Hierarchical Bayesian Approaches to Seismic Imaging and Other Geophysical Inverse Problems

by

Sam Ahmad Zamanian

B.S., Biomedical Engineering
Johns Hopkins University (2005)

S.M., Electrical Engineering and Computer Science
Massachusetts Institute of Technology (2007)

Electrical Engineer
Massachusetts Institute of Technology (2014)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2014

Author . . . .
Department of Electrical Engineering and Computer Science
August 13, 2014

Certified by . . . .
Michael C. Fehler
Senior Research Scientist
Thesis Supervisor

Accepted by . . . .
Leslie A. Kolodziejski
Professor of Electrical Engineering and Computer Science
Chair, Department Committee on Graduate Students

In loving memory of my late mother,

Farideh Payandeh Zamanian,

and,

to my father,

Ahmad Zamanian.

Hierarchical Bayesian Approaches to Seismic Imaging and Other Geophysical Inverse Problems

by

Sam Ahmad Zamanian

Submitted to the Department of Electrical Engineering and Computer Science
on August 13, 2014, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

In many geophysical inverse problems, smoothness assumptions on the underlying geologic model are utilized to mitigate the effects of poor data coverage and observational noise and to improve the quality of the inferred model parameters. In the context of Bayesian inference, these smoothness assumptions take the form of a prior distribution on the model parameters. Conventionally, the regularization parameters defining these assumptions are fixed independently from the data or tuned in an *ad hoc* manner. However, it is often the case that the smoothness properties of the true earth model are not known *a priori*, and furthermore, these properties may vary spatially. In the seismic imaging problem, for example, where the objective is to estimate the earth's reflectivity, the reflectivity model is smooth along a particular reflector but exhibits a sharp contrast in the direction orthogonal to the reflector. In such cases, defining a prior using predefined smoothness assumptions may result in posterior estimates of the model that incorrectly smooth out these sharp contrasts.

In this thesis, we explore the application of Bayesian inference to different geophysical inverse problems and seek to address issues related to smoothing by appealing to the hierarchical Bayesian framework. We capture the smoothness properties of the prior distribution on the model by defining a Markov random field (MRF) on the set of model parameters and assigning weights to the edges of the underlying graph; we refer to these parameters as the *edge strengths* of the MRF. We investigate two cases where the smoothing is specified *a priori* and introduce a method for estimating the edge strengths of the MRF.

In the first part of this thesis, we apply a Bayesian inference framework (where the edge strengths of the MRF are predetermined) to the problem of characterizing the fractured nature of a reservoir from seismic data. Our methodology combines different features of the seismic data, particularly P-wave reflection amplitudes and scattering attributes, to allow for estimation of fracture properties under a larger physical regime than would be attainable using only one of these data types. Through this application, we demonstrate the capability of our parameterization of the prior distribution with edge strengths to both enforce smoothness in the estimates of the fracture properties and capture *a priori* information about geological features in the model (such as a discontinuity that may arise in the presence of a fault). We solve the inference problem via loopy belief propagation to approximate the posterior marginal distributions of the fracture properties, as well as their *maximum a posteriori* (MAP) and Bayes least squares estimates.

In the second part of the thesis, we investigate how the parameters defining the prior distribution are connected to the model covariance and address the question of how to optimize these parameters in the context of the seismic imaging problem. We formulate the seismic imaging problem within the hierarchical Bayesian setting, where the edge strengths are treated as random variables to be inferred from the data, and provide a framework for computing the marginal MAP estimate of the edge strengths by application of the expectation-maximization (E-M) algorithm. We validate our methodology on synthetic datasets arising from 2-D models. The images we obtain after inferring the edge strengths exhibit the desired spatially-varying smoothness properties and yield sharper, more coherent reflectors.

In the final part of the thesis, we shift our focus and consider the problem of time-lapse seismic processing, where the objective is to detect changes in the subsurface over a period of time using repeated seismic surveys. We focus on the realistic case where the surveys are taken with differing acquisition geometries. In such situations, conventional methods for processing time-lapse data involve inverting surveys separately and subtracting the inversion models to estimate the change in model parameters; however, such methods often perform poorly as they do not correctly account for differing model uncertainty between surveys due to differences in illumination and observational noise. Applying the machinery explored in the previous chapters, we formulate the time-lapse processing problem within the hierarchical Bayesian setting and present a framework for computing the marginal MAP estimate of the time-lapse change model using the E-M algorithm. The results of our inference framework are validated on synthetic data from a 2-D time-lapse seismic imaging example, where the hierarchical Bayesian estimates significantly outperform conventional time-lapse inversion results.

Thesis Supervisor: Michael C. Fehler
Title: Senior Research Scientist

Acknowledgments

There are a countless number of people (actually countable and finite) to whom I owe a great a deal of gratitude for their support during the progress of my doctoral studies.

I would like to thank my thesis advisor, Mike Fehler, who has truly been an amazing mentor during these four years at MIT's Earth Resources Laboratory. Mike was generous enough to take me on as a graduate student from a different department and has been both very patient and encouraging as I developed my knowledge of geophysics and searched for problems in the field that would both interest and challenge me. Often times, due to our different technical backgrounds, Mike and I would tend to think about problems from different perspectives, but this has made our technical discussions very enjoyable and all the more insightful. As an advisor, Mike really has gone above and beyond the call of duty, and it is obvious he truly enjoys working with his students; beyond our technical discussions, our conversations about work life, student life, and family life (among other lives) have always been enjoyable and have taught me a lot about life in general. Thank you, Mike; I truly could not have asked for a better advisor.

I am much indebted to Bill Rodi who, in addition to serving on my doctoral committee, has been like an informal co-advisor. Sometimes five days a week (and many times, seven), I would bump into Bill in the ERL coffee room, and what would begin as an informal chat about some topic in inverse theory (or some other mathematical point) would often spiral into an all-out technical discussion on the whiteboard (conveniently located right next to the coffee room). Bill's excitement about the problems we would discuss was always tangible (and often evidenced by the fact that he was willing to sacrifice his coffee break, plus usually another hour that I'm sure he had scheduled for some other task, to discuss them). He really is a wealth of knowledge at ERL in the field of geophysical inverse theory, and I learned a great deal from him in our numerous discussions. In addition to his technical expertise, his sense of humor is also a force to be reckoned with.

I am also very grateful for the encouragement and suggestions provided by the other members of my doctoral committee: Bill Freeman, Alan Willsky, and Jonathan Kane. I benefited immensely from Bill and Alan's vast expertise in Bayesian inference and graphical models. Alan has made himself available for technical advice on multiple occasions from an early point in my graduate student career. It was he who suggested early on that I study probabilistic graphical models more deeply to better motivate my research in geophysical inverse problems. He subsequently served on my RQE committee and made useful suggestions on the research presented therein. His advice undoubtedly helped guide my studies in a direction that led to the bulk of the research presented in this thesis. Jonathan Kane, a fellow Bayesian geophysicist, provided many helpful suggestions, including the suggestion to explore an extension of the least-squares migration problem. I must also thank Jonathan for introducing me to the people involved in research at Shell, which has helped open doors allowing me to take the next step in my research career.

I am also thankful to my collaborators, Dan Burns, Xinding Fang, and Di Yang. Dan provided useful advice for the fracture characterization work of Chapter 3, and Xinding provided the synthetic data used for testing the method developed in this chapter. Di provided the acoustic data used in Chapter 5 and also helped motivate the idea for the time-lapse study in Chapter 6. I also benefited greatly from my discussions with other faculty and research staff at MIT, including Alison Malcolm, Oleg Poliannikov, Tianrun Chen, Yingcai Zheng, Dale Morgan, and Steve Brown of ERL, Devavrat Shah of LIDS, and John Fisher of CSAIL. In addition, various discussions with Michael Prange of Schlumberger and Anu Chandran, Henning Kuehl, Vanessa Goh, and Ken Matson of Shell have been very beneficial.

During my PhD, I was fortunate enough to have the opportunity to twice TA 6.041 (Probabilistic Systems Analysis), both times under John Tsitsiklis. I learned a great deal from John, both from his style of teaching and from his approach to probability and mathematics. My experience as a TA was certainly an unforgettable one and was made all the more unforgettable by the good times shared with my fellow 6.041 TA's: Aliaa Atwi, Uzoma Orji, and Shashank Dwivedi.

I would be remiss if I did not acknowledge the support of George Verghese and Thomas Heldt, who had advised me when I was a master's student and continued to provide encouragement and advice during my doctoral studies. George, in particular, generously offered his advice both as I was preparing to return to MIT to complete my PhD and as I was preparing for the RQE.

Thanks are due to Sue Turbak of ERL for being a wonderful administrator (and, as she would put it, "lab mom") as well as to the new ERL administrator, Natalie Counts. I am also grateful to Anna Shaughnessy for her efforts in keeping ERL and its graduate students well-connected to members of the energy industry. I also thank Janet Fischer of the EECS Graduate Office for her assistance and guidance throughout the doctoral program.

My friends at ERL have made my time there extremely enjoyable and have helped make bearable even the most adversarial slumps in my research. It is hard to imagine how I would have been able to complete these four years without their companionship and camaraderie. During my stay at ERL, I've had the pleasure of keeping the good company of Sudhish Bakku, Di Yang, Alan Richardson, Nasruddin Nazerali, Hua Wang, Andrey Shabelansky, Xinding Fang, Abdulaziz AlMuhaidib, Junlin Li, Fuxian Song, Bram Willemsen, Diego Concha, Bongani Mashele, Greg Ely, Jared Atkinson, Saleh Al Nasser, Haijing Zhang, Gabi Melo Silva, Yulia Agramakova, Ruel Jerry, Xuefang Shang, Chunquan Yu, Hussam Busfar, Mirna Slim, Leonardo Zepada, Beatrice Parker, Chen Gu, Jing Liu, Nate Dixon, Scott Burdick, Noa Bechor Ben Dov, Farrokh Sheibani, Haoyue Wang, Yuval Tal, Martina Coccia, Manuel Torres, and Eva Golos. I have thoroughly enjoyed my many late night lab dinners with Sudhish and his fiancée Sara (which would often times include home-cooked food that Sara would generously bring to the lab). My numerous discussions with Sudhish, Di, Alan, and Nasruddin about politics, culture, gastronomy, and philosophy (and, of course, geophysics) have always been very enjoyable. I will never forget the final month of my PhD studies (which incidentally occurred at the same time as Ramadan), every night of which was spent writing this dissertation; both Nasruddin and Sudhish were also immersed in their research at this time, hence also staying up well into the night, and they both faithfully kept me company during those many long nights of writing. The 3 AM breakfasts Nasruddin and I shared that month will always remain a fond memory.

The time I spent at MIT and in the Boston area has allowed me to make many other good friends here, including Kashif Sheikh, Omair Saadat, Ebraheim Ismail, Mahdi Ghassemi, Asad Lodhia, Hamza Fawzi, and Faisal Kashif. My discussions with Asad about stochastic PDEs and with Hamza about convex optimization have helped me gain further insight into my research. I'd particularly like to thank Faisal, not only for his warm friendship, but also for providing valuable guidance as a senior student at multiple points along my journey in graduate school. I'd also like to thank my good friends from my undergraduate days at Johns Hopkins for their continued friendship and support (including Nabil Rab, Bilal Farooqi, Sameer Ahmed, Zain Syed, Abdulrasheed Alabi, Safi Shareef, Usman Zaheer, Nurain Fuseini, and Imad Qayyum); one can be certain he has found true friends when they continue to call after almost a decade.

Now I come to the point where I mention those nearest and dearest to me. My father, Ahmad Zamanian, has always been a source of inspiration in my life. Also an engineer, he intrigued me from a young age with his technical prowess; I still remember being perplexed by the math puzzles he would serve me when I was still in elementary school. Even today, when we talk on the phone, every now and then he might still relate to me new math puzzles (but now I can at least solve them half the time as well as sometimes respond with a puzzle of my own). His love and encouragement have helped me reach where I am today; he remains an inspiration to me in all aspects of life, and I continue to learn from him every time we talk. My beloved late mother, Farideh Payandeh Zamanian, whom I miss dearly, has been a constant support throughout my life. I am the person I am today because of her selfless love and support. In every point of my life, she had always pushed me to exceed what I thought were my limits and taught me not to place barriers on my potential. Her memory will always remain dear in my heart, and I dedicate this thesis to her and to my father. I am also grateful to my sisters, Neeki and Donna, for their lifelong friendship and support. I also thank Grazyna and her family for their kind support during my PhD.

I cannot begin to express my thanks to my beloved wife, Sabah, who has made a tremendous sacrifice by joining me on this arduous journey. She has put up with my numerous all-nighters and, due to my lack of availability, has often borne alone much of the responsibility of taking care of our family. Her dedication, love, and support have helped me persevere in my studies, even through the most difficult times. Thank you, Sabah, for everything you've done for me and our family; I look forward to our continued journey together as we begin the next chapter of our lives. My two daughters, Maryam and Zaynab, have been a source of immense joy and are the light of my life. Seeing them always brings a smile to my face, and I thank God for these precious gifts. They have certainly been a strong motivation for me to push myself to find the light at the end of the tunnel and finish my PhD. While Zaynab is likely too young to remember her time here, I hope Maryam will recall fond memories of her childhood living at Westgate on the MIT campus. I would also like to thank my mother-in-law, Zahida, for her support (and for coming to MIT to help us with baby Zaynab when she was born).

The research presented in this thesis was supported by the MIT Energy Initiative, Shell International E&P Inc., and Total S.A.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

A fundamental issue in the field of geophysics is the problem of inferring physical properties of the earth from data collected at or near the earth's surface. The laws of physics generally provide a relationship between the data and the earth properties of interest (which we refer to as the earth model), and the problem of computing these data given a particular earth model is referred to as the *forward problem*. By contrast, the *inverse problem* is the task of estimating these properties given a particular set of measurements [69]. Examples of geophysical inverse problems include, amongst others, electromagnetic inversion [33, 34, 47, 54, 57], inversion of gravity data for density [38, 40, 52, 75], seismic traveltime tomography [15, 26, 51, 79, 80, 83], and reflection seismic imaging [5, 11, 12, 19, 29, 36, 37, 39, 49].

