over 32 CPU hours to estimate about 11,000 parameters (velocity parameters + source correction parameters), and about 7.5 CPU hours to estimate about 8977 velocity parameters (using 10,000 MCMC runs). On the same infrastructure the INLA program takes about 33 CPU hours to run the full model (VSR) with about 13,000 parameters (velocity parameters + source and receiver fields on $\mathbb{S}^2$-grids). Estimating model (V), which contains only the 8977 velocity parameters, takes about 4 CPU hours

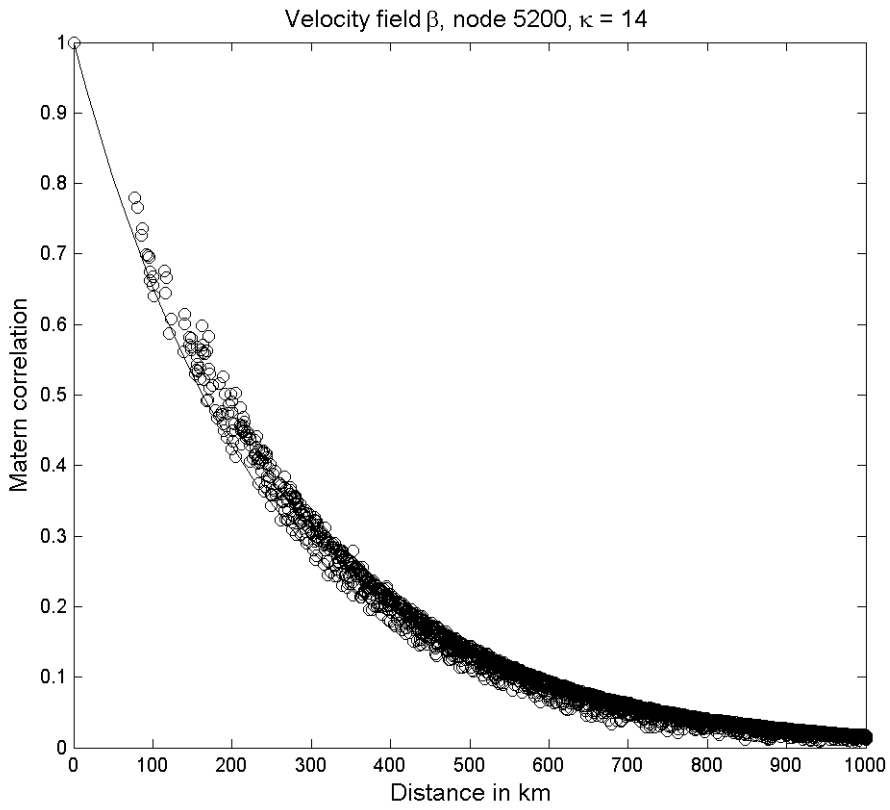

Figure 4.7: Empirical values (circles) obtained from the approximated Matérn correlation of the 3-D velocity field for a synthetic value of $\kappa = 14$. This confirms that our 3-D Matérn model approximates the theoretical function correctly.

with INLA. This shows that for this type of seismic tomography problems, the INLA algorithm has a computational advantage over the MCMC method with a speedup of about 1.5 - 2 times while delivering similar posterior mean images of the velocity fields.

4.5.2 Estimated Matérn correlation of the velocity, source and receiver fields

Figure 4.7 illustrates the approximated Matérn correlation of one vertex in the 3-D velocity field for a synthetic value of $\kappa = 14$ and the corresponding theoretical Matérn correlation. This confirms that our 3-D Matérn model approximates the theoretical function correctly. Figure 4.8 shows the Matérn correlation of one vertex in each of the estimated velocity,