In these and other geophysical inverse problems, the data are often either insufficient to fully constrain the earth model or are corrupted with observational noise. This can be problematic because incorrect models may fit the data as well as (or, in the case of noisy data, perhaps better than) the true earth model. This problem can be averted by introducing *a priori* information about the model into the inversion procedure. For example, one might expect that the model parameters do not vary rapidly in space, and hence smoother models would be given preference over those which exhibit sharp contrasts. In the context of Bayesian inference, which can be viewed as a probabilistic framework for inversion, this is achieved by considering the model parameters as random variables with a prior probability distribution which captures one's belief about the model prior to observing the data.

The choice of the prior distribution is not a trivial one and can significantly impact the posterior estimates of the model obtained from the Bayesian inference framework. A prior distribution that prefers smooth models, for example, may be inappropriate if the true earth model does indeed contain sharp discontinuities, and using such a prior may result in a relatively low posterior probability being assigned to the true model. Furthermore, we often encounter situations where the model may exhibit spatially-varying smoothness properties that may not be known *a priori*. For example, in the seismic imaging problem, where the objective is to obtain an image of the earth's reflectivity from seismic data, the model (i.e. the seismic image) will inherently have spatially-varying smoothness properties: the image is smooth along a particular reflector but exhibits a sharp contrast in the direction orthogonal to the reflector. However, neither the location nor the orientation of the reflectors are known *a priori* (if they were, we would not be inferring them).

In this thesis, we are concerned with the question of how to learn the optimal prior from the data in such a setting, where we focus particularly on the seismic imaging problem described above and the associated smoothness properties of the seismic image. While the idea of learning an optimal prior from the data may seem to run counter to the Bayesian philosophy (since, after all, the prior distribution is intended to capture one's state of belief prior to observing the data), the problem can still be formulated within the Bayesian setting in what is known as the *hierarchical Bayesian* framework. As the name suggests, here a hierarchy of random variables is introduced by treating the parameters defining the original prior distribution (of the earth model) as random variables themselves, endowed with their own prior distribution; i.e., we place a prior on the prior. These new random variables can then be inferred from the data and used to define a prior for the earth model in the original inference problem. This particular approach of estimating the prior distribution from the data is sometimes referred to as the *empirical Bayes* method [45].

Even as we seek to let the data dictate an optimal prior, there remain problem-specific design choices to be made. In particular, we must choose how to parameterize the prior on the model in a meaningful way that captures its spatially-varying smoothness properties. We address this by defining a Markov random field (MRF) on the model vector and parameterizing the edges of its underlying graphical model (which, as will be seen in Chapter 2, is a 2-D grid graph); we refer to these edge parameters as *edge strengths*, which we define precisely in Chapter 2.

1.1 Thesis Outline and Summary of Contributions

We note that the chapters of this thesis have been written in such a way that they are arranged in self-contained units, so the interested reader is able to go straight to the unit of interest. In particular, Chapter 2, Chapter 3, and Chapter 6 are each self-contained units, and Chapters 4-5 form a self-contained unit. Due to arranging the thesis in this format, the reader may find a small portion of the background sections in some chapters to be slightly redundant. Below we give a brief outline of the thesis along with our main contributions.

Chapter 2 : A Brief Primer on Inverse Problems, Bayesian Inference, and Graphical Models

In Chapter 2, we give a brief tutorial on inverse problems and Bayesian inference as well as introduce, by way of example, the form of the smoothness enforcing prior distribution on the model. We begin the chapter by discussing the deterministic formulation of regularized inversion. We then proceed to the probabilistic formulation of Bayesian inference and show how regularized inversion can be viewed as a special case of Bayesian inference. Lastly we review the concept of a probabilistic graphical model and show how we can obtain a meaningful interpretation of the prior distribution of the model through this framework. Here we define the concept of an edge strength and show numerically how different choices for the edge strengths can affect the prior covariance of the model.

Chapter 3: Bayesian Fracture Characterization

Chapter 3 serves as our first study into the application of Bayesian inference methods to a geophysical inverse problem. In this chapter, we develop the application of a (non-hierarchical) Bayesian framework to the problem of fracture characterization from seismic data. Here the model consists of the fracture properties (particularly the fracture orientation and excess compliance) of a 2-D reservoir that is localized in depth, and the measured data are taken to be features extracted from the seismic traces, particularly the P-wave amplitude variation with offset and azimuth [42, 60, 62, 63] and the fracture transfer function [22], which measures the change in the scattered seismic energy after the seismic wavefield passes through the reservoir. When fractures are closely spaced relative to the seismic wavelength, the fractured medium tends to exhibit anisotropy [66], and hence the P-wave reflection amplitude data are more sensitive to the fracture properties at small fracture spacings. On the other hand, when the fracture spacing is on the order of the seismic wavelength, the fractures tend to instead act as scatterers [22, 78], thus the scattered seismic energy is more sensitive to the presence and orientation of fractures in this regime of fracture spacings. The Bayesian framework allows us to combine these data to give estimates of the fracture properties over a larger regime of fracture spacings than would otherwise be attainable while also providing a measure of uncertainty in the estimates. We derive the likelihood models for these data via physical models for anisotropy [60, 66] and fracture scattering [22, 78]. The fracture properties are modeled as discretely-valued random variables with a prior distribution described by the same 2-D grid MRF introduced in Chapter 2. We solve the inference problem via loopy belief propagation [48, 55, 56] to obtain the posterior marginal distributions of the fracture properties, as well as their *maximum a posteriori* (MAP) and Bayes least squares (posterior mean) estimates. While we do not attempt to infer the edge strengths of the graph in this chapter, we do describe and briefly explore how one may incorporate *a priori* geological knowledge into the inference procedure by manipulation of the edge strengths. This exploratory problem motivated our deeper study into how one might invert for these edge strengths in a hierarchical Bayesian setting, leading us to the work of Chapter 4.

Chapter 4: Least-Squares Migration with a Hierarchical Bayesian Framework

In Chapter 4, we turn to the problem of seismic imaging (also referred to as migration), where the model is now the seismic image (i.e. the earth's reflectivity model), and the data consist of full seismic waveforms measured by a set of seismic receivers at the surface. While migration is traditionally performed via back-propagation of the recorded seismic wavefield into the model domain [11], recent efforts [19, 36, 49] attempt to give the image as the solution to a least-squares inverse problem. This approach to the seismic imaging problem is referred to as *least-squares migration* (LSM). To remain analogous to LSM, we model the image with a Gaussian prior distribution and give the data as a linear function of the image corrupted by additive Gaussian noise. The dependence of the data on the image is given by the Kirchhoff modeling operator, which can be viewed as a ray-theoretic single-scattering approximation to the integral wave-equation. The Gaussian prior on the image is again described by the parameterized MRF from Chapter 2, but now we wish to infer the edge strengths from the data in the hierarchical Bayesian setting. To do so, we endow the edge strengths with their own prior and obtain the marginal MAP estimate for the edge strengths via the expectation-maximization (E-M) algorithm [17, 46]. We verify our procedure on 2-D synthetic datasets. The images we obtain after inferring the edge strengths exhibit the desired spatially-varying smoothness properties: the images are generally smooth along the reflectors but are allowed to vary sharply at pixels adjacent to a reflector, so as not to smooth out the discontinuity. In contrast, the images obtained when the edge strengths are fixed are either too smooth or overly noisy.

Chapter 5: Interpretation and Estimation of Regularization Parameters

While we inferred the edge strengths in the previous chapter, the remaining parameters defining the prior were decided in a somewhat *ad hoc* fashion. In Chapter 5, we present two approaches to picking these remaining parameters more rigorously. In the first part of this chapter, we elucidate further on the connections between these parameters and the prior model covariance that results from a particular choice. To do this analytically, we follow the methodology of Rodi and Myers [59] and Simpson et al. [68], passing from a random vector representation of the model defined on a discrete spatial grid to the limit of a continuous random field representation. In the limiting case, we can obtain the covariance of the model as the Green's function of a differential operator, where the properties of this covariance function depend on the parameters defining the prior distribution.

In the second part of this chapter, we seek to extend the inference framework of the previous chapter to include these parameters. Here we perform the inference via the variational Bayesian method [3], which can be viewed as a generalization of the E-M algorithm [3, 4], and validate our methodology on a synthetic dataset.

Chapter 6: Hierarchical Bayesian Time-Lapse Seismic Processing

In Chapter 6, we apply the hierarchical Bayesian framework and algorithmic machinery explored in the previous chapters to the problem of time-lapse seismic inversion. Here the goal is to detect changes in the subsurface over a period of time by taking repeated seismic surveys (a first "baseline" survey and a second "monitor" survey). Conventional methods for time-lapse inversion typically involve subtracting repeated datasets and inverting the differenced data to obtain the time-lapse change [8, 30, 81]; these methods, however, require identical acquisition geometries between subsequent seismic surveys, which is often difficult to achieve. In the realistic case of differing acquisition geometries, one common approach is to invert the datasets separately and estimate the time-lapse change model as the difference in the inversion models. However, this method often performs poorly due to differences in illumination and observational noise between the datasets. To correctly treat the case of differing acquisition geometries between the baseline and monitor surveys, we cast the problem in the hierarchical Bayesian setting and seek the marginal MAP estimate for the time-lapse change in the model parameters. We again solve the Bayesian inference problem using the E-M algorithm, which, in this case, iterates between performing subsequent updates to the background model and the time-lapse change. We verify the inference results on a time-lapse seismic imaging example involving synthetic datasets with different acquisition geometries.

Chapter 7: Conclusions

In Chapter 7, we summarize the major contributions of our work, give concluding remarks, and suggest some avenues for future research.

Chapter 2

A Brief Primer on Inverse Problems, Bayesian Inference, and Graphical Models

2.1 An Introduction to Inverse Problems

In an inverse problem, the objective is to infer from a set of observed data $\mathbf{d}$ some unknown attributes of interest which we refer to as the model parameters $\mathbf{m}$. For the inverse problems of interest in this thesis, $\mathbf{m} \in \mathbb{R}^N$ is an N-dimensional model vector of geophysical properties (such as acoustic reflectivity) defined over a 2-D (or potentially 3-D) spatial grid of the subsurface, and $\mathbf{d} \in \mathbb{R}^K$ is a K-dimensional data vector consisting of either a set of seismic traces or features extracted from the seismic dataset (such as P-wave reflection amplitudes, for example).

Typically, a forward modeling operator F relates the model parameters to the observed data, i.e.

$$\mathbf{d} = F(\mathbf{m}) + \mathbf{n}, \tag{2.1}$$

where $\mathbf{n}$ is a noise term introduced to capture measurement and modeling errors. Solving the inverse problem then involves finding the model parameters that minimize some data misfit function $\Omega(\mathbf{m}, \mathbf{d})$, usually taken to be some norm of the *residual*

$\mathbf{r} = \mathbf{d} - A\mathbf{m}$. For example, one might search for the model that gives the best fit to the data in the least-squares sense, wherein the data misfit function is given by the squared ℓ^2-norm of the residual

$$\Omega(\mathbf{m}, \mathbf{d}) = \frac{1}{2}\|\mathbf{d} - F(\mathbf{m})\|_2^2. \tag{2.2}$$

The model $\mathbf{m}_{\text{LS}}$ that minimizes (2.2) is consequently known as the *least-squares solution*

$$\mathbf{m}_{\text{LS}} = \arg\min_{\mathbf{m}} \frac{1}{2}\|\mathbf{d} - F(\mathbf{m})\|_2^2. \tag{2.3}$$

Assuming F is continuously differentiable, we can derive the first-order necessary condition on $\mathbf{m}_{\text{LS}}$ by setting the gradient (with respect to $\mathbf{m}$) of $\Omega(\mathbf{m}, \mathbf{d})$ in (2.2) to 0:

$$\nabla_{\mathbf{m}}\Omega(\mathbf{m}, \mathbf{d}) = -A(\mathbf{m})^T (\mathbf{d} - F(\mathbf{m})) = 0, \tag{2.4}$$

where $A(\mathbf{m})$ is the Jacobian of F. When $F(\mathbf{m})$ is a linear function of the model (so that $F(\mathbf{m}) = A\mathbf{m}$ and the Jacobian A does not vary with $\mathbf{m}$), then (2.4) implies that $\mathbf{m}_{\text{LS}}$ must be a solution to the linear system

$$A^T A\mathbf{m}_{\text{LS}} = A^T\mathbf{d}. \tag{2.5}$$

If $K \geq N$ and the columns of A are linearly independent (i.e. A is full column rank), then $A^T A$ is positive definite and (2.5) has a unique solution:

$$\mathbf{m}_{\text{LS}} = (A^T A)^{-1} A^T\mathbf{d}. \tag{2.6}$$

Unfortunately, even when the solution is unique, many inverse problems exhibit ill-posedness, where even a small amount of noise in the data can lead to large errors in the estimated model parameters [77]. In order to solve ill-posed problems and prevent overfitting of noisy data, additional information about the model parameters can be introduced in a process known as regularization [50, 77]. This is typically achieved by adding to the data misfit function a regularization function $\Phi(\mathbf{m})$ that depends solely on the model parameters $\mathbf{m}$ and penalizes models that are inconsistent this information. The regularized solution $\mathbf{m}_{\text{reg}}$ then minimizes the combined penalty terms:

$$\mathbf{m}_{\text{reg}} = \arg\min_{\mathbf{m}} \Phi(\mathbf{m}) + \Omega(\mathbf{m}, \mathbf{d}). \tag{2.7}$$

For example, if the model is known to be spatially smooth *a priori*, an appropriate regularization function may penalize local differences in the model parameters such as

$$\begin{aligned}\Phi(\mathbf{m}) &= \frac{1}{2}\lambda \sum_{(i,j)\in\mathcal{E}} \beta_{ij}(m_i - m_j)^2 & (2.8)\\ &= \frac{1}{2}\lambda \mathbf{m}^T D(\boldsymbol{\beta})\mathbf{m}, & (2.9)\end{aligned}$$

where the differencing weights $\beta_{ij} \in [0,1]$ capture the degree of smoothness we expect between m_i and m_j, $\mathcal{E}$ is the set of pairs of indices that correspond to spatial neighbors, $D(\boldsymbol{\beta})$ is a differencing matrix capturing this operation defined by the vector $\boldsymbol{\beta} = \{\beta_{ij} : (i,j) \in \mathcal{E}\}$, and $\lambda > 0$ is a trade-off parameter that assigns the maximum weight given to penalizing these differences. Alternatively, if the model is believed to be near some reference model $\mathbf{m}_0$, the regularization function might quantify the distance between the model parameters and $\mathbf{m}_0$, e.g.

$$\Phi(\mathbf{m}) = \frac{1}{2}\lambda\|\mathbf{m} - \mathbf{m}_0\|_2^2. \tag{2.10}$$

Both (2.9) and (2.10) are examples of *Tikhonov regularization* [70, 77], where the regularization function is taken to be a quadratic positive semi-definite function of the model. The Tikhonov regularization function can be expressed in a general form as

$$\Phi(\mathbf{m}) = \frac{1}{2}(\mathbf{m} - \mathbf{m}_0)^T Q(\mathbf{m} - \mathbf{m}_0), \tag{2.11}$$

where the regularization matrix Q is symmetric and positive semi-definite.

When the data misfit function $\Omega(\mathbf{m}, \mathbf{d})$ is taken to be the squared ℓ^2-norm of the residual (as in (2.2)), the model $\mathbf{m}_{\text{RLS}}$ that solves the Tikhonov regularized minimization problem of (2.7) is referred to as the *regularized least-squares* solution

$$\mathbf{m}_{\text{RLS}} = \arg\min_{\mathbf{m}} \frac{1}{2}(\mathbf{m}-\mathbf{m}_0)^T Q(\mathbf{m}-\mathbf{m}_0) + \frac{1}{2}\|\mathbf{d} - F(\mathbf{m})\|_2^2. \tag{2.12}$$

(2.12) can be made even more general by instead considering a matrix-weighted quadratic function of the residual $\mathbf{r}^T P\mathbf{r}$ (for some symmetric positive definite matrix P) for the data misfit term, so that solving (2.7) will yield the *regularized weighted least-squares* model $\mathbf{m}_{\text{RWLS}}$:

$$\mathbf{m}_{\text{RWLS}} = \arg\min_{\mathbf{m}} \frac{1}{2}(\mathbf{m}-\mathbf{m}_0)^T Q(\mathbf{m}-\mathbf{m}_0) + \frac{1}{2}\left(\mathbf{d} - F(\mathbf{m})\right)^T P\left(\mathbf{d} - F(\mathbf{m})\right). \tag{2.13}$$

As will be seen shortly, the matrices Q and P (when both are positive definite) can be interpreted as inverse covariance matrices of the model $\mathbf{m}$ (prior to observing the data) and the noise $\mathbf{n}$, respectively. As before, when F is a linear function of the model (described by the matrix A), applying the stationarity condition on $\mathbf{m}_{\text{RWLS}}$ yields an analytical solution to (2.13):

$$\mathbf{m}_{\text{RWLS}} = \left(A^T PA + Q\right)^{-1}\left(A^T P\mathbf{d} + Q\mathbf{m}_0\right), \tag{2.14}$$

where the above requires that $A^T PA + Q$ be positive definite.

2.2 Bayesian Inference

Having described the fundamental components of an inverse problem, we turn to the setting of Bayesian inference which provides a useful and mathematically rigorous framework for inferring the model parameters from the data and handling the relevant uncertainties in both the model and data. In what follows, we give a brief overview of the concepts of Bayesian inference and probabilistic graphical models. A more thorough treatment of these concepts can be found in Gelman et al. [24] and Koller and Friedman [35].

For the sake of clarity, we pause here to make a few comments about the notation and terminology we will use. In what follows and throughout this thesis, we use sans-serif font for random variables (e.g. m_i), **boldface** font for vectors (e.g. $\mathbf{m}$), and combine both fonts for random vectors (e.g. $\boldsymbol{\mathsf{m}}$). When we refer to the probability distribution (or simply distribution) of a random vector $\boldsymbol{\mathsf{m}}$, we are referring to its probability mass function (in the case of a discretely-valued random vector) or its probability density function (in the case of a continuously-valued random vector). We denote the distribution of $\boldsymbol{\mathsf{m}}$ by $p_{\boldsymbol{\mathsf{m}}}(\mathbf{m})$, or, when the meaning is clear from the context, we drop the subscript and simply write $p(\mathbf{m})$. We use analogous notation for conditional distributions: e.g. for the conditional distribution of $\boldsymbol{\mathsf{m}}$ given $\boldsymbol{\mathsf{d}}$, we will write $p_{\boldsymbol{\mathsf{m}}|\boldsymbol{\mathsf{d}}}(\mathbf{m}|\mathbf{d})$ or simply $p(\mathbf{m}|\mathbf{d})$.

In the context of Bayesian inference, both the model and data are treated as random vectors, denoted by $\boldsymbol{\mathsf{m}}$ and $\boldsymbol{\mathsf{d}}$. Our belief about the model parameters $\boldsymbol{\mathsf{m}}$ prior to observing the data is encoded via $p(\mathbf{m})$, the *prior distribution* on $\boldsymbol{\mathsf{m}}$. Similarly, our belief about what the data $\boldsymbol{\mathsf{d}}$ will be, given a particular model $\boldsymbol{\mathsf{m}} = \mathbf{m}$, is captured by the *likelihood model* $p(\mathbf{d}|\mathbf{m})$, which can be thought of as a stochastic forward model for the data. These two probability distributions fully specify the probabilistic model for $\boldsymbol{\mathsf{m}}$ and $\boldsymbol{\mathsf{d}}$ and are used to compute the *posterior distribution* on the model parameters $p(\mathbf{m}|\mathbf{d})$ via Bayes' rule:

$$p(\mathbf{m}|\mathbf{d}) = \frac{p(\mathbf{m})p(\mathbf{d}|\mathbf{m})}{\int_{\mathbb{R}^N} p(\mathbf{m}')p(\mathbf{d}|\mathbf{m}')\,\mathrm{d}\mathbf{m}'}\,. \tag{2.15}$$

If $\boldsymbol{\mathsf{m}}$ is discretely-valued, the integral in the denominator of (2.15) would be replaced by a summation. The posterior distribution is the complete solution to the problem of inferring $\boldsymbol{\mathsf{m}}$ from $\boldsymbol{\mathsf{d}}$, and updates our belief about the model upon having observed the data. When $\boldsymbol{\mathsf{m}}$ is high-dimensional, it may not be possible to tractably explore the entire posterior distribution and often one will instead seek a point estimator of the model $\boldsymbol{\mathsf{m}}$. A Bayesian point estimator $\hat{\mathbf{m}}(\cdot)$ is a function of the data $\boldsymbol{\mathsf{d}}$ that is obtained by minimizing an expected Bayes cost function $B(\hat{\mathbf{m}}(\boldsymbol{\mathsf{d}}), \boldsymbol{\mathsf{m}})$ which encodes the cost of estimating $\boldsymbol{\mathsf{m}}$ as $\hat{\boldsymbol{\mathsf{m}}}(\boldsymbol{\mathsf{d}})$. Hence we have:

$$\hat{\mathbf{m}}(\cdot) = \arg\min_{f(\cdot)} \mathbb{E}_{p(\mathbf{m},\mathbf{d})}\left[B(f(\boldsymbol{\mathsf{d}}), \boldsymbol{\mathsf{m}})\right] \tag{2.16}$$

$$= \arg\min_{f(\cdot)} \mathbb{E}_{p(\mathbf{d})}\left[\mathbb{E}_{p(\mathbf{m}|\mathbf{d})}\left[B(f(\boldsymbol{\mathsf{d}}), \boldsymbol{\mathsf{m}}) \,|\, \boldsymbol{\mathsf{d}}\right]\right], \tag{2.17}$$

where $\mathbb{E}_p$ denotes the expectation operator with respect to distribution p and (2.17) is due to the law of iterated expectations. As seen by (2.17), the total expected Bayes cost of (2.16) is a non-negative combination of the conditional expected Bayes costs given the different realizations of the data. Hence the expected Bayes cost (2.16) can be minimized by designing $\hat{\mathbf{m}}(\cdot)$ to minimize the conditional expected Bayes costs for each realization of the data $\boldsymbol{\mathsf{d}} = \mathbf{d}$, so that

$$\hat{\mathbf{m}}(\mathbf{d}) = \arg\min_{\mathbf{m}'} \mathbb{E}_{p(\mathbf{m}|\mathbf{d})}\left[B(\mathbf{m}', \boldsymbol{\mathsf{m}}) \,|\, \boldsymbol{\mathsf{d}} = \mathbf{d}\right]. \tag{2.18}$$

The particular Bayesian point estimator we obtain from (2.18) is determined by the choice of the Bayes cost function $B(\cdot, \cdot)$. If B quantifies the squared ℓ^2-norm of the estimation error, so that

$$B(\hat{\mathbf{m}}, \boldsymbol{\mathsf{m}}) = \|\hat{\mathbf{m}} - \boldsymbol{\mathsf{m}}\|_2^2, \tag{2.19}$$

solving (2.18) yields the *Bayes least-squares* (BLS) estimator $\mathbf{m}_{\text{BLS}}$, which turns out to be the posterior mean of $\boldsymbol{\mathsf{m}}$:

$$\mathbf{m}_{\text{BLS}}(\mathbf{d}) = \mathbb{E}[\boldsymbol{\mathsf{m}}|\mathbf{d}]. \tag{2.20}$$

The well-known *maximum a posteriori* (MAP) estimator $\mathbf{m}_{\text{MAP}}$, which maximizes the posterior distribution $p(\mathbf{m}|\mathbf{d})$, is obtained by uniformly penalizing all non-zero estimation errors; to be precise, if, for some $\epsilon > 0$, B is taken to be

$$B(\hat{\mathbf{m}}, \boldsymbol{\mathsf{m}}) = \begin{cases} 1 & \|\hat{\mathbf{m}} - \boldsymbol{\mathsf{m}}\| > \epsilon \\ 0 & \|\hat{\mathbf{m}} - \boldsymbol{\mathsf{m}}\| \leq \epsilon \end{cases} \tag{2.21}$$

then taking the limit as $\epsilon \to 0$ in (2.21) yields the MAP estimator as the solution to (2.18), which is given by:

$$\mathbf{m}_{\text{MAP}}(\mathbf{d}) = \arg\max_{\mathbf{m}} p(\mathbf{m}|\mathbf{d}). \tag{2.22}$$

The connection between the Bayesian inference setting and the deterministic inversion framework described in Section 2.1 is perhaps best seen through the MAP estimator. The MAP estimator can be equivalently expressed as the maximizer of the log posterior distribution, since log is a monotonic function; hence we can rewrite (2.22) as

$$\begin{aligned} \mathbf{m}_{\text{MAP}}(\mathbf{d}) &= \arg\max_{\mathbf{m}} \; \log p(\mathbf{m}|\mathbf{d}) && (2.23) \\ &= \arg\max_{\mathbf{m}} \; \log p(\mathbf{m}) + \log p(\mathbf{d}|\mathbf{m}), && (2.24) \end{aligned}$$

where we have employed Bayes' rule (2.15) and dropped the denominator which does not depend on $\mathbf{m}$. Noting that maximizing a function is the same as minimizing the negative of that function, (2.24) becomes

$$\mathbf{m}_{\text{MAP}}(\mathbf{d}) = \arg\min_{\mathbf{m}} \; \{-\log p(\mathbf{m}) - \log p(\mathbf{d}|\mathbf{m})\}. \tag{2.25}$$

Comparing the form of the MAP estimate in (2.25) to that of the regularized solution from the deterministic inversion framework in (2.7), we see that the two are mathematically equivalent. In particular, we can interpret the regularization and data misfit functions of (2.7) as negative log prior and likelihood models, respectively:

$$\Phi(\mathbf{m}) = -\log p(\mathbf{m}) + const., \tag{2.26}$$

$$\Omega(\mathbf{m}, \mathbf{d}) = -\log p(\mathbf{d}|\mathbf{m}) + const. \tag{2.27}$$

The regularized weighted least-squares solution of (2.13) can be interpreted as a Bayesian MAP estimate when the prior model and noise are both Gaussian. In particular, we endow $\mathbf{m}$ with a Gaussian prior distribution, having prior mean $\mathbf{m}_0$ and covariance matrix C (i.e. $\mathbf{m} \sim \mathcal{N}(\mathbf{m}_0, C)$) so that

$$p(\mathbf{m}) = \frac{\exp\left\{-\frac{1}{2}\left(\mathbf{m} - \mathbf{m}_0\right)^T C^{-1} \left(\mathbf{m} - \mathbf{m}_0\right)\right\}}{(2\pi)^{N/2} \left|C\right|^{1/2}}, \tag{2.28}$$

where $|\cdot|$ denotes the determinant, and let the data be described by

$$\mathbf{d} = F(\mathbf{m}) + \mathbf{n}, \tag{2.29}$$

where $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \Sigma)$ is a zero-mean Gaussian noise vector independent from the model $\mathbf{m}$, so that

$$p(\mathbf{d}|\mathbf{m}) = \frac{\exp\left\{-\frac{1}{2}\left(\mathbf{d} - F(\mathbf{m})\right)^T \Sigma^{-1} \left(\mathbf{d} - F(\mathbf{m})\right)\right\}}{(2\pi)^{K/2} \left|\Sigma\right|^{1/2}}. \tag{2.30}$$