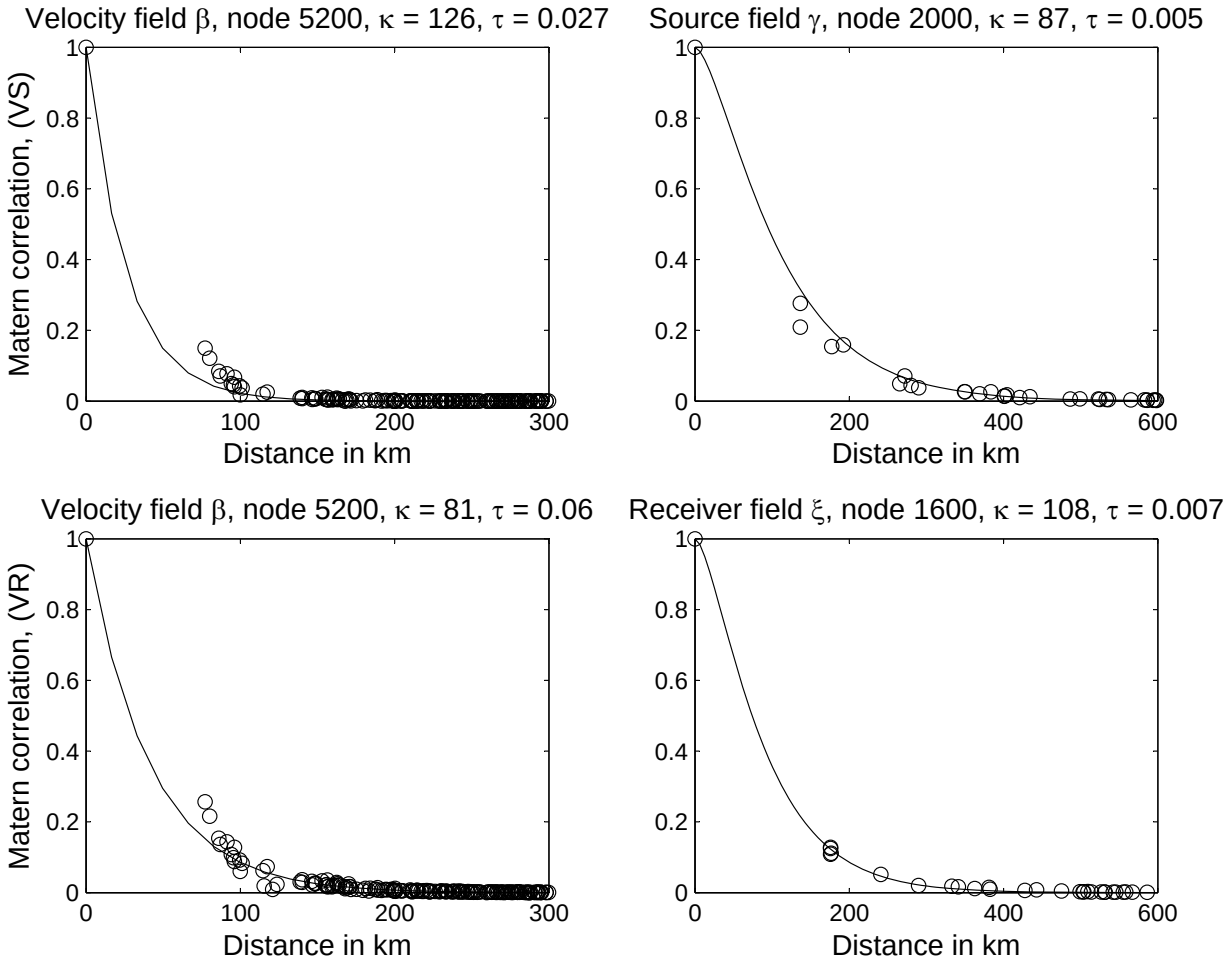

Figure 4.8: Estimated Matérn correlations from (VS) and (VR) for the real data: node #5200 of the velocity field $\boldsymbol{\beta}_{\text{usa}}$ (located at 160 km depth) with posterior mean values $\hat{\kappa}_\beta = 126$, $\hat{\tau}_\beta = 0.027$ for (VS) and $\hat{\kappa}_\beta = 81$, $\hat{\tau}_\beta = 0.06$ for (VR), respectively (left panels); results from (VS): node #2000 of the source field $\boldsymbol{\gamma}$ with $\hat{\kappa}_\beta = 87$ and $\hat{\tau}_\beta = 0.005$ (right top); and results from (VR): Node #1600 of the receiver field $\boldsymbol{\xi}$ with $\hat{\kappa}_\xi = 108$ and $\hat{\tau}_\xi = 0.007$ (right bottom). Estimated correlations on each mesh node obtained from the approximated Matérn correlation matrix using the SPDE approach (circles) are close to their corresponding theoretical values (solid lines) which confirms the correct approximation of our 3-D Matérn model.

source and receiver fields from the (VS) and (VR) models. Matérn correlation of node #5200 of the velocity field with posterior mean estimates reveals spatial correlation within 150 km, and the correlation falls to close to zero at a distance around 200 km (left panels).

The source and receiver fields (with estimates $\hat{\kappa}_\gamma = 87$ and $\hat{\tau}_\gamma = 0.005$ for the source field and $\hat{\kappa}_\xi = 108$ and $\hat{\tau}_\xi = 0.007$ for the receiver field) exhibit similar spatial correlations compared to the 3-D velocity field, which can be also confirmed by comparing the relative magnitude of κ's: the larger the value of κ is, the weaker the spatial correlation. Weak spatial correlations present in the source and receiver fields imply that the source correction that had been applied a priori, based on a 3-D model of the continental, crust and source corrections explained the data well at most locations.

The mesh vertices, for which the empirical Matérn correlations are illustrated here, are selected arbitrarily within an area away from the boundaries of the meshes. We note that the empirical correlations can vary from the theoretical correlations if the vertices are located at the mesh boundaries. Generally, variances at vertices close to boundaries are much larger due to the Neumann boundary conditions (Lindgren et al., 2011; Simpson et al., 2011). This phenomenon means that the approximation of the 3-D GF is a solution to the SPDE, but not a stationary solution (constant variance everywhere inside the domain). One possibility to overcome this non-stationarity is to construct a larger meshed domain than the target domain to run the inference. The original radius of our tetrahedral mesh for the globe is 6420 km, 50 km larger than the true radius of the earth (6371 km). However, this may not be sufficient to generate a purely stationary field for our target region. Theoretical solutions to avoid variance inhomogeneity is a topic of current research.

4.6 Discussion and outlook

This chapter studies high-dimensional linear inverse problems by means of Bayesian spatial modeling techniques. We construct a new, computationally efficient approximation to the velocity field on a $\mathbb{R}^3$-manifold using the stochastic partial differential equations (SPDE) approach by Lindgren et al. (2011). As an exemplary application, we estimate the 3-D seismic wave velocity field in the upper mantle beneath the western United States, while also allowing for spatial correlation structure of the data errors, which are expressed through the source and receiver fields over the earth's surface. In our tomographic problem, Bayesian uncertainty quantification of the random fields is carried out jointly by the integrated nested Laplace approximation (INLA) from (Rue et al., 2009) with the

R-INLA software. Our application to upper mantle structure involves about 13,000 parameters and over 53,000 observations. Simulation studies and real data analysis showed that, depending on the dimensionality of the parameters, our approach using INLA performs approximately 1.5-2 times faster than the MCMC method in Zhang et al. (2013a) (Chapter 3), while delivering comparable posterior tomographic images with better statistical fits. Spatial pattern detected on the source and receiver fields imply that there exist additional systematic errors at certain sources or receivers locations that cannot be solely explained by the physical model and source corrections.

Our current model using INLA is limited to the assumption that the posterior distribution of the parameters has one posterior mode, in order to achieve accurate approximation of the MCMC estimates. If the posterior distribution is close to Gaussian, the INLA approximation is exact (Rue et al., 2009). It is not applicable to estimating parameter uncertainties in complex physical models, e.g. PDE systems, where multiple modes exist in the posterior distributions. For a literature study on Bayesian methods for large-scale PDE systems we refer to the discussion section in Chapter 3. An extension of INLA methodology for mixture-distributions with multiple modes remains a current research topic. INLA works for the class of Gaussian latent models with nonlinear likelihood and/or non-Gaussian responses. For non-Gaussian data errors, such as t-distributed error, the INLA method is computationally still too time consuming for our application. Martin et al. (2012) extends the INLA theory beyond the scope of latent Gaussian models, where the latent field can have a near-Gaussian distribution, another ongoing topic of research.

The SPDE approach of Lindgren et al. (2011) can be extended in several directions: Spatial modeling techniques using the SPDE approach can be flexibly extended to non-stationary, non-isotropic Gaussian fields, and space-time models on manifolds, by defining a few additional space-dependent control parameters for the hyper-parameters of the random fields (Simpson et al., 2012) (e.g., using a spline function to approximate the hyper-parameters). Bolin (2012) extends the SPDE approach to non-Gaussian fields. Other research focuses on the extension of the Matérn class approximation to other classes of the continuous Gaussian fields when α in the SPDE equation (4.7) is not an integer. These developments might be applied to our 3-D random field application in the future. For a comprehensive discussion on the SPDE approach we refer to Lindgren (2012); Simpson et al. (2012).

In conclusion, our application is the first example for modeling continuous Gaussian fields in a 3-D space. It allows for spatial correlations of the data errors capturing spatial dependency of data errors caused by earthquake sources and receivers simultaneously.

Our application demonstrates an example of very high-dimensional parameter estimation carried out within the INLA framework.

Acknowledgement

We thank Prof. Lindgren for his valuable comments and advice. The numerical simulations were carried out with the R-INLA software (www.r-inla.org) on the computational facilities of the Department of Earth and Environmental Sciences at the Ludwig-Maximilians-Universität München and the Leibniz Supercomputing Center in Garching. The first author is funded by the project B3 of the Munich Center of Advanced Computing (MAC), Technische Universität München.

Appendix A

Calculation of the precision matrix elements on 3-D manifold

The calculation of the $\tilde{C}$ and G matrices are accomplished by using the theory of the finite element method (FEM). Advanced implementation techniques can be seen in Alberty et al. (2002). Here, we show the explicit expression of the $\tilde{C}$ and G matrices for one tetrahedron as shown in Figure A.1. Although the explicit expressions of C and B matrices as part of the precision matrix Q are not applied in our study, we show them in this appendix. Since they are important part of the SPDE approach developed by Lindgren et al. (2011), these explicit expressions can be used for further research.

First, we denote the four vertices in a tetrahedron as

$$\boldsymbol{a}_1 := (x_1, y_1, z_1)', \ \boldsymbol{a}_2 := (x_2, y_2, z_2)', \ \boldsymbol{a}_3 := (x_3, y_3, z_3)' \text{ and } \boldsymbol{a}_4 := (x_4, y_4, z_4)'.$$

We also define the *directional* vectors $\boldsymbol{v}_{ij}$ as $\boldsymbol{v}_{ij} := \boldsymbol{a}_i - \boldsymbol{a}_j$ for $i, j = 1, ..., 4$ as illustrated in Figure A.1. Every tetrahedral element in the original mesh will be calculated based on a reference tetrahedron with the (ξ, γ, η)-coordinates. Furthermore, the cross product of two vectors $\boldsymbol{u} := (u_x, u_y, u_z)'$ and $\boldsymbol{v} := (v_x, v_y, v_z)'$ is defined by

$$\boldsymbol{u} \times \boldsymbol{v} := \boldsymbol{e}_1(u_y v_z - u_z v_y) + \boldsymbol{e}_2(u_z v_x - u_x v_z) + \boldsymbol{e}_3(u_x v_y - u_y v_x),$$

where $\boldsymbol{e}_1 := (1, 0, 0)'$, $\boldsymbol{e}_2 := (0, 1, 0)'$ and $\boldsymbol{e}_3 := (0, 0, 1)'$ are standard basis vectors. The cross product of two vectors is always perpendicular to both vectors with the orientation determined by the right-hand rule (Arfken, 1985; Bronshtein et al., 2007).

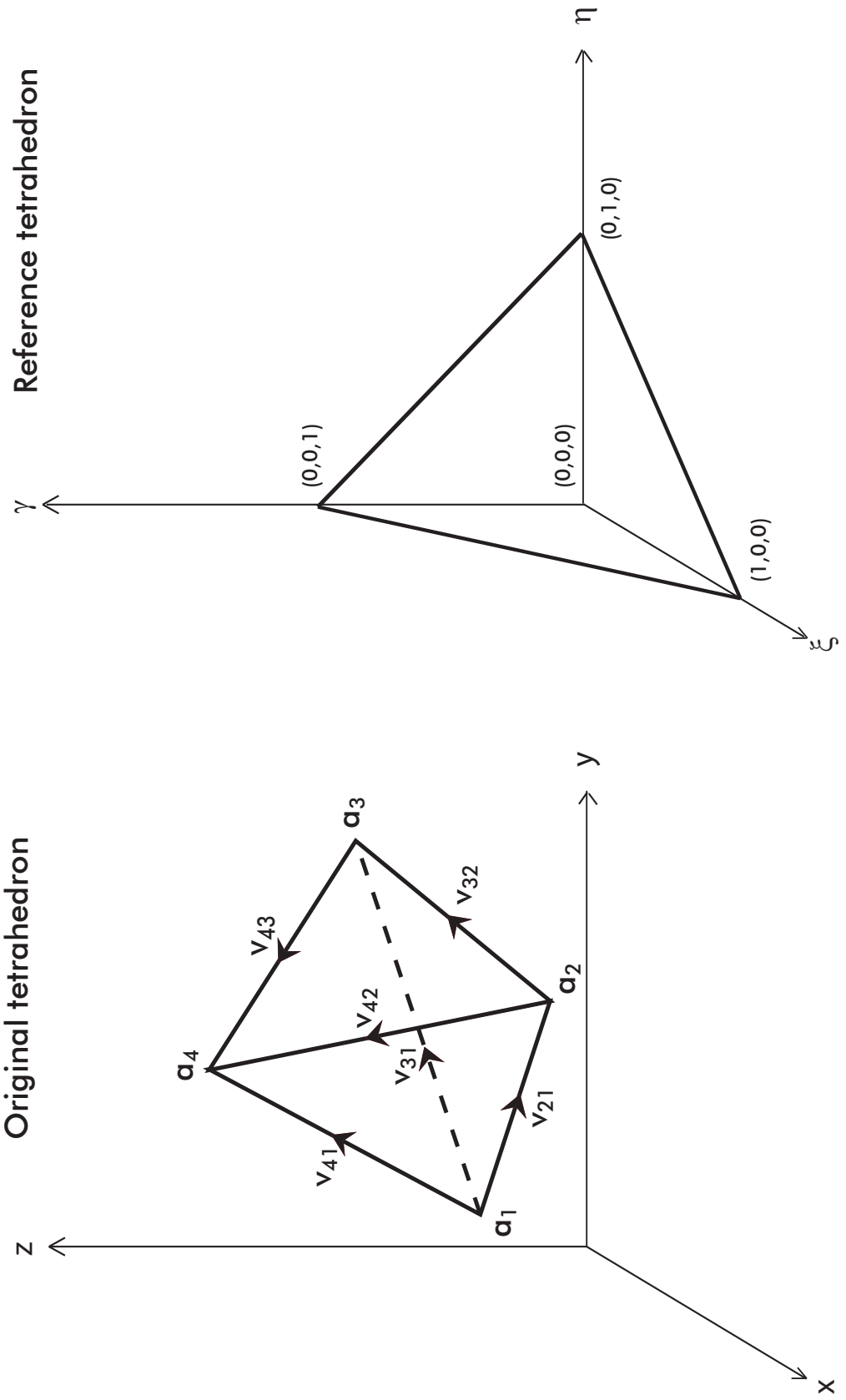

Figure A.1: Left: original tetrahedron; right: reference tetrahedron

A.1 Definitions of the basis functions and their gradient.

Let $\varphi_i(x, y, z) : \mathbb{R}^3 \to \mathbb{R}$ be the piecewise linear basis function in the original coordinates as defined in (4.6). Further, we denote the piecewise linear basis functions in the reference coordinate as $\tilde{\varphi}_i(\xi, \eta, \gamma) : \mathbb{R}^3 \to \mathbb{R}$, $i = 1, ..., 4$. They are defined as

$$\tilde{\varphi}_1(\xi, \eta, \gamma) := 1 - \xi - \eta - \gamma, \quad \tilde{\varphi}_2(\xi, \eta, \gamma) := \xi, \quad \tilde{\varphi}_3(\xi, \eta, \gamma) := \eta, \quad \tilde{\varphi}_4(\xi, \eta, \gamma) := \gamma. \tag{A.1}$$