The posterior distribution is then given by

$$p(\mathbf{m}|\mathbf{d}) = \frac{\exp\left\{-\frac{1}{2}\left(\mathbf{m} - \mathbf{m}_0\right)^T C^{-1} \left(\mathbf{m} - \mathbf{m}_0\right) - \frac{1}{2}\left(\mathbf{d} - F(\mathbf{m})\right)^T \Sigma^{-1} \left(\mathbf{d} - F(\mathbf{m})\right)\right\}}{(2\pi)^{(N+K)/2} \left|C\right|^{1/2} \left|\Sigma\right|^{1/2} p(\mathbf{d})} \tag{2.31}$$

$$\propto \exp\left\{-\frac{1}{2}\left(\mathbf{m} - \mathbf{m}_0\right)^T C^{-1} \left(\mathbf{m} - \mathbf{m}_0\right) - \frac{1}{2}\left(\mathbf{d} - F(\mathbf{m})\right)^T \Sigma^{-1} \left(\mathbf{d} - F(\mathbf{m})\right)\right\}, \tag{2.32}$$

so that the Bayesian MAP estimate of the model is

$$\mathbf{m}_{\mathrm{MAP}} = \arg\max_{\mathbf{m}} \log p(\mathbf{m}|\mathbf{d}) \tag{2.33}$$

$$= \arg\min_{\mathbf{m}} \frac{1}{2}\left(\mathbf{m} - \mathbf{m}_0\right)^T C^{-1} \left(\mathbf{m} - \mathbf{m}_0\right) + \frac{1}{2}\left(\mathbf{d} - F(\mathbf{m})\right)^T \Sigma^{-1} \left(\mathbf{d} - F(\mathbf{m})\right). \tag{2.34}$$

Comparing (2.34) with the regularized weighted least-squares cost function of (2.13), we see that (for positive definite Q and P in (2.13)) the two are equivalent and $\mathbf{m}_{\mathrm{RWLS}}$ is the Bayesian MAP estimate when the prior model and noise are Gaussian as described above. Furthermore, we can interpret the regularization matrix Q as the inverse prior covariance matrix (also called the prior precision matrix) C^{-1} and the weighting matrix P (from the quadratic data misfit function) as the inverse covariance matrix of the noise [70].

2.3 Probabilistic Graphical Models, Markov Random Fields, and Covariance

We return to the examples of Tikhonov regularization given in (2.9) and (2.10). In particular, we consider the case when the regularization function $\Phi(\mathbf{m})$ both enforces smoothness in the model parameters and penalizes the magnitude of the model:

$$\Phi(\mathbf{m}) = \frac{1}{2}\lambda \left(\sum_{(i,j)\in\mathcal{E}} \beta_{ij}(m_i - m_j)^2 + \epsilon \sum_{i\in\mathcal{V}} m_i^2 \right) \tag{2.35}$$

$$= \frac{1}{2}\mathbf{m}^T \left(\lambda \left(D(\boldsymbol{\beta}) + \epsilon I\right)\right) \mathbf{m}, \tag{2.36}$$

where $\lambda > 0$, $\epsilon > 0$, $\boldsymbol{\beta} = \{\beta_{ij} : (i,j) \in \mathcal{E}\}$ with each $\beta_{ij} \in [0,1]$, the set $\mathcal{V} = \{1, \ldots, N\}$ indexes the components of $\mathbf{m}$, and $\mathcal{E}$ and the differencing matrix $D(\boldsymbol{\beta})$ take the same meaning as in (2.9). In the Bayesian inference context, this is equivalent to modeling $\mathbf{m}$ *a priori* as Gaussian

$$\mathbf{m} \sim \mathcal{N}(\mathbf{0}, C) \tag{2.37}$$

where the prior precision matrix $Q = C^{-1}$ is given by

$$Q = C^{-1} = \lambda \left(D(\boldsymbol{\beta}) + \epsilon I\right). \tag{2.38}$$

It is clear from (2.35) that the differencing coefficients β_{ij} determine how strongly to penalize differences between m_i and m_j and hence the resulting smoothness of the regularized solution. In the Bayesian context, $\boldsymbol{\beta}$ similarly affects the smoothness of $\mathbf{m}_{\text{MAP}}$ through the prior on $\mathbf{m}$. Indeed the prior distribution plays a key role in determining the smoothness properties of the model parameters.

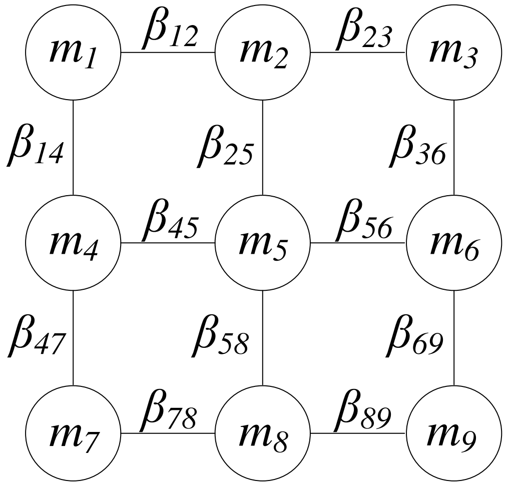


Figure 2-1: The Markov random field imposed on $\mathbf{m}$ by fixing $\boldsymbol{\beta}$ prior to observing the data $\mathbf{d}$, for a simple nine pixel image.

2.3.1 Probabilistic Graphical Models

The expressive formalism of probabilistic graphical models provides a useful analytical framework for both understanding the interdependencies imposed by a particular distribution as well as implementing efficient algorithms to solve the underlying inference problem (some of which are explored in Chapter 3). We define an undirected graphical model $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, with a set of vertices (or nodes) $\mathcal{V}$, which index the random variables m_i comprising the random vector $\mathbf{m}$, and a set of edges $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$, represented as pairs of vertices in $\mathcal{V}$ that encode dependencies between the random variables. We say that the random vector $\mathbf{m}$ forms a Markov random field (MRF) over $\mathcal{G}$ if it is *Markov* on $\mathcal{G}$, meaning that, for any disjoint subsets of nodes $S, T, U \subset \mathcal{V}$ such that S separates T from U on $\mathcal{G}$ (i.e. if we remove the nodes in S from $\mathcal{G}$ then there is no remaining path from any node in T to any node in U), then conditioned on $\mathbf{m}_S$, $\mathbf{m}_T$ is conditionally independent of $\mathbf{m}_U$ (which we write as $\mathbf{m}_T \perp\!\!\!\perp \mathbf{m}_U \,|\, \mathbf{m}_S$). Here we have defined $\mathbf{m}_S \triangleq \{\mathsf{m}_i : i \in S\}$ as the set of random variables corresponding to the nodes in S (similarly for $\mathbf{m}_T$ and $\mathbf{m}_U$). An example of an undirected graphical model is depicted in Figure 2-1.

To make the connection between a particular distribution and a graphical model, we turn to the Hammersley-Clifford theorem [13, 28, 35]. The Hammersley-Clifford theorem states that a distribution is Markov on an undirected graphical model $\mathcal{G}$ if it factorizes over the maximal cliques of that graph, and, for distributions that are strictly positive, the converse statement also holds. Here a *clique* $\mathcal{C}$ of a graph $\mathcal{G}$ is any subset of its vertices ($\mathcal{C} \subset \mathcal{V}$) which are fully-connected, meaning every vertex in $\mathcal{C}$ shares an edge from $\mathcal{E}$ with every other vertex in $\mathcal{C}$. The *maximal cliques* of a graph $\mathcal{G}$ are its largest possible cliques: $\mathcal{C}$ is a maximal clique of $\mathcal{G}$ if it fails to remain a clique when even one more vertex from $\mathcal{V} \setminus \mathcal{C}$ is added to $\mathcal{C}$. We say that a distribution $p_{\mathsf{m}}(\mathbf{m})$ *factorizes* over the maximal cliques of $\mathcal{G}$ if it can be written as a product of functions of the random variables in each maximal clique $\mathcal{C}$ of $\mathcal{G}$: $p_{\mathsf{m}}(\mathbf{m}) = \prod_{\mathcal{C}} f_{\mathcal{C}}(\mathbf{m}_{\mathcal{C}})$.

For a Gaussian random vector m with precision matrix (or inverse covariance matrix) Q, the Hammersley-Clifford theorem implies that m is Markov on the graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ if and only if $Q_{ij} = 0$ whenever there is no edge between nodes i and j (i.e. whenever Q is at least as sparse as the edge set $\mathcal{E}$) [35]. Hence, a Gaussian random vector with a given precision matrix Q induces a natural graph based on the sparsity pattern of Q.

The graph shown in Figure 2-1 is the natural graph induced by the Gaussian prior described in Equations (2.37)-(2.38), and the sets $\mathcal{V}$ and $\mathcal{E}$ used in (2.35) are precisely the vertex and edge sets of this graph. The edges of the graph in Figure 2-1 are labeled by the differencing weights β_{ij} because, in a sense, the β_{ij} determine the "strength" of each edge. To be precise, $\boldsymbol{\beta}$ captures the conditional dependence structure of m, such that $\beta_{ij} = 0$ implies that there is no edge between m_i and m_j and hence $\mathsf{m}_i \perp\!\!\!\perp \mathsf{m}_j \,|\, \{\mathsf{m}_k : k \neq i, j\}$, and a larger value for β_{ij} induces a stronger conditional correlation between m_i and m_j. For this reason, we sometimes refer to the elements of $\boldsymbol{\beta}$ as the *edge strengths* of $\mathcal{G}$ and to $D(\boldsymbol{\beta})$ as the weighted graph Laplacian of $\mathcal{G}$ (weighted by $\boldsymbol{\beta}$). (We note to the reader that, although we have defined the notion of *edge strengths* in this chapter, we will re-introduce and review this concept in Chapter 4.)

2.3.2 Edge Strengths and Covariance

This graph theoretic approach to defining a distribution via a Markov random field contrasts with the more typical method for defining priors in geophysical inverse problems using stationary covariance functions (typically over an assumed Gaussian random field) [70]. A covariance function is often characterized by its variance and correlation length, the latter being a characteristic length defining the rate of decay of the covariance function. While defining a covariance function explicitly encodes the covariances between any two points in space, an MRF instead encodes the conditional independencies and (when parameterized with edge strengths) the local smoothness structure of the model, thereby implicitly defining a covariance function.

To show the effect of $\boldsymbol{\beta}$ on the covariance of m in the Gaussian example described by (2.37)-(2.38), we compute one row of the covariance matrix $C(\boldsymbol{\beta})$ and draw samples of m for different choices of $\boldsymbol{\beta}$. We note that since $C_{ij} = \mathrm{cov}(\mathsf{m}_i, \mathsf{m}_j)$, the ith row of the covariance matrix gives the covariance of the entire model with m_i. The covariances and samples are computed with m defined over a 101-by-101 node grid (with grid spacing set to 1 m) and with setting $\lambda = 1$ and $\epsilon = 10^{-3}$ and where the covariances with the central point in the grid are computed. We consider two cases: the first where each β_{ij} is set to a common value β, to illustrate how varying the β_{ij} uniformly affects the model covariance, and the second where some of the β_{ij} are set to 0 and the remainder are set to 1, to illustrate how the β_{ij} can capture spatially-varying smoothness in the model.

Figure 2-2 displays the first case where all the β_{ij} are set to a common value, where this value is varied from 1 to 0. As shown in the covariance plots, as β decreases, the resulting covariance function becomes taller and skinnier: decreasing β results in an increased model variance and decreased correlation length. $\beta = 0$ corresponds to the case of a completely disconnected graph, so that all the m_i are independent, and hence the covariance function is 0 everywhere except at the central point, which corresponds to the variance of m_i. The draws from the Gaussian distributions similarly show this. When $\beta = 1$, the sample is highly correlated in space and has a relatively smaller magnitude (due to the lower variance), however as β is made smaller, we observe that the sample becomes spatially less correlated and the magnitude of its range increases. At $\beta = 0$, the sample is of white noise and hence completely uncorrelated in space.

The second case, where we set some of the β_{ij} to 0 and the rest to 1 is shown in Figure 2-3. In particular we consider the choice of $\boldsymbol{\beta}$ we may want to use if we suspect a horizontal discontinuity in the model at $z = 40.5$ m and centered at $x = 51$ m; in this case, we would set $\beta_{ij} = 0$ for the edge strengths corresponding to vertical edges connecting nodes at $z = 40$ m and $z = 41$ m along the length of the discontinuity. We show the covariances and sample draws for discontinuities of different lengths: 21 m, 51 m, and an infinite length discontinuity. As evidenced by the plots, the covariance is significantly reduced across the discontinuity where the β_{ij} are set to 0; however, for the finite-length discontinuities, the covariance remains non-zero across the discontinuity, as there is still a path on the graph $\mathcal{G}$ connecting the nodes to the central node, through which they remain correlated to the central node. For the infinite-length discontinuity, no such path exists as the nodes above the discontinuity are completely disconnected from those below the discontinuity; hence the two sets of nodes are independent, and the covariances of the nodes above the discontinuity with the central node are identically 0. This is similarly observed in the sample draws: the draw corresponding to the infinite-length discontinuity shows that the nodes above and below the discontinuity are uncorrelated. For the draws corresponding to the finite-length discontinuities, one observes that the sample values are allowed to contrast significantly across the discontinuity, however a reduced correlation still exists through paths of nodes that avoid the discontinuity.

The preceding examples have served to illustrate the significance of the edge strengths $\boldsymbol{\beta}$ in defining the prior distribution for $\mathbf{m}$ and to point out the flexibility of this construction for treating inverse problems with models that have spatially-varying smoothness properties. In Chapter 5, we undertake a more rigorous investigation of the parameters $\boldsymbol{\beta}$, λ, and ϵ and their relationship with the model covariance. We note here that the graphical model depicted in Figure 2-1 with edge set $\mathcal{E}$ connecting only a node's four nearest neighbors is one of many possible graphs that can be used in our

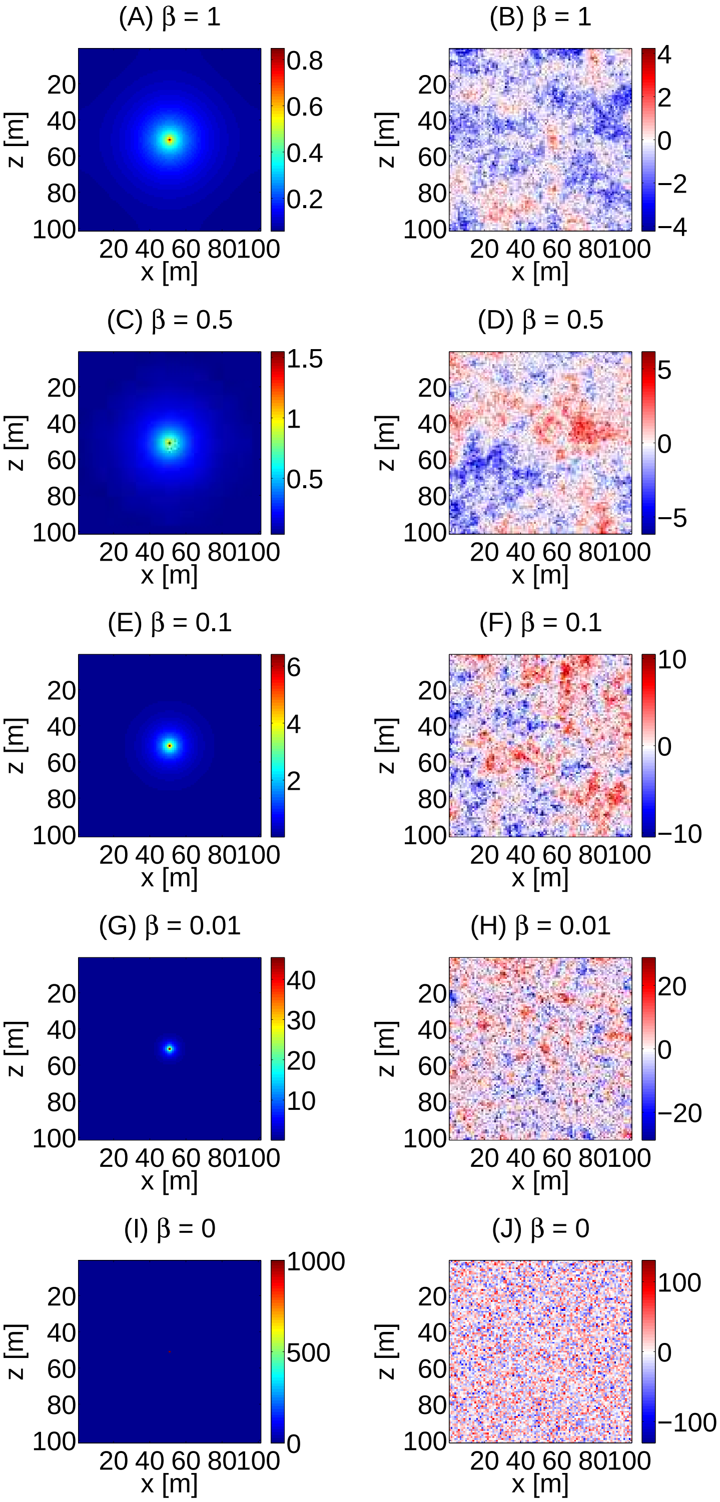


Figure 2-2: Computed covariance functions (first column) and sample draws (second column) from $\mathcal{N}(\mathbf{0}, (\lambda(D(\boldsymbol{\beta}) + \epsilon I))^{-1})$ with each $\beta_{ij} = \beta$ for (A-B) $\beta = 1$, (C-D) $\beta = 0.5$, (E-F) $\beta = 0.1$, (G-H) $\beta = 0.01$, (I-J) $\beta = 0$.

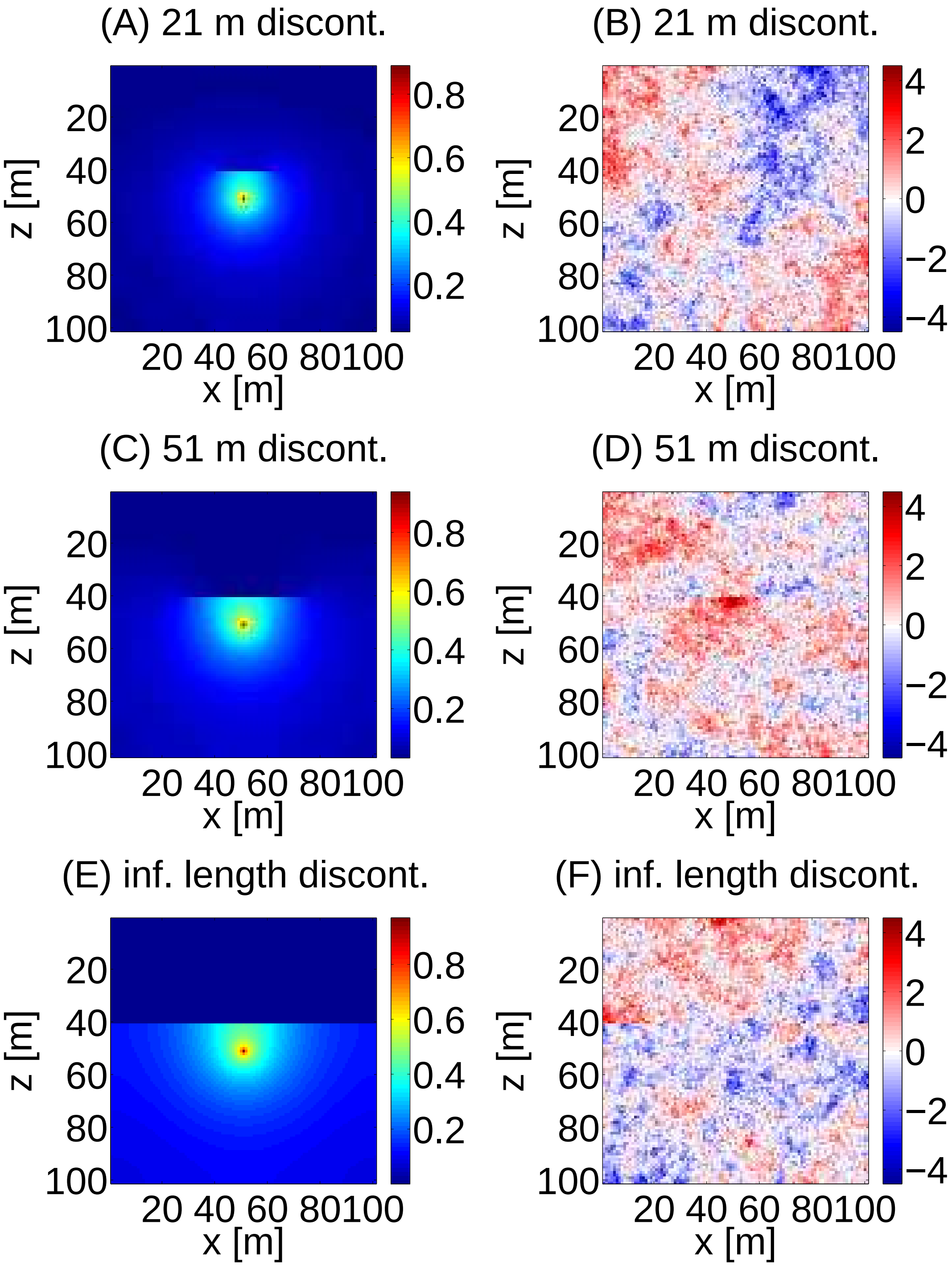


Figure 2-3: Computed covariance functions (first column) and sample draws (second column) from $\mathcal{N}(\mathbf{0}, (\lambda(D(\boldsymbol{\beta}) + \epsilon I))^{-1})$ with $\beta_{ij} = 0$ along a horizontal discontinuity at $z = 40.5$ m (centered at $x = 51$ m) of length (A-B) 21 m, (C-D) 51 m, (E-F) infinite length. $\beta_{ij} = 1$ elsewhere.

construction. In particular, we can generalize the graph in Figure 2-1 by appending $\mathcal{E}$ so that a node is connected to all other nodes within some specified radius. Picking a radius of 1 grid node will give the graph of Figure 2-1; a radius greater than $\sqrt{2}$ but less than 2 will additionally induce diagonal connections, and so on.

In the remainder of this thesis, we explore different inference problems that use this construction with $\boldsymbol{\beta}$. Chapter 3 is an exploratory study in Bayesian inference on the geophysical inverse problem of fracture characterization; there the model parameters are treated as discrete random variables, and the prior distribution is defined using the same graphical model as in Figure 2-1, with $\boldsymbol{\beta}$ fixed. Our investigation of this inference problem made apparent the significance of the choice of the edge strengths $\boldsymbol{\beta}$ to the solution of the problem. Hence, in Chapter 4, we turn to the problem of *estimating* the edge strengths $\boldsymbol{\beta}$ from the data in the context of the seismic imaging problem known as least-squares migration. Therein the model is treated as a Gaussian random vector (following the description given in Equations (2.37)-(2.38)). In Chapter 5, we extend the work of Chapter 4 by investigating the connections between the parameters $\boldsymbol{\beta}$, λ, and ϵ and the model covariance and further generalize the methodology developed in Chapter 4 to additionally estimate λ and the noise variance. Chapter 6 focuses on the problem of time-lapse seismic processing and utilizes some of the machinery discussed in Chapter 4 to correctly frame and solve the inference problem; however, in that chapter, while the construction with $\boldsymbol{\beta}$ is utilized, the focus is on correctly dealing with uncertainty about how the model evolves with time rather than on the edge strengths $\boldsymbol{\beta}$.

Chapter 3

Bayesian Fracture Characterization

3.1 Summary

In this chapter, we describe a methodology for quantitatively characterizing the fractured nature of a hydrocarbon or geothermal reservoir from surface seismic data under a Bayesian inference framework. The method combines different kinds of measurements of fracture properties to find a best fit model while providing estimates of the uncertainty of model parameters. Fractures provide pathways for fluid flow in a reservoir, and hence, knowledge about a reservoir's fractured nature can be used to enhance production from the reservoir. The fracture properties of interest in this study (to be inferred) are fracture orientation and excess compliance, where each of these properties are assumed to vary spatially over a 2-D horizontal grid which is assumed to represent the top of a reservoir. The Bayesian framework in which the inference problem is cast has the key benefits of (1) utilization of a prior model that allows geological information to be incorporated, (2) providing a straightforward means of incorporating all measurements (across the 2-D spatial grid) into the estimates at each grid point, (3) allowing different types of measurements to be combined under a single inference procedure, and (4) providing a measure of uncertainty in the estimates. The observed data are taken from a 2-D array of surface seismic receivers responding to an array of surface sources. Well understood features from the seismic traces are extracted and treated as the observed data, namely the P-wave

reflection amplitude variation with acquisition azimuth and offset (amplitude versus azimuth data) and fracture transfer function data. Amplitude versus azimuth data are known to be more sensitive to fracture properties when the fracture spacing is significantly smaller than the seismic wavelength, whereas fracture transfer function data are more sensitive to fracture properties when the fracture spacing is on the order of the seismic wavelength. Combining these two measurements has the benefit of allowing inferences to be made about fracture properties over a larger range of fracture spacing than otherwise attainable. Geophysical forward models for the measurements are used to arrive at likelihood models for the data, and the prior distribution for the fracture variables is obtained by defining a Markov random field over the horizontal 2-D grid on which we wish to obtain fracture properties. The fracture variables are then inferred by application of loopy belief propagation to yield approximations for the posterior marginal distributions of the fracture properties, as well as the *maximum a posteriori* and Bayes least squares (posterior mean) estimates of these properties. Verification of the inference procedure is performed using a synthetic dataset, where the estimates are shown to be at or near ground truth for both fracture orientation (at the full range of fracture spacings) and fracture excess compliance (at small fracture spacings).

3.2 Introduction

Fractures are cracks in the earth's crust through which fluid, such as oil, natural gas, or brine, can flow. Knowledge about the presence and properties of fractures in a reservoir can be extremely valuable, as such information can be used to determine pathways for fluid flow and to optimize production from the reservoir [1, 62]. Since the presence of fractures in an elastic medium can alter the compliance of the medium and fractures often have a preferred alignment relative to *in situ* stress, fractures can cause the medium to exhibit anisotropy [66]. This anisotropy has been exploited to give different techniques for determining fracture properties from seismic data, such as reflection amplitude versus offset and azimuth analysis [42, 60, 62, 63] and shear wave birefringence [23]. These methods, however, are only valid when the fracture spacing is small in comparison to the seismic wavelength, so that seismic waves average over the fractures [22, 78]. The equivalent anisotropic medium assumption breaks down when the spacing between the fractures increases to being on the order of the seismic wavelength. For the case of larger fracture spacings, Willis et al. [78] proposed a technique, referred to as the scattering index method, to estimate the azimuthal orientation (or strike) of a fracture system based on the scattered seismic energy. Fang et al. [22] described a series of modifications to this technique to give a more robust methodology for determining fracture orientation, which is referred to as the fracture transfer function (FTF) method.

In the way of statistical inference methods applied to geophysical problems, Eidsvik et al. [20] gave a Bayesian framework for determining rock facies and saturating fluid by integrating a forward rock physics model with spatial statistics of rock properties. Specifically in the area of fracture characterization, Ali and Jakobsen [1] used a Bayesian inference framework to infer fracture orientation and density from seismic velocity and attenuation anisotropy data. Sil and Srinivasan [67] applied a similar Bayesian inference methodology to determine fracture strikes from seismic and well data. All of the aforementioned statistical studies solved the inference problem via Markov chain Monte Carlo (MCMC), a sampling technique which stochastically searches the model space. Furthermore, the data models used in all of these studies follow from the assumption that a medium with closely spaced fractures is an equivalent anisotropic medium.