For expressing the four vertices of the tetrahedron in the (x, y, z)-coordinates by the standard (ξ, η, γ)-coordinates, we consider the mapping $F : \mathbb{R}^3 \to \mathbb{R}^3$. In particular for $\boldsymbol{p} := (x, y, z)'$ define $\boldsymbol{p} = F(\tilde{\boldsymbol{p}})$ with $\tilde{\boldsymbol{p}} := (\xi, \eta, \gamma)'$, then F is given by,

$$F: \quad \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} x_2 - x_1 & x_3 - y_1 & x_4 - z_1 \\ y_2 - y_1 & y_3 - y_1 & y_4 - z_1 \\ z_2 - z_1 & z_3 - y_1 & z_4 - z_1 \end{pmatrix} \begin{pmatrix} \xi \\ \eta \\ \gamma \end{pmatrix} + \begin{pmatrix} x_1 \\ y_1 \\ z_1 \end{pmatrix} =: J\tilde{\boldsymbol{p}} + \begin{pmatrix} x_1 \\ y_1 \\ z_1 \end{pmatrix}$$

where J is the Jacobian matrix. The inverse of F is $\tilde{\boldsymbol{p}} = F^{-1}(\boldsymbol{p})$. Then the generic basis function in the original coordinates can be also written as $\varphi_i(\boldsymbol{p}) = \tilde{\varphi}_i(F^{-1}(\boldsymbol{p})) = \tilde{\varphi}_i(\tilde{\boldsymbol{p}})$. The gradient of the basis functions φ_i can be obtained using the chain rules, that is,

$$\nabla\varphi_i(\boldsymbol{p}) = \left(\frac{\partial F}{\partial \tilde{\boldsymbol{p}}}\right)^{-1} \nabla\tilde{\varphi}_i(F^{-1}(\boldsymbol{p})) := J^{-1}\nabla\tilde{\varphi}_i(\tilde{\boldsymbol{p}}). \tag{A.2}$$

It can also rewritten as

$$J = \Big[\boldsymbol{v}_{21}\, \boldsymbol{v}_{31}\, \boldsymbol{v}_{41}\Big]'$$

with directional vectors $\boldsymbol{v}_{21} := \boldsymbol{a}_2 - \boldsymbol{a}_1$, $\boldsymbol{v}_{31} := \boldsymbol{a}_3 - \boldsymbol{a}_1$ and $\boldsymbol{v}_{41} := \boldsymbol{a}_4 - \boldsymbol{a}_1$ in $\mathbb{R}^3$, as shown in Figure A.1. Furthermore, the inverse of the Jacobian matrix can be expressed using the cross products of vectors, i.e.

$$J^{-1} = \frac{1}{6|K|}\left[\boldsymbol{b}_2\ \boldsymbol{b}_3\ \boldsymbol{b}_4\right],$$

due to the identity $\det(J) = 6|K|$, where $|K|$ denote the volume of the tetrahedron K (Bronshtein et al., 2007). Here, $\boldsymbol{b}_i \in \mathbb{R}^3$ is the cross product of the two vectors that span the triangle opposite to the vertex i in a tetrahedron, $i = 1, ..., 4$, (Bronshtein et al., 2007). Further, we denote by K and $\tilde{K}$ the tetrahedra in the original and reference coordinates, respectively. By putting J^{-1} into (A.2) we then have

$$\nabla\varphi_i(\boldsymbol{p}) = \frac{1}{6|K|}\,[\boldsymbol{b}_2\ \boldsymbol{b}_3\ \boldsymbol{b}_4]\,\nabla\tilde{\varphi}_i(\boldsymbol{p}).$$

A.2 Calculation of the $\tilde{C}_{ii}$ element.

Recall that $\varphi_i := \varphi_i(x, y, z) : \mathbb{R}^3 \to \mathbb{R}$ and $\tilde{\varphi}_i := \tilde{\varphi}_i(\xi, \eta, \gamma) : \mathbb{R}^3 \to \mathbb{R}$ are the basis function in the original coordinates (x, y, z) and in the reference coordinates (ξ, η, γ), respectively, as defined in (A.1). The corresponding gradients of φ_i and $\tilde{\varphi}_i$ are $\nabla\varphi_i : \mathbb{R}^3 \to \mathbb{R}^3$ and $\nabla\varphi_i : \mathbb{R}^3 \to \mathbb{R}^3$, respectively. For the precision matrix Q_{ij} with one tetrahedron as shown in Figure A.1 we need the matrix $\tilde{C}$:

$$\tilde{C}_{ii} := \iiint\limits_K \varphi_i dz dy dx.$$

In particular, we have

$$\tilde{C}_{ii} = \iiint\limits_K \varphi_i\, dz dy dx = \iiint\limits_{\tilde{K}} \tilde{\varphi}_i \det(J)\, d\gamma d\eta d\xi = \int\limits_0^1 [\int\limits_0^{1-\xi} [\int\limits_0^{1-\xi-\eta} \tilde{\varphi}_i \det(J)\, d\gamma] d\eta] d\xi$$

For $i = 1$, this simplifies to $\tilde{C}_{ii} = \det(J) \int\limits_0^1 [\int\limits_0^{1-\xi} [\int\limits_0^{1-\xi-\eta} (1 - \xi - \eta - \gamma) d\gamma] d\eta] d\xi = \frac{|K|}{4}$. After similar algebraical calculation for $i = 2, 3, 4$ it follows that

$$\tilde{C} = \frac{|K|}{4} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

A.3 Calculation of the G_{ij} element.

Given the gradients of $\tilde{\varphi}$, i.e. $\nabla\tilde{\varphi}_1 = \nabla(1-\xi-\eta-\gamma)' = (-1,-1,-1)'$, $\nabla\tilde{\varphi}_2 = (1,0,0)'$, $\nabla\tilde{\varphi}_3 = (0,1,0)'$ and $\nabla\tilde{\varphi}_4 = (0,0,1)'$, the gradient of φ_i is $\nabla\varphi_i = J^{-1}\nabla\tilde{\varphi}_i$, $i = 1,...,4$, which simplifies to:

$$
\begin{aligned}
i=1: \quad & \nabla\varphi_1 = J^{-1}\nabla\tilde{\varphi}_1 = -\frac{1}{\det(J)}(\boldsymbol{b}_2+\boldsymbol{b}_3+\boldsymbol{b}_4) = -\frac{1}{\det(J)}(\boldsymbol{v}_{43}\times\boldsymbol{v}_{23}) = \frac{1}{\det(J)}\boldsymbol{b}_1 \\
& \text{with } \boldsymbol{b}_1 := -(\boldsymbol{v}_{43}\times\boldsymbol{v}_{23}) = \boldsymbol{v}_{43}\times\boldsymbol{v}_{32}, \\
i=2: \quad & \nabla\varphi_2 = J^{-1}\nabla\tilde{\varphi}_2 = \frac{1}{\det(J)}\boldsymbol{b}_2, \quad \boldsymbol{b}_2 := \boldsymbol{v}_{31}\times\boldsymbol{v}_{41}, \\
i=3: \quad & \nabla\varphi_3 = J^{-1}\nabla\tilde{\varphi}_3 = \frac{1}{\det(J)}\boldsymbol{b}_3, \quad \boldsymbol{b}_3 := \boldsymbol{v}_{41}\times\boldsymbol{v}_{21}, \\
i=4: \quad & \nabla\varphi_4 = J^{-1}\nabla\tilde{\varphi}_4 = \frac{1}{\det(J)}\boldsymbol{b}_4, \quad \boldsymbol{b}_4 := \boldsymbol{v}_{21}\times\boldsymbol{v}_{31}.
\end{aligned}
$$

Then, the G matrix is given by,

$$
\begin{aligned}
G_{ij} &:= \iiint_K \nabla\varphi_i'\nabla\varphi_j \, dzdydx \\
&= \iiint_{\tilde{K}} (J^{-1}\nabla\tilde{\varphi}_i)' \, (J^{-1}\nabla\tilde{\varphi}_j) \det(J) \, d\gamma d\eta d\xi = \frac{1}{\det(J)}\boldsymbol{b}_i'\boldsymbol{b}_j \iiint_{\tilde{K}} d\gamma d\eta d\xi.
\end{aligned}
$$