The aim of our study is to utilize a Bayesian framework to combine different data which, on their own, are informative about certain regimes of fracture spacing, to be able to estimate different fracture properties (particularly, excess fracture compliance and fracture orientation) at a wider range of fracture spacings than otherwise attainable if one data type is used. The Bayesian framework in which the problem is cast furthermore makes it straightforward to encode prior knowledge about either geological features of the reservoir, such as the existence of a discontinuity arising from a geological fault, or known information about the fracture properties or their spatial correlation, where we assume that the fracture properties can vary spatially over a 2-D horizontal grid.

We first describe the physical parameters that we use to characterize a fracture system in a reservoir whose properties vary with space. We then present our approach for casting the fracture characterization problem within a Bayesian framework in which we assume that fracture properties within the reservoir are spatially correlated. We capture this spatial correlation by defining a Markov random field (MRF) over the grid of fracture properties, which we use to arrive at the prior distribution for the fracture properties. We then introduce the types of seismic data that we assume are available to characterize fractures, particularly amplitude versus azimuth [60, 63] and fracture transfer function [22], and describe the methods by which these data are used to place constraints on the properties. The inversion for model parameters is accomplished via loopy belief propagation (LBP) [48, 55, 56], a numerically scalable approximate inference procedure that yields both the posterior marginal distributions of the properties at each point in space in addition to the Bayes least squares (BLS) and *maximum a posteriori* (MAP) estimates of the fracture properties. Finally, we demonstrate the applicability of the method for inferring fracture properties using synthetic seismic data.

3.3 Description of the Problem

Consider a set of seismic measurements taken from a 2-D array of surface receivers over a layered medium responding to a set of surface seismic sources. We are interested in inferring from the seismic data whether or not fractures are present in a particular layer of the medium (e.g. the reservoir) and, if so, the properties of the fractures. A simple example of this setup, where the medium consists of flat homogeneous layers, is displayed in Figure 3-1. In particular, we would like to infer fracture orientation $\boldsymbol{\varphi} = [\varphi_{ij}]$ and the (base 10) log excess fracture compliance $\mathbf{z} = [z_{ij}]$ spatially over a 2-D m-by-n horizontal grid $\mathbf{L} = \{1, \ldots, m\} \times \{1, \ldots, n\}$ (where i and j index the axes of the grid $\mathbf{L}$). Each grid point corresponds to a square of area ℓ^2, so that the entire

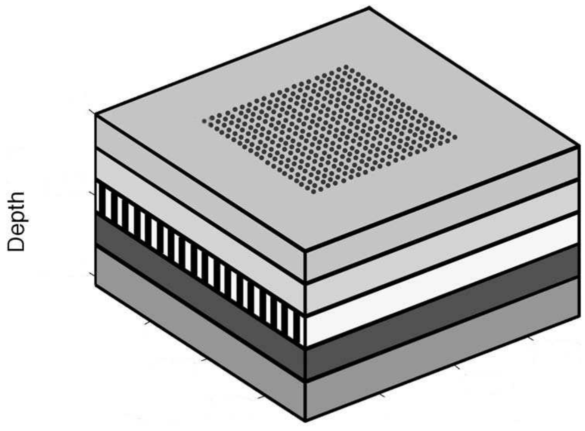


Figure 3-1: A simple model of the problem setting. The formation consists of five flat homogeneous layers with fractures that may be present in the third layer and measurements obtained from the 2-D array of surface seismic receivers. Figure modified from Willis et al. [78].

grid corresponds to a region of area $mn\ell^2$. Excess fracture compliance is defined as the overall additional medium compliance (having units of Pa^{-1}) due to the presence of fractures and is the ratio of the compliance of the individual fractures (in m/Pa) to the fracture spacing (in m) [18, 66]. We make the simplifying assumptions (1) that the fractures are vertical, so that the fracture orientation φ_{ij} is simply the azimuth (or strike) of the fractures (with respect to North), and (2) that the normal and tangential excess compliances are equal, which may represent gas-filled fractures [64], hence we need only infer a single log excess compliance value z_{ij} for each grid node. It is possible that the ratio of normal to tangential fracture compliances may deviate from unity due to mineralization [64]; while this ratio has been measured using shear wave splitting data [74], we are not able to uniquely resolve this ratio from AvAz data. If this ratio is known to differ from unity for a particular fracture system, we may proceed with our analysis inferring for only the normal compliance and setting the tangential compliance according to this ratio. An excess compliance value of 0 at a particular grid point is taken to mean there are no fractures at that grid point (rendering the value for azimuth arbitrary and meaningless). In order to compare zero and non-zero compliance on a logarithmic scale, we treat an excess compliance of zero as 10^{-13} Pa^{-1}, which is geophysically reasonable as this is an insignificantly small value for excess compliance and results in a negligible effect on seismic wave propagation. We assume that the dataset is rich enough so that for each grid point in $\mathbf{L}$, there are corresponding source-receiver pairs that sample the point at multiple offsets and acquisition azimuths. We further assume that the background velocity structure of the medium is well understood.

In order to relate the fracture properties $\mathbf{m} = (\mathbf{z}, \boldsymbol{\varphi}) = [\mathbf{m}_{ij}]$ to the seismic trace dataset, it is necessary to model seismic data as a function of the fracture properties. Unfortunately, modeling the entire seismic trace dataset requires a full elastic 3-D forward simulation of the seismic wavefield, and the computational cost associated with the repeated simulations required to invert for the fracture properties is prohibitively high, so we instead resort to simulating well understood features of the seismic trace dataset and treat these features as our observed data $\mathbf{d}$. In particular, we choose to model P-wave reflection amplitude as a function of acquisition azimuth (at a fixed angle of incidence), also known as amplitude versus azimuth (AvAz) data [60, 63], and fracture transfer function (FTF) data, as defined by Fang et al. [22]. We refer to these observed data with variables $\mathbf{d}^{\mathrm{AvAz}}$ and $\mathbf{d}^{\mathrm{FTF}}$, respectively, and let $\mathbf{d} = (\mathbf{d}^{\mathrm{AvAz}}, \mathbf{d}^{\mathrm{FTF}})$. Detailed descriptions of the data and their forward models are detailed in Section 3.4.2. Both $\mathbf{d}^{\mathrm{AvAz}} = \left[\mathbf{d}_{ij}^{\mathrm{AvAz}}\right]$ and $\mathbf{d}^{\mathrm{FTF}} = \left[\mathbf{d}_{ij}^{\mathrm{FTF}}\right]$ are defined over the grid $\mathbf{L}$, in a manner such that to each grid node of fracture properties $\mathbf{m}_{ij}$ there is an associated data vector $\mathbf{d}_{ij}$.

3.4 Bayesian Inference Framework

In order to arrive at an estimate of the fracture properties from the seismic data, we employ a Bayesian inference framework. As mentioned earlier, the Bayesian framework is chosen as it allows us to naturally encode prior information about the fracture properties (and their spatial variation), combine different types of data, and quantify the uncertainty associated with the inferred quantities. The fracture properties and seismic data are treated as random variables, and a stochastic model is used to give the joint distribution of the fracture properties and seismic data $(\mathbf{m}, \mathbf{d}) = ((\mathbf{z}, \boldsymbol{\varphi}), \mathbf{d})$. In particular, we model the fracture properties as discrete random variables where the domain for each of the variables is given by: $10^{\mathbf{z}_{ij}} \in \mathcal{Z} = \{10^{-9.0}, 10^{-9.1}, \ldots, 10^{-12.0}, 10^{-13}\}$ (in units of Pa^{-1}) and $\varphi_{ij} \in \mathcal{F} = \{0°, 20°, \ldots, 160°\}$, $\forall (i,j) \in \mathbf{L}$. The use of discrete random variables makes the inference problem amenable to the general framework of message-passing inference algorithms described in Section 3.4.3, where the intervals of discretization were picked based on the level of resolution we might reasonably expect to achieve using seismic measurements. The posterior distribution of the fracture properties given the data $p(\mathbf{m}|\mathbf{d})$ is given by Bayes' rule:

$$\begin{aligned} p(\mathbf{m}|\mathbf{d}) &= \frac{p(\mathbf{m})p(\mathbf{d}|\mathbf{m})}{\sum_{\mathbf{m}'} p(\mathbf{m}')p(\mathbf{d}|\mathbf{m}')} \\ &\propto p(\mathbf{m})p(\mathbf{d}|\mathbf{m}) \end{aligned} \tag{3.1}$$

where $p(\mathbf{m})$ and $p(\mathbf{d}|\mathbf{m})$ are the prior distribution of the fracture properties and the distribution of the seismic data given the fracture properties, respectively.

While the posterior distribution of Equation 3.1 is the complete solution to the Bayesian inference problem, exploring this distribution can be intractable due to the high-dimensionality of the fracture properties $\mathbf{m}$. To glean meaningful inferences from the posterior distribution, one may either choose to obtain point estimators of $\mathbf{m}$ from the posterior or to obtain marginal posterior distributions over some tractably explorable subsets of the random variables in $\mathbf{m}$. While point estimators are useful when a single answer to the inference problem is desired, they do not capture the associated estimation uncertainties (which are described by the marginal posterior distributions). Among the most common point estimators are the MAP estimate $\hat{\mathbf{m}}_{\text{MAP}}$ and the BLS (or posterior mean) estimate $\hat{\mathbf{m}}_{\text{BLS}}$. The MAP estimate of the fracture properties minimizes the probability of estimation error and is the overall configuration of the fracture properties that maximizes the posterior distribution, that is

$$\hat{\mathbf{m}}_{\text{MAP}} = \arg\max_{\mathbf{m}} p(\mathbf{m}|\mathbf{d}) \tag{3.2}$$

In contrast, the BLS estimate of the fracture properties minimizes the expected value of the squared estimation error and is given by the expected value of the fracture properties given the data. The posterior marginal distribution for the fracture properties at a particular node m_{ij} is given by summation of the posterior distribution over all other variables $\mathbf{m}_{-ij} \triangleq \{\mathbf{m}_{kl} : (k,l) \in \mathbf{L} \setminus \{(i,j)\}\}$. So, for example, the posterior marginal for the log excess compliance $p(z_{ij}|\mathbf{d})$ is given by

$$p(z_{ij}|\mathbf{d}) = \sum_{\varphi_{ij}} \sum_{\mathbf{m}_{-ij}} p(z_{ij}, \varphi_{ij}, \mathbf{m}_{-ij}|\mathbf{d}). \tag{3.3}$$

Computing the posterior marginals has the additional benefit of yielding (with minimal additional computation) the BLS estimates. For example, for the log excess compliance at node (i,j), we have

$$\hat{z}_{ij,\mathrm{BLS}} = \mathbb{E}\left[\mathsf{z}_{ij}|\mathsf{d} = \mathbf{d}\right] = \sum_{z_{ij}} z_{ij}\, p(z_{ij}|\mathbf{d}). \tag{3.4}$$

For any reasonably large number of grid nodes mn, the maximization and summation in Equations 3.2 and 3.3, respectively, are intractable, hence we must turn to approximate inference algorithms to perform the estimation. We discuss the inference algorithms used to approximate the MAP estimates and posterior marginals in Section 3.4.3.

3.4.1 Prior Model

Assuming that the fracture properties will not change rapidly with position, it is reasonable to make the properties at one point depend on its nearest neighbors in space. We capture this spatial dependence mathematically by carefully constructing an appropriate prior model for the fracture properties. We arrive at a prior model by defining the set of fracture properties m as a Markov random field over an undirected graphical model $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ on the 2-D grid $\mathbf{L}$ [35]. Here $\mathcal{V}$ is the set of vertices or nodes of the graphical model, which correspond to partitions of the random variables in m. We associate with each grid node $(i,j) \in \mathbf{L}$ a vertex in $\mathcal{V}$ corresponding to the

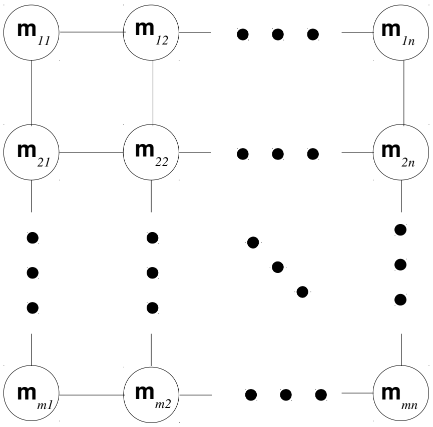


Figure 3-2: Undirected graphical model $\mathcal{G}$ over which m is Markov, prior to observing the seismic data. The model is based on the assumption that the values of the fracture properties at one location are dependent on those of its nearest neighbors.

pair of random variables $\mathsf{m}_{ij} = (\mathsf{z}_{ij}, \varphi_{ij})$, so that $\mathcal{V} \equiv \mathbf{L}$. The set of edges of the graph is $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$, represented as pairs of nodes, which encode dependencies between the random variables.

We define the edge set $\mathcal{E}$ over the 2-D grid $\mathbf{L}$ so that a particular node shares edges with its four neighbors on the grid $\mathbf{L}$. The graphical model for m prior to observing the data is shown in Figure 3-2. Intuitively, such a graph structure means that given the fracture properties of the four nearest neighbors of a particular node, knowledge of the fracture properties of the medium elsewhere on the grid will have no impact on our belief about the properties at that node. Since we expect the properties of the medium at a particular point in space to be similar to its surrounding properties, this suggests a prior distribution that penalizes differences between a node and its neighbors. In particular we define the prior distribution for the fracture parameters to be

$$p(\mathbf{z}) \propto \exp\left\{ - \sum_{(ij,kl)\in\mathcal{E}} \beta_{z_{ij,kl}} (z_{ij} - z_{kl})^2 \right\} \tag{3.5}$$

and

$$p(\boldsymbol{\varphi}) \propto \exp\left\{-\sum_{(ij,kl)\in\mathcal{E}} \beta_{\varphi_{ij,kl}}(\varphi_{ij}-\varphi_{kl})^2\right\}, \tag{3.6}$$

where $\beta_{z_{ij,kl}}$ and $\beta_{\varphi_{ij,kl}}$ are smoothness parameters. Note that the fracture orientations must be manipulated with a modulo operation to bring the difference within the interval $[-90°, 90°)$, as the azimuth is identical modulo 180°.

We define the smoothness parameters in terms of an overall (spatially-varying) smoothness parameter $\beta_{ij,kl}$ after normalizing by the bin sizes for each variable (so that the degree of smoothness is not dependent on the units of the variables) via $\beta_{z_{ij,kl}} = \beta_{ij,kl}/(0.1)^2$ and $\beta_{\varphi_{ij,kl}} = \beta_{ij,kl}/(20)^2$. Allowing spatial variation in the smoothness parameter allows us to encode *a priori* information about discontinuities in the medium, such as those which may arise from a geological fault. In general, we pick a single value $\beta_{\text{c}} > 0$ for the smoothness parameter along edges where there are no known discontinuities and set the smoothness parameter to zero along edges where a fault is known to exist (thereby removing those edges from the graph). We experiment with different choices for β_c in Section 3.5.2. Defining $\mathcal{E}_f$ to be the set of edges where a fault is known to exist, then

$$\beta_{ij,kl} = \begin{cases} \beta_c & \text{if } ((i,j),(k,l)) \in \mathcal{E}\setminus\mathcal{E}_f \\ 0 & \text{if } ((i,j),(k,l)) \in \mathcal{E}_f \end{cases} \tag{3.7}$$

Treating the two different fracture properties as independent *a priori* gives the overall prior distribution for the fracture properties as

$$p(\mathbf{m}) = p(\mathbf{z}, \boldsymbol{\varphi}) = p(\mathbf{z})p(\boldsymbol{\varphi}). \tag{3.8}$$

Indeed, we see that $p(\mathbf{m})$ factorizes over the maximal cliques of $\mathcal{G}$ (which are precisely the edges $\mathcal{E}$) and thus, by the Hammersley-Clifford theorem (see Ch. 2), $\mathbf{m}$ is Markov on $\mathcal{G}$.

3.4.2 Likelihood Model

As discussed in Section 3.3, the seismic data used in this study are AvAz and FTF data, which are extracted from the seismic trace dataset and denoted by $\mathbf{d}^{\mathrm{AvAz}}$ and $\mathbf{d}^{\mathrm{FTF}}$, respectively. We make the assumption that given the fracture parameters $\mathbf{m}$, the two types of seismic data $\mathbf{d}^{\mathrm{AvAz}}$ and $\mathbf{d}^{\mathrm{FTF}}$ are conditionally independent, and hence

$$p(\mathbf{d}|\mathbf{m}) = p(\mathbf{d}^{\mathrm{AvAz}}|\mathbf{m})p(\mathbf{d}^{\mathrm{FTF}}|\mathbf{m}). \tag{3.9}$$

In the remainder of this section, we discuss how we model the data to arrive at the likelihood models $p(\mathbf{d}^{\mathrm{AvAz}}|\mathbf{m})$ and $p(\mathbf{d}^{\mathrm{FTF}}|\mathbf{m})$.

Amplitude versus Azimuth Data

We suppose we have data for the amplitudes of P-P arrivals reflected from the top of the fractured layer at a full range of acquisition azimuths and source-receiver offsets for each grid point $(i, j) \in \mathbf{L}$. We can use ray-tracing to determine spatially the grid point corresponding to each source-receiver pair as well as to map the offsets to incidence angles for the wave incident on the top of the fractured layer (where the incidence angle is the angle the incident wave makes with the vertical axis). This gives, for each grid point (i, j) a set of P-P reflection amplitudes that vary with incidence angle $\theta \in \Theta$ and acquisition azimuth (relative to North) $\phi^{\mathrm{Acq}} \in \Phi^{\mathrm{Acq}}$, where Θ and Φ^{Acq} are the sets of incidence angles and acquisition azimuths over which the data has been obtained; we denote the reflection amplitudes by $\hat{R}_{ij}^{PP}(\theta, \phi^{\mathrm{Acq}})$. For concreteness, suppose the acquisition azimuths we have are precisely the set $\Phi^{\mathrm{Acq}} = \{0°, 10°, \ldots, 170°\}$. In order to be able to compare these amplitudes to the P-wave reflection coefficient, for each incidence angle θ, we normalize the amplitudes by the mean amplitude (taken over acquisition azimuths). This allows us to compare the variation of the reflection coefficient with azimuth (rather than its absolute value):

$$\mathbf{d}_{ij}^{\mathrm{AvAz}} = d_{ij,\theta,\phi^{\mathrm{Acq}}}^{\mathrm{AvAz}} = \frac{\hat{R}_{ij}^{PP}(\theta, \phi^{\mathrm{Acq}})}{\frac{1}{|\Phi^{\mathrm{Acq}}|}\sum_{\phi\in\Phi^{\mathrm{Acq}}} \hat{R}_{ij}^{PP}(\theta, \phi)} \tag{3.10}$$

In order to arrive at a forward model for the P-wave reflection coefficient of the interface above the fractured layer as a function of acquisition azimuth, we make various simplifying assumptions about the formation and the fractured medium. The layers above the fractured layer are assumed to be isotropic and homogeneous and the background medium of the layer in which the fractures exist is assumed to be homogeneous and isotropic with known medium parameters. We assume that the presence of fractures in the fractured layer causes the layer to behave as an equivalent anisotropic medium, which is a geophysically valid assumption when the fracture spacing is small compared to the seismic wavelength [65, 78]. In this case, it is reasonable to assume that the presence of a parallel set of vertical fractures causes the medium to exhibit horizontal transverse isotropy (HTI) with a symmetry axis normal to the strike of the fractures [60, 72]. A transverse isotropic medium with a given symmetry axis means seismic wave propagation in all directions that form the same angle with the symmetry axis is equivalent. As such, in an HTI medium resulting from a set of parallel vertical fractures, the plane normal to the symmetry axis (and parallel to the fractures) is referred to as the isotropy plane, as wave propagation is equivalent in all directions in this plane [72].

The P-wave reflection coefficient of an interface is defined as the ratio of the reflected P-wave amplitude to the incident P-wave amplitude on the interface. Rüger [60] derives the P-wave polarization vector and P-wave phase velocities in an HTI medium and uses these to solve a system of perturbation equations for the reflection and transmission coefficients at the interface of two HTI media having the same symmetry axes. The resultant P-wave reflection coefficient is given as a function of the incidence phase angle (θ, the angle the incident P-wave makes with the vertical axis) and the azimuthal phase angle (ϕ, the azimuth of the incident P-wave relative to the symmetry axis), and in terms of the isotropic background and anisotropy parameters, as

$$
\begin{aligned}
R^{PP}(\theta,\phi) = \frac{1}{2}\frac{\Delta Z}{\bar{Z}} + \frac{1}{2}\left(\frac{\Delta\alpha}{\bar{\alpha}} - \left(\frac{2\bar{\beta}}{\bar{\alpha}}\right)^2 \frac{\Delta G}{\bar{G}} + \left(\Delta\delta^{(V)} - 2\left(\frac{2\bar{\beta}}{\bar{\alpha}}\right)^2 \Delta\gamma^{(V)}\right)\cos^2\phi\right)\sin^2\theta \\
+ \frac{1}{2}\left(\frac{\Delta\alpha}{\bar{\alpha}} + \Delta\epsilon^{(V)}\cos^4\phi + \Delta\delta^{(V)}\sin^2\phi\cos^2\phi\right)\sin^2\theta\tan^2\theta,
\end{aligned}
\tag{3.11}
$$

where α is the vertical P-wave velocity, β is the vertical velocity of the S-wave polarized parallel to the isotropy plane, ρ is the medium density, $Z = \rho\alpha$ is the vertical P-wave impedance, and $G = \rho\beta^2$ is the vertical shear modulus; these parameters are all from the background isotropic model and are assumed to be known in our analysis. Rüger [61] and Liu and Martinez [41] note that this linearized equation for the P-wave reflection coefficient is accurate for small medium contrasts and weak anisotropy at angles of incidence less than 35°. In cases where a more accurate model is required, one may wish to use the approximations of the P-wave reflection coefficient given by Ursin and Haugen [73] or Pšenčík and Martins [58]. The parameters $\delta^{(V)}, \epsilon^{(V)}, \gamma^{(V)}$ are the Thomsen anisotropy parameters defined with respect to the vertical axis [71]; these parameters are identically zero for an isotropic medium, but will depend on the fracture properties for an HTI medium. The parameters in Equation 3.11 are defined in terms of their relative differences between the upper and lower media $\Delta(\cdot)$ and their average values $\bar{(\cdot)}$. So, for example, $\Delta\alpha = \alpha_2 - \alpha_1$ and $\bar{\alpha} = (\alpha_1 + \alpha_2)/2$, where α_1 and α_2 are the vertical P-wave velocities of the upper and lower media, respectively. Since the axis of symmetry is normal to the strike of the fractures (thus having an azimuth relative to North of $\varphi_{ij} + 90°$), then with ϕ^{Acq} as the known azimuth of the incident P-wave relative to North, we have

$$
\phi = \phi^{\text{Acq}} - \varphi_{ij} - 90°. \tag{3.12}
$$

Furthermore, we set the incidence angle θ to the values computed for the AvAz data.

We use the linear slip model of Schoenberg and Sayers [66] to express the Thomsen anisotropy parameters of the fractured medium in terms of the excess fracture compliance. The details of this derivation are given in Appendix A. Combining this with (3.11) gives the forward model for the P-P reflection coefficient as a function of the fracture parameters at node (i, j), which we denote by $R_{ij}^{PP}(\theta, \phi^{\text{Acq}}, \varphi_{ij}, z_{ij})$. To make this comparable to the data $\mathbf{d}_{ij}^{\text{AvAz}}$ defined in (3.10), we process it in the same manner by normalizing by the mean reflection coefficient, for each incidence angle θ, over all acquisition azimuths, giving

$$\bar{R}_{ij}^{PP}(\theta, \phi^{\text{Acq}}, \varphi_{ij}, z_{ij}) = \frac{R_{ij}^{PP}(\theta, \phi^{\text{Acq}}, \varphi_{ij}, z_{ij})}{\frac{1}{|\Phi^{\text{Acq}}|}\sum_{\phi \in \Phi^{\text{Acq}}} R_{ij}^{PP}(\theta, \phi, \varphi_{ij}, z_{ij})} \tag{3.13}$$

as the deterministic forward model for $d_{ij,\theta,\phi^{\text{Acq}}}^{\text{AvAz}}$. We note that if the excess fracture compliance is zero, then the P-P reflection coefficient is constant with respect to acquisition azimuth, and hence does not vary with the fracture orientation φ_{ij}. This is consistent with our interpretation of zero compliance to mean the absence of fractures, which indeed renders the value of φ_{ij} arbitrary.

We arrive at a stochastic model for the data by assuming the output of the forward model is perturbed by zero-mean additive independent, identically distributed (i.i.d.) Gaussian noise, so that

$$\mathsf{d}_{ij,\theta,\phi^{\text{Acq}}}^{\text{AvAz}} = \bar{R}_{ij}^{PP}(\theta, \phi^{\text{Acq}}, \varphi_{ij}, z_{ij}) + \mathsf{w}_{ij,\theta,\phi^{\text{Acq}}} \tag{3.14}$$

where the $\mathsf{w}_{ij,\theta,\phi^{\text{Acq}}}$ are mutually independent Gaussian random variables distributed as $\mathcal{N}(0, \sigma_{ij,\text{AvAz}}^2)$. We estimate the variance $\sigma_{ij,\text{AvAz}}^2$ using synthetic data obtained from a finite-difference simulation of the seismic wavefield; the details of the synthetic data are described in Section 3.5.1. Processing of the synthetic data gives a set of single observations of the data $d_{ij,\theta,\phi^{\text{Acq}}}^{\text{AvAz}}$ at a single grid node and at a range of incidence angles and acquisition azimuths, where the fracture properties (z_{ij}, φ_{ij}) are known, thus giving independent samples for the noise

$$\mathbf{w}_{ij} = w_{ij,\theta,\phi^{\text{Acq}}} = d_{ij,\theta,\phi^{\text{Acq}}}^{\text{AvAz}} - \bar{R}_{ij}^{PP}(\theta, \phi^{\text{Acq}}, \varphi_{ij}, z_{ij}). \tag{3.15}$$

We estimate $\sigma^2_{ij,\mathrm{AvAz}}$ via its maximum likelihood (ML) estimator, which is given by

$$\hat{\sigma}^2_{ij,\mathrm{AvAz},ML} = \hat{\sigma}^2_{ij,\mathrm{AvAz},ML}\left(\mathbf{w}_{ij}\right) = \frac{1}{|\Theta|\,|\Phi^{\mathrm{Acq}}|}\sum_{\theta\in\Theta}\sum_{\phi\in\Phi^{\mathrm{Acq}}} w^2_{ij,\theta,\phi}. \tag{3.16}$$

Note that, in contrast to the ML estimator of the combined variance and mean of a normal random variable, the ML estimator in Equation 3.16 is unbiased, that is $\mathbb{E}\left[\hat{\sigma}^2_{ij,\mathrm{AvAz},ML}\left(\mathbf{w}_{ij}\right)\right] = \sigma^2_{ij,\mathrm{AvAz}}$. Having fully described the stochastic model for the data, we are now in a position to give an expression for the likelihood model for $\mathbf{d}^{\mathrm{AvAz}}$, which is:

$$p(\mathbf{d}^{\mathrm{AvAz}}|\mathbf{m}) = \prod_{(i,j)\in\mathbf{L}}\left(\prod_{\theta\in\Theta}\prod_{\phi^{\mathrm{Acq}}\in\Phi^{\mathrm{Acq}}}\mathcal{N}\left(d^{\mathrm{AvAz}}_{ij,\theta,\phi^{\mathrm{Acq}}};\, \bar{R}^{PP}_{ij}(\theta,\phi^{\mathrm{Acq}},\varphi_{ij},z_{ij}), \hat{\sigma}^2_{ij,\mathrm{AvAz},ML}\right)\right), \tag{3.17}$$

where $\mathcal{N}(\,\cdot\,;\mu,\sigma^2)$ is the Gaussian probability density function (PDF) with mean μ and variance σ^2.