Since the term $\iiint_{\tilde{K}} d\xi d\eta d\gamma = \frac{1}{6}$ represents the volume of the standard reference tetrahedron, we have $G_{ij} = \frac{1}{36|K|}\boldsymbol{b}_i'\boldsymbol{b}_j$ or

$$
G = \frac{1}{36|K|}\begin{pmatrix}
\|\boldsymbol{b}_1\|^2 & \boldsymbol{b}_1'\boldsymbol{b}_2 & \boldsymbol{b}_1'\boldsymbol{b}_3 & \boldsymbol{b}_1'\boldsymbol{b}_4 \\
\boldsymbol{b}_2'\boldsymbol{b}_1 & \|\boldsymbol{b}_2\|^2 & \boldsymbol{b}_2'\boldsymbol{b}_3 & \boldsymbol{b}_2'\boldsymbol{b}_4 \\
\boldsymbol{b}_3'\boldsymbol{b}_1 & \boldsymbol{b}_3'\boldsymbol{b}_2 & \|\boldsymbol{b}_3\|^2 & \boldsymbol{b}_3'\boldsymbol{b}_4 \\
\boldsymbol{b}_4'\boldsymbol{b}_1 & \boldsymbol{b}_4'\boldsymbol{b}_2 & \boldsymbol{b}_4'\boldsymbol{b}_3 & \|\boldsymbol{b}_4\|^2
\end{pmatrix}.
$$

A.4 Calculation of the C_{ii} element.

The calculation of the matrix $C_{ij} := \langle \varphi_i, \varphi_j \rangle_\Omega$ can be done as follows:

$$\begin{aligned} C_{ij} &:= \langle \varphi_i, \varphi_j \rangle_\Omega \\ &= \iiint\limits_K \varphi_i \varphi_j \, dzdydx = \iiint\limits_{\tilde{K}} \tilde{\varphi}_i \tilde{\varphi}_j \det(J) \, d\gamma d\eta d\xi = \int\limits_0^1 \int\limits_0^{1-\xi} \int\limits_0^{1-\xi-\eta} \tilde{\varphi}_i \tilde{\varphi}_j \det(J) \, d\gamma d\eta d\xi. \end{aligned}$$

This implies that

$$C = \frac{|K|}{20} \begin{pmatrix} 2 & 1 & 1 & 1 \\ 1 & 2 & 1 & 1 \\ 1 & 1 & 2 & 1 \\ 1 & 1 & 1 & 2 \end{pmatrix}$$

holds.

A.5 Calculation of the B_{ij} element.

Let $\boldsymbol{r}(\boldsymbol{t})$ be a differentiable vector at point $\boldsymbol{t} \in \mathbb{R}^3$. Further, let $\boldsymbol{v}(\boldsymbol{t}) := \boldsymbol{r}'(\boldsymbol{t})$ be derivative of $\boldsymbol{r}(\boldsymbol{t})$ and $T(\boldsymbol{t}) := \boldsymbol{v}(\boldsymbol{t})/\|\boldsymbol{v}(\boldsymbol{t})\|$ be the unit tangent vector. Then, the unit normal vector is given by $\vec{\boldsymbol{n}} := \vec{\boldsymbol{n}}(\boldsymbol{t}) := T'(\boldsymbol{t})/\|T'(\boldsymbol{t})\|$.
Here, $\langle \cdot \rangle_{\partial\Omega}$ denotes the inner product of scalar- or vector-valued functions on the boundary $\partial\Omega$ of the domain Ω. Then, the element of the matrix B_{ij}, defined by

$$B_{ij} := \langle \varphi_i, \partial_n \varphi_j \rangle_{\partial\Omega} := \langle \varphi_i, (\nabla \varphi_j)' \overrightarrow{\boldsymbol{n}} \rangle_{\partial\Omega},$$

can be determined as

$$\begin{aligned} B_{ij} := \langle \varphi_i, (\nabla \varphi_j)' \overrightarrow{\boldsymbol{n}} \rangle_{\partial\Omega} &= \iint\limits_{\partial(K_i \cap K_j) \cap \partial\Omega} \varphi_i \, (\nabla \varphi_j)' \overrightarrow{\boldsymbol{n}}(\boldsymbol{s}_T) dS_T(\boldsymbol{s}) \\ &= \sum_{T_k \in \, (\partial(K_i \cap K_j) \cap \partial\Omega)} (\nabla \varphi_j)' \overrightarrow{\boldsymbol{n}}(\boldsymbol{s}_T) \iint\limits_{\partial(K_i \cap K_j) \cap \partial\Omega} \varphi_i dS_T(\boldsymbol{s}) \end{aligned}$$

$$= \sum_{T_k} (J^{-1}\nabla\tilde{\varphi}_j)'\overrightarrow{\boldsymbol{n}}(\boldsymbol{s}_T) \iint\limits_{\partial(K_i \cap K_j)\cap\partial\Omega} \varphi_i dS_T(\boldsymbol{s}),$$

where K_i denotes the tetrahedra that includes the vertex i and T_k is the triangle opposite to the vertex k for $k = 1, ..., 4$. Since we have

$$\iint\limits_{\partial(K_i \cap K_j)\cap\partial\Omega} \varphi_i dS_T(\boldsymbol{s}) = \frac{1}{3}|T_k| = \frac{\|\boldsymbol{b}_k\|}{6},$$

then B_{ij} can be written as

$$B_{ij} = \sum_{T_k} -\frac{1}{\det(J)}\boldsymbol{b}_j'\boldsymbol{b}_k\left(\frac{1}{3}\frac{\|\boldsymbol{b}_k\|}{2}\right) = -\frac{1}{36|K|}\sum_{T_k}\boldsymbol{b}_j'\boldsymbol{b}_k\|\boldsymbol{b}_k\|,$$

where $|T_k|$ denotes area of the triangle T_k. The vector $\boldsymbol{b}_k$ denote the cross product of two vectors that span the triangle $T_k \in \partial(K_i \cap K_j) \cap \partial\Omega$.
In the example with one tetrahedra as illustrated in Figure A.1, the triangle $T_4 \in (\partial(K_i \cap K_j) \cap \partial\Omega)$ is the plane opposite to the vertex a_4, then the contributions to the matrix B from one triangle T_4 (for $k = 4$), denoted as $B(T_4)$, is calculated as:

$$B(T_4)_{ij} = -\frac{1}{36\|K\|}\boldsymbol{b}_j'\boldsymbol{b}_4\|\boldsymbol{b}_4\|, \quad \text{for } i = 1, ..., 3 \text{ and } j = 1, ..., 4.$$

The elements $B(T_4)_{4j} = 0$ for all $j = 1, ..., 4$, since the vertex 4 is opposite to T_4 and therefore $\varphi_4 = 0$ on the triangle T_4. In matrix form, we have

$$B(T_4) = -\frac{\|\boldsymbol{b}_4\|}{36|K|}\begin{pmatrix} \boldsymbol{b}_1'\boldsymbol{b}_4 & \boldsymbol{b}_2'\boldsymbol{b}_4 & \boldsymbol{b}_3'\boldsymbol{b}_4 & \boldsymbol{b}_4'\boldsymbol{b}_4 \\ \boldsymbol{b}_1'\boldsymbol{b}_4 & \boldsymbol{b}_2'\boldsymbol{b}_4 & \boldsymbol{b}_3'\boldsymbol{b}_4 & \boldsymbol{b}_4'\boldsymbol{b}_4 \\ \boldsymbol{b}_1'\boldsymbol{b}_4 & \boldsymbol{b}_2'\boldsymbol{b}_4 & \boldsymbol{b}_3'\boldsymbol{b}_4 & \boldsymbol{b}_4'\boldsymbol{b}_4 \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \end{pmatrix}.$$

Similarly to the calculation of $B(T_4)$ we obtain the summands of B for T_k, $k = 1, ..., 4$, therefore the matrix B can be written as:

$$B = -\frac{1}{36|K|} \cdot \left(\|\boldsymbol{b}_1\| \begin{bmatrix} \mathbf{0}' \\ \boldsymbol{b}_1' \\ \boldsymbol{b}_1' \\ \boldsymbol{b}_1' \end{bmatrix} \mathbb{1}_{\{T_1 \in \partial\Omega\}} + \|\boldsymbol{b}_2\| \begin{bmatrix} \boldsymbol{b}_2' \\ \mathbf{0}' \\ \boldsymbol{b}_2' \\ \boldsymbol{b}_2' \end{bmatrix} \mathbb{1}_{\{T_2 \in \partial\Omega\}} + \|\boldsymbol{b}_3\| \begin{bmatrix} \boldsymbol{b}_3' \\ \boldsymbol{b}_3' \\ \mathbf{0}' \\ \boldsymbol{b}_3' \end{bmatrix} \mathbb{1}_{\{T_3 \in \partial\Omega\}} + \|\boldsymbol{b}_4\| \begin{bmatrix} \boldsymbol{b}_4' \\ \boldsymbol{b}_4' \\ \boldsymbol{b}_4' \\ \mathbf{0}' \end{bmatrix} \mathbb{1}_{\{T_4 \in \partial\Omega\}} \right) \cdot [\boldsymbol{b}_1 \ \boldsymbol{b}_2 \ \boldsymbol{b}_3 \ \boldsymbol{b}_4] \,.$$

The matrix can be also expressed as

$$B = -\frac{1}{36|K|} \begin{bmatrix} \mathbf{0}' & \boldsymbol{b}_2' & \boldsymbol{b}_3' & \boldsymbol{b}_4' \\ \boldsymbol{b}_1' & \mathbf{0}' & \boldsymbol{b}_3' & \boldsymbol{b}_4' \\ \boldsymbol{b}_1' & \boldsymbol{b}_2' & \mathbf{0}' & \boldsymbol{b}_4' \\ \boldsymbol{b}_1' & \boldsymbol{b}_2' & \boldsymbol{b}_3' & \mathbf{0}' \end{bmatrix} \begin{pmatrix} \|\boldsymbol{b}_1\| \mathbb{1}_{\{T_1 \in \partial\Omega\}} \\ \|\boldsymbol{b}_2\| \mathbb{1}_{\{T_2 \in \partial\Omega\}} \\ \|\boldsymbol{b}_3\| \mathbb{1}_{\{T_3 \in \partial\Omega\}} \\ \|\boldsymbol{b}_4\| \mathbb{1}_{\{T_4 \in \partial\Omega\}} \end{pmatrix} [\boldsymbol{b}_1 \ \boldsymbol{b}_2 \ \boldsymbol{b}_3 \ \boldsymbol{b}_4] \,.$$

Bibliography

Alberty, J., C. Carstensen, S. A. Funken, and R. Klose (2002). Matlab implementation of the finite element method in elasticity. *Computing 69*, 239 – 263.

Amestoy, P. R., T. A. Davis, and I. S. Duff (1996). An approximate minimum degree ordering algorithm. *SIAM Journal on Matrix Analysis and Applications 17*, 886 – 905.

Andrieu, C. and J. Thoms (2008). A tutorial on adaptive mcmc. *Statistics and Computing 18*, 343 – 373.

Arfken, G. (1985). *Mathematical Methods for Physicists*, Chapter 1.4, pp. 18 – 26. Orlando, FL: Academic Press.

Aso, T., B. Gustavsson, K. Tanabe, U. Brändström, T. Sergienko, and I. Sandahl (2008). A proposed Bayesian model on the generalized tomographic inversion of aurora using multi-instrument data. *IRF Sci. Rep. 292*, 105 – 109.

Aster, R., B. Borchers, and C. Thurber (2005). *Parameter Estimation and Inverse Problems (International Geophysics)* (Har/Cdr ed.). Academic Press.

Banerjee, S., A. E. Gelfand, and B. P. Carlin (2003). *Hierarchical Modeling and Analysis for Spatial Data. Monographs on Statistics and Applied Probability.* Chapman & Hall/CRC Monographs on Statistics & Applied Probability.

Banerjee, S., A. E. Gelfand, A. O. Finley, and H. Sang (2008). Gaussian predictive process models for large spatial data sets. *Journal of Royal Statistical Society: Series B 70*(4), 825 – 848.

Bardsley, J. M. (2012). Mcmc-based image reconstruction with uncertainty quantification. *SIAM Journal of Scientific Computing 34*(3), A1316 – A1332.

Bodin, T. and M. Sambridge (2009). Seismic tomography with the reversible jump algorithm. *Geophysical Journal International 178*(3), 1411 – 1436.

Bodin, T., M. Sambridge, N. Rawlinson, and P. Arroucau (2012). Transdimensional tomography with unknown data noise. *Geophysical Journal International 189*(3), 1536 – 1556.

Bodin, T., M. Sambridge, H. Tkalčić, P. Arroucau, K. Gallagher, and N. Rawlinson (2012). Transdimensional inversion of receiver functions and surface wave dispersion. *Journal of Geophysical Research 117*(B02301).

Bolin, D. (2012). Spatial Matérn fields driven by non-Gaussian noise. Technical report, Mathematical Statistics Centre for Mathematical Sciences, Lund University, Sweden. arXiv:1206.0622.

Bronshtein, I. N., K. A. Semendyayev, G. Musiol, and H. Mühlig (2007). *Handbook of Mathematics* (5 ed.)., Chapter 3, pp. 213. Springer.

Brooks, S., A. Gelman, G. Jones, and X.-L. Meng (Eds.) (2011). *Handbook of Markov Chain Monte Carlo (Chapman & Hall/CRC Handbooks of Modern Statistical Methods)* (1 ed.). Chapman and Hall/CRC.

Bui-Thanh, T., O. Ghattas, and D. Higdon (2011). Adaptive Hessian-based non-stationary Gaussian process response surface method for probability density approximation with application to Bayesian solution of large-scale inverse problems. Technical report, Texas University at Austin, Institut for computational engineering and sciences.

Burdick, S., C. Li, V. Martynov, T. Cox, J. Eakins, T. Mulder, L. Astiz, F. L. Vernon, G. L. Pavlis, and R. D. van der Hilst (2008). Upper mantle heterogeneity beneath North America from travel time tomography with global and USArray transportable array data. *Seismological Research Letters 79*(3), 389 – 392.

Cameletti, M., F. Lindgren, D. Simpson, and H. Rue (2012). Spatio-temporal modeling of particular matter concentration through the SPDE approach. *AStA Adv. Stat. Anal.*. DOI 10.1007/s10182-012-0196-3.

Celeux, G., F. Forbes, C. P. Robert, D. M. Titterington, I. Futurs, and I. Rhone-alpes (2006). Deviance information criteria for missing data models. *Bayesian Analysis 4*, 651 – 674.

Congdon, P. (2003). *Applied Bayesian Modelling* (1 ed.). Wiley.

Cressie, N. (1993). *Statistics for Spatial Data*. Wiley Series in Probability and Statistics.

Cui, T., C. Fox, and M. J. O'Sullivan (2011). Bayesian calibration of a large-scale geothermal reservoir model by a new adaptive delayed acceptance Metropolis Hastings algorithm. *Water Resources Research 47*(W10521), 26 pp.

Dahlen, F. A., S.-H. Hung, and G. Nolet (2000). Fréhet kernels for finite-frequency traveltimes-i. theory. *Geophysical Journal International 141*(1), 157 – 174.

Dębski, W. (2010). Seismic tomography by monte carlo sampling. *Pure and Applied Geophysics 167*(1), 131 – 152.

Dziewonski, A. M. and D. L. Anderson (1981). Preliminary reference Earth model. *Physics of the Earth and Planetary Interiors 25*, 297 – 356.

Fichtner, A., B. L. N. Kennett, H. Igel, and H.-P. Bunge (2009). Full waveform tomography for upper-mantle structure in the australasian region using adjoint methods. *Geophysical Journal International 179*, 1703 – 1725.

Flath, H. P., L. C. Wilcox, V. Akcelik, J. Hill, B. van Bloemen Waanders, and O. Ghattas (2011). Fast algorithms for Bayesian ucertainty quantification in large-scale linear inverse problems based on low-rank partial Hessian approximations. *SIAM Journal on Scientific Computing 33*, 407 – 432.

Frühwirth-Schnatter, S. and S. Pyne (2010). Bayesian inference for finite mixtures of univariate and multivariate skew-normal and skew-t distributions. *Biostatistics 11*(2), 317 – 336.

Furrer, R., R. Knutti, S. R. Sain, D. W. Nychka, and G. A. Meehl (2007). Spatial patterns of probabilistic temperature change projections from a multivariate Bayesian analysis. *Geophys. Res. Lett. 34*(6), L06711.

Gockenbach, M. S. (2006). *Understanding and Implementing the Finite Element Method*. SIAM.

Haario, H., M. Laine, A. Mira, and E. Saksman (2006). Dram: Efficient adaptive mcmc. *Statistics and Computing 16*, 339 – 354.

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

Hämäläinen, M. S. and R. J. Ilmoniemi (1994). Interpreting magnetic fields of the brain: minimum norm estimates. *Medical and Biolgical Engineering and Computing 32*(1), 35 – 42.

Harrison, L. M. and G. G. Green (2010). A Bayesian spatiotemporal model for very large data sets. *Neuroimage 50*(3), 1126 – 41.

Herbei, R., I. W. McKeague, and K. G. Speer (2008). Gyres and jets: Inversion of tracer data for ocean circulation structure. *Journal of Physical Oceanography 38*, 1180 – 1202.

Hoerl, E. and R. W. Kennard (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics 12*(1), 55 – 67.

Hughes, T. J. R. (1987). *The Finite Element Method. Linear Static and Dynamic Finite Element Analysis*. Englewood Cliffs, New Jersey: Prentice-Hall.

Illian, J. B., S. H. Sørbye, and H. Rue (2012). A toolbox for fitting complex spatial point process models using integrated nested Laplace approximation (INLA). *Annals of Applied Statistics 6*(4), 1499 – 1530.

Kaipio, J. and E. Somersalo (2007). Statistical inverse problems: Discretization, model reduction and inverse crimes. *Journal of Computational and Applied Mathematics 198*, 493 – 504.

Kass, R. E., B. P. Carlin, A. Gelman, and R. M. Neal (1998). Markov Chain Monte Carlo in practice: A roundtable discussion. *The American Statistician 52*(2), 93 – 100.

Kennett, B. L. N. and E. R. Engdahl (1991). Traveltimes for global earthquake location and phase identification. *Geophysical Journal International 105*, 429 – 465.

Khan, A., A. Zunino, and F. Deschamps (2011). The thermo-chemical and physical structure beneath the North American continent from Bayesian inversion of surface-wave phase velocities. *Journal of Geophysical Research 116*(B09304), 23pp.

Kolehmainen, V., A. Vanne, S. Siltanen, S. Järvenpää, J. P. Kaipio, M. Lassas, and M. Kalke (2007). Bayesian inversion method for 3D dental X-ray imaging. *e & i Elektrotechnik und Informationstechnik 124*, 248–253.

Lieberman, C., K. Willcox, and O. Ghattas (2010). Parameter and state model reduction for large-scale statistical inverse problems. *SIAM Journal on Scientific Computing 32*(5), 2523 – 2542.

Lindgren, F. (2012). Continuous domain spatial models in R-INLA. *The ISBA Bulletin 19*(4).

Lindgren, F., H. Rue, and J. Lindström (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the SPDE approach (with discussion). *Journal of Royal Statistical Society: Series B 73*(4), 423 – 498.

Liu, J. S. (2008). *Monte Carlo Strategies in Scientific Computing*. Springer.

Martin, J., L. C. Wilcox, C. Burstedde, and O. Ghattas (2012). A stochastic Newton MCMC method for large-scale statistical inverse problems with application to seismic inversion. *SIAM Journal on Scientific Computing 34*(3), A1460 – A1487.

Mercerat, E. D. and G. Nolet (2013). On the linearity of cross-correlation delay times in finite-frequency tomography. *Geophysical Journal International 192*(2), 681 – 687.

Mohammad-Djafari, A. (2012). Bayesian approach with prior models which enforce sparsity in signal and image processing. *EURASIP Journal on Advances in Signal Processing 2012*(52).

Möller, A., A. Lenkoski, and T. L. Thorarinsdottir (2012). Multivariate probabilistic forecasting using Bayesian model averaging and copulas. *Quarterly Journal of the Royal Meteorological Society*.

Mosca, I., L. Cobden, A. Deuss, J. Ritsema, and J. Trampert (2012). Seismic and mineralogical structures of the lower mantle from probabilistic tomography. *Journal of Geophysical Research 117*(B06304), 26pp.

Mosegaard, K. and A. Tarantola (1995). Monte Carlo sampling of solutions to inverse problems. *Journal of Geophysical Research 100*(B7), 12431 – 12447.

Mosegaard, K. and A. Tarantola (2002). *International Handbook of Earthquake and Engineering Seismology*, Chapter Probabilistic Approach to Inverse Problems, pp. 237 – 265. Academic Press.

Neal, R. M. (2010). Mcmc using hamiltonian dynamics. In S. Brooks, A. Gelman, G. Jones, and X.-L. Meng (Eds.), *Handbook of Markov Chain Monte Carlo*, Chapter 5. Chapman & Hall/CRC Press.

Nolet, G. (1987). Seismic wave propagation and seismic tomography. In G. Nolet (Ed.), *Seismic Tomography*, pp. 1 – 24. Dordrecht: D. Reidel Publishing Company.

Nolet, G. (2008). *A Breviary of Seismic Tomography: Imaging the Interior of the Earth and Sun*. Cambridge University Press.

Paige, C. C. and M. A. Saunders (1982). LSQR: An algorithm for sparse linear equations and sparse least squares. *TOMS 8*(1), 43–71.

Pettitt, N. A., I. S. Weir, and A. G. Hart (2002). A conditional autoregressive Gaussian process for irregularly spaced multivariate data with application to modelling large sets of binary data. *Statistics and Computing 12*, 353 – 367.

Rue, H. and L. Held (2005). *Gaussian Markov Random Fields: Theory And Applications (Monographs on Statistics and Applied Probability)*. Chapman & Hall/CRC.

Rue, H. and S. Martino (2007). Approximate Bayesian inference for hierarchical Gaussian Markov random fields models. *Journal of Statistical Planning and Inference 137*(10), 3177 – 3192. Special Issue: Bayesian Inference for Stochastic Processes.

Rue, H., S. Martino, and N. Chopin (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B 71*(2), 319 – 392.

Sahu, S. K., D. K. Dey, and M. D. Branco (2003). A new class of multivariate skew distributions with applications to Bayesan regression models. *The Canadian Journal of Statistics 31*(2), 129 – 150.

Sambridge, M. and O. Gudmundsson (1998). Tomographic systems of equations with irregular cells. *Journal of Geophysical Research 103*(B1), 773 – 781.

Sambridge, M. and K. Mosegaard (2002). Monte Carlo methods in geophysical inverse problems. *Reviews of Geophysics 40*(3), 1 – 29.

Schmandt, B. and E. Humphreys (2010). Complex subduction and small-scale convection revealed by body-wave tomography of the western United States upper mantle. *Earth and Planetary Science Letters 297*(3-4), 435 – 445.

Schrödle, B. and L. Held (2011). Spatio-temporal disease mapping using INLA. *Environmetrics 22*(6), 725 – 734.

Sieminski, A., Q. Liu, J. Trampert, and J. Tromp (2007). Finite-frequency sensitivity of body waves to anisotropy based upon adjoint methods. *Geophysical Journal International 171*, 368 – 389.

Sigloch, K. (2008). *Multiple-frequency body-wave tomography.* Ph. D. thesis, Princeton University.

Sigloch, K. (2011). Mantle provinces under North America from multi-frequency P-wave tomography. *Geochemistry, Geophysics, Geosystems 12*(Q02W08).

Sigloch, K., N. McQuarrie, and G. Nolet (2008). Two-stage subduction history under North America inferred from finite-frequency tomography. *Nature GEO 1*, 458 – 462.

Sigloch, K. and G. Nolet (2006). Measuring finite-frequency body-wave amplitudes and traveltimes. *Geophysical Journal International 167*, 271 – 287.

Simpson, D. P., J. B. Illian, F. Lindgren, S. H. Sørbye, and H. Rue (2011). Going off grid: Computationally efficient inference for log-Gaussian Cox processes. Preprint statistics no.10/2011, Department of Mathematical Sciences, Norwegian University of Science and Technology.

Simpson, D. P., F. Lindgren, and H. Rue (2011). In order to make spatial statistics computationally feasible, we need to forget about the covariance function. *Environmetrics 23*(1), 65 – 74. Special Issue: Spatio-Temporal Stochastic Modelling.

Simpson, D. P., F. Lindgren, and H. Rue (2012). Think continuous: Markovian Gaussian models in spatial statistics. *Spatial Statistics 1*, 16 – 29.

Simpson, D. P., I. W. Turner, and A. N. Pettitt (2008). Fast sampling from a Gaussian Markov random field using Krylov subspace approaches. QUT ePrints ID 14376, Queensland Univ. of Technology, Brisbane, Australia.

Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. van der Linde (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B 64*(4), 583 – 639.

Stein, M. L. (1999). *Interpolation of Spatial Data.* New York: Springer-Verlag.

Stramer, O. and R. L. Tweedie (1999). Langevin-Type Models II: Self-Targeting Candidates for MCMC Algorithms. *Methodology and Computing in Applied Probability 1*(3), 307 – 328.

Swindel, B. F. (1976). Good ridge estimators based on prior information. *Communications in Statistics - Theory and Methods 5*(11), 1065 – 1075.

Tarantola, A. (2004). *Inverse Problem Theory and Methods for Model Parameter Estimation* (1 ed.). SIAM: Society for Industrial and Applied Mathematics.

Ter Braak, C. J. F. (2006). A Markov chain Monte Carlo version of the genetic algorithm differential evolution: easy Bayesian computing for real parameter spaces. *Statistics and Computing 16*(3), 239 – 249.

Tian, Y., R. Montelli, G. Nolet, and F. A. Dahlen (2007). Computing traveltime and amplitude sensitivity kernels in finit-frequency tomography. *Journal of Computational Physics 226*, 2271 – 2288.

Tian, Y., K. Sigloch, and G. Nolet (2009). Multiple-frequency SH-wave tomography of the western US upper mantle. *Geophysical Journal International 178*(3), 1384 – 1402.

Tierney, L. and J. B. Kadane (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association 81*(393), 82 – 86.

Tromp, J., C. Tape, and Q. Liu (2005). Seismic tomography, adjoint methods, time reversal and banana-doughnut kernels. *Geophysical Journal International 160*(1), 195– 216.

Uutela, K., M. Häämäläinen, and E. Somersalo (1999). Visualization of magnetoencephalographic data using minimum current estimates. *NeuroImage 10*, 173 – 180.

Vrugt, J. A., C. J. F. ter Braak, C. G. H. Diks, B. A. Robinson, J. M. Hyman, and D. Higdon (2009). Accelerating Markov chain Monte Carlo simulation by differential evolution with self-adaptive randomized subspace sampling. *International Journal of Nonlinear Sciences & Numerical Simulation 10*(3), 273 – 290.

Watzenig, D. and C. Fox (2009). A review of statistical modelling and inference for electrical capacitance tomography. *Measurement Science and Technology 20*(052002), 22pp.

Whittle, P. (1954). On stationary processes in the plane. *Biometrika 41*, 434 – 449.

Whittle, P. (1963). Stochastic processes in several dimensions. *Bull. Inst. Internat. Statist. 40*, 974 – 994.

Wilkinson, D. J. and S. K. H. Yeung (2002). Conditional simulation from highly structured gaussian systems, with application to blocking-MCMC for the Bayesian analysis of very large linear models. *Statistics and Computing 12*(3), 287–300.

Zhang, R., C. Czado, and K. Sigloch (2013a). A Bayesian linear model for the high-dimensional inverse problem of seismic tomography. *Annals of Applied Statistics 7*(2), 1111 – 1138.

Zhang, R., C. Czado, and K. Sigloch (2013b). Bayesian spatial modelling for high-dimensional seismic inverse problems. Submitted to Royal Journal of Statistical Society: Series C.