Fracture Transfer Function Data

We further suppose that we have what Fang et al. [22] describe as fracture transfer function data. We briefly describe the definition of the FTF data and how it is computed, which will result in a natural choice for our data $\mathbf{d}^{\mathrm{FTF}}$ and its likelihood model $p(\mathbf{d}^{\mathrm{FTF}}|\mathbf{m})$.

Intuitively, the fracture transfer function is the transfer function from the seismic wavefield reflected off the top of the fractured layer to the wavefield propagating out of the fractured layer after reflecting off the bottom of this layer. In other words, it quantifies the redistribution of energy of the reflected and scattered seismic wavefield after passing through the fractured layer. A cartoon depicting this is shown in Figure 3-3.

FTF is inherently a function of the propagation azimuth of the incident and reflected waves. At fracture spacings on the order of the seismic wavelength, the orien-

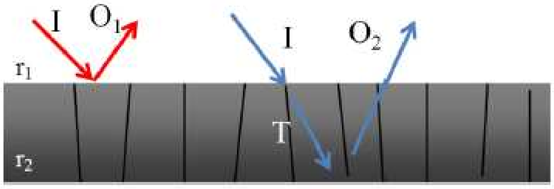


Figure 3-3: A cartoon depicting the meaning of fracture transfer function for layer r_2. $I(\omega)$ is the incident wavefield, $T(\omega)$ is the transmitted wavefield into the fractured layer, and $O_1(\omega)$ and $O_2(\omega)$ are the waves reflected by layers above and below the fracture zone, respectively. Theoretically, the the fracture transfer function at angular frequency ω is defined as $FTF(\omega) = \frac{O_2(\omega)}{O_1(\omega)}$. Figure adapted from Fang et al. [22].

tation of the fractures relative to the propagation azimuth has a significant effect on the amplitude of the scattered wavefield reflected off the bottom of the fractured layer. In particular, when fractures are parallel to the propagation azimuth, the fractures tend to act as waveguides, directing more of the scattered energy back to the surface in the direction away from the source. However, when the fractures are normal to the propagation azimuth, the scattered energy is less coherent as the fractures tend to scatter energy in both forward and backward directions. With this in mind, we expect FTF to be maximized at propagation azimuths parallel to the fractures.

According to the methodology given by Fang et al. [22] and Fang et al. [21], the FTF at a particular spatial grid point (i, j) is estimated from surface seismic data by first determining (via ray-tracing) all source-receiver pairs corresponding to grid point (i, j). Then, for all source-receiver pairs within the same acquisition azimuth bin ϕ^{Acq}, normal moveout to zero offset is applied to the seismic traces which are then stacked. The result of this procedure gives a single, stacked seismic trace for each acquisition azimuth. The arrivals on the traces corresponding to reflections off the top and bottom of the fractured layer are then located in the stacked trace and windowed, giving windowed arrivals for each acquisition azimuth $o_1^{ij}(t, \phi^{\text{Acq}})$ and $o_2^{ij}(t, \phi^{\text{Acq}})$, respectively. The Fourier transforms $O_1^{ij}(\omega, \phi^{\text{Acq}})$ and $O_2^{ij}(\omega, \phi^{\text{Acq}})$ of the windowed arrivals are taken, and we compute the fracture transfer function at angular frequency ω and acquisition azimuth ϕ^{Acq} as

$$FTF^{ij}(\omega, \phi^{\text{Acq}}) = \frac{O_2^{ij}(\omega, \phi^{\text{Acq}})}{O_1^{ij}(\omega, \phi^{\text{Acq}})}. \tag{3.18}$$

This is reduced to a function of only acquisition azimuth by integrating out the angular frequency via a weighted integral. The idea is that frequencies at which there is greater variability in FTF with acquisition azimuth should be given more weight, hence a frequency weighting function $W^{ij}(\omega)$ is defined as the standard deviation of $FTF^{ij}(\omega, \cdot)$ with respect to acquisition azimuth, so that

$$\overline{FTF}^{ij}(\phi^{\text{Acq}}) = \int_{\omega} FTF^{ij}(\omega, \phi^{\text{Acq}}) W^{ij}(\omega)\, d\omega. \tag{3.19}$$

Due to the reasons mentioned above, we expect $\overline{FTF}^{ij}(\phi^{\mathrm{Acq}})$ to be maximized at φ_{ij} if fractures are present in the medium. On the other hand, in the absence of fractures, we expect there will not be a unique maximizer for $\overline{FTF}^{ij}(\phi^{\mathrm{Acq}})$. Hence, it is natural to define the FTF data used in our analysis as $\boldsymbol{\mathsf{d}}_{ij}^{\mathrm{FTF}} = \big(d_{ij,1}^{\mathrm{FTF}}, d_{ij,2}^{\mathrm{FTF}}\big)$, where $d_{ij,1}^{\mathrm{FTF}} \in \{0,1\}$ is an indicator variable set to 0 when there is no unique maximizer (within a numerical threshold) for $\overline{FTF}^{ij}(\phi^{\mathrm{Acq}})$, and set to 1 otherwise and where $d_{ij,2}^{\mathrm{FTF}}$ is set to the acquisition azimuth that maximizes $\overline{FTF}^{ij}(\phi^{\mathrm{Acq}})$:

$$d_{ij,2}^{\mathrm{FTF}} \triangleq \underset{\phi^{\mathrm{Acq}} \in \Phi^{\mathrm{Acq}}}{\arg\max}\, \overline{FTF}^{ij}(\phi^{\mathrm{Acq}}). \tag{3.20}$$

If there is no unique maximizing ϕ^{Acq}, we arbitrarily set $d_{ij,2}^{\mathrm{FTF}}$ to any one maximizing value.

We define a stochastic forward model for $\boldsymbol{\mathsf{d}}_{ij}^{\mathrm{FTF}}$ by first assuming that, given the fracture properties $\boldsymbol{\mathsf{m}}_{ij}$ at node (i,j), $\boldsymbol{\mathsf{d}}_{ij}^{\mathrm{FTF}}$ is conditionally independent of the fracture properties and FTF data at the remaining nodes.

We define ζ_{ij} as the probability that $\mathsf{d}_{ij,1}^{\mathrm{FTF}}$ correctly predicts whether or not there are fractures present at node (i,j). Then, given the fracture properties, we model $\mathsf{d}_{ij,1}^{\mathrm{FTF}}$ as a Bernoulli random variable. Thus, given $\mathsf{z}_{ij} = -13$ (i.e. zero excess fracture compliance) then $\mathsf{d}_{ij,1}^{\mathrm{FTF}} = 0$ with probability ζ_{ij} and $\mathsf{d}_{ij,1}^{\mathrm{FTF}} = 1$ with probability $1-\zeta_{ij}$, and given $\mathsf{z}_{ij} > -13$ (i.e. non-zero excess fracture compliance) then $\mathsf{d}_{ij,1}^{\mathrm{FTF}} = 1$ with probability ζ_{ij} and $\mathsf{d}_{ij,1}^{\mathrm{FTF}} = 0$ with probability $1-\zeta_{ij}$. That is,

$$p(d_{ij,1}^{\mathrm{FTF}} \,|\, \mathbf{m}_{ij}\,;\, \zeta_{ij}) = \begin{cases} \zeta_{ij} & \text{if } d_{ij,1}^{\mathrm{FTF}} = \mathbb{1}_{\{z_{ij} > -13\}} \\ 1-\zeta_{ij} & \text{if } d_{ij,1}^{\mathrm{FTF}} = 1 - \mathbb{1}_{\{z_{ij} > -13\}} \end{cases}, \tag{3.21}$$

where $\mathbb{1}_{\{\cdot\}}$ is the indicator function defined as

$$\mathbb{1}_{\{A\}} \triangleq \begin{cases} 1 & \text{if } A \\ 0 & \text{otherwise} \end{cases}.$$

Now, given $\mathsf{d}_{ij,1}^{\mathrm{FTF}}$ and the fracture properties, if either $\mathsf{z}_{ij} = -13$, so that there are no fractures present, or if $\mathsf{d}_{ij,1}^{\text{FTF}} = 0$, so that a unique preferential scattering direction was not identified, then any value for $\mathsf{d}_{ij,2}^{\text{FTF}} \in [0^\circ, 180^\circ)$ is arbitrary, so we model $\mathsf{d}_{ij,2}^{\text{FTF}}$ as uniform on the set $[0^\circ, 180^\circ)$. Otherwise, if both $\mathsf{z}_{ij} > -13$ and $\mathsf{d}_{ij,1}^{\text{FTF}} = 1$, then $\mathsf{d}_{ij,2}^{\text{FTF}}$ should be near the true fracture orientation φ_{ij}. As with the AvAz data, we model this by introducing additive zero-mean independent Gaussian noise, so that conditioned on the event $\{\mathsf{z}_{ij} \neq 0, \mathsf{d}_{ij,1}^{\text{FTF}} = 1\}$, we have:

$$\mathsf{d}_{ij,2}^{\text{FTF}} = \varphi_{ij} + \mathsf{v}_{ij} \tag{3.22}$$

where $\mathsf{v}_{ij} \sim \mathcal{N}(0, \sigma_{ij,\text{FTF}}^2)$ and $\{\mathsf{v}_{ij} : (i,j) \in L\}$ is a collection of mutually independent random variables. This gives the conditional distribution for $d_{ij,2}^{\text{FTF}}$ as

$$p\left(d_{ij,2}^{\text{FTF}} \mid d_{ij,1}^{\text{FTF}}, \mathbf{m}_{ij}\,;\, \sigma_{ij,\text{FTF}}^2\right) = \begin{cases} \frac{1}{180}\mathbb{1}_{\left\{d_{ij,2}^{\text{FTF}} \in [0,180)\right\}} & \text{if } z_{ij} = -13 \text{ or } d_{ij,1}^{\text{FTF}} = 0 \\ \mathcal{N}\left(d_{ij,2}^{\text{FTF}}\,;\, \varphi_{ij}, \sigma_{ij,\text{FTF}}^2\right) & \text{if } z_{ij} > -13 \text{ and } d_{ij,1}^{\text{FTF}} = 1 \end{cases}. \tag{3.23}$$

As with the AvAz data, we use synthetic data from finite-difference simulations of the seismic wavefield (described in Section 3.5.1), to obtain a set of K independent samples $\mathbf{D}_{ij} = (\mathbf{m}_{ij}^{(k)}, \mathbf{d}_{ij}^{\text{FTF}(k)})_{k=1,\ldots,K}$ of the fracture properties and FTF data at a single grid point, arising from the simulation of K different fracture models. The ML estimate for ζ_{ij} is simply the fraction of times $d_{ij,1}^{\text{FTF}(k)}$ correctly takes the value 0 (when $z_{ij}^{(k)} = -13$) or 1 (when $z_{ij}^{(k)} > -13$). For the synthetic data we have obtained, this fraction turns out to be 1. However, in order to preserve stochasticity in detecting the presence of fractures from FTF data, we instead estimate ζ_{ij} under a Bayesian approach by treating it as random variable with a prior distribution that is uniform over $[0,1]$. Having observed K correct observations (and 0 incorrect observations) of $\mathsf{d}_{ij,1}^{\text{FTF}}$, the posterior distribution for ζ_{ij} is a Beta distribution $\zeta_{ij} \sim \text{Beta}(K+1, 1)$.

Integrating out ζ_{ij} in the likelihood model for $\mathsf{d}^{\mathrm{FTF}}_{ij,1}$, given the data, we have

$$\begin{aligned}
p\left(d^{\mathrm{FTF}}_{ij,1}\,|\,\mathbf{m}_{ij}, \mathbf{D}_{ij}\right) &= \int_0^1 p(\zeta_{ij}|\mathbf{m}_{ij}, \mathbf{D}_{ij})\, p(d^{\mathrm{FTF}}_{ij,1}\,|\,\mathbf{m}_{ij}, \mathbf{D}_{ij}, \zeta_{ij})\, d\zeta_{ij} && (3.24)\\
&= \int_0^1 p(\zeta_{ij}|\mathbf{D}_{ij})\, p(d^{\mathrm{FTF}}_{ij,1}\,|\,\mathbf{m}_{ij}, \zeta_{ij})\, d\zeta_{ij} && (3.25)\\
&= \begin{cases} \frac{1}{B(K+1,1)} \int_0^1 \zeta_{ij}^{K+1}(1-\zeta_{ij})^0\, d\zeta_{ij} & \text{if } d^{\mathrm{FTF}}_{ij,1} = \mathbb{1}_{\{z_{ij}>-13\}} \\ \frac{1}{B(K+1,1)} \int_0^1 \zeta_{ij}^{K}(1-\zeta_{ij})^1\, d\zeta_{ij} & \text{if } d^{\mathrm{FTF}}_{ij,1} = 1 - \mathbb{1}_{\{z_{ij}>-13\}} \end{cases} && (3.26)\\
&= \begin{cases} \frac{B(K+2,1)}{B(K+1,1)} & \text{if } d^{\mathrm{FTF}}_{ij,1} = \mathbb{1}_{\{z_{ij}>-13\}} \\ \frac{B(K+1,2)}{B(K+1,1)} & \text{if } d^{\mathrm{FTF}}_{ij,1} = 1 - \mathbb{1}_{\{z_{ij}>-13\}} \end{cases} && (3.27)\\
&= \begin{cases} \frac{K+1}{K+2} & \text{if } d^{\mathrm{FTF}}_{ij,1} = \mathbb{1}_{\{z_{ij}>-13\}} \\ \frac{1}{K+2} & \text{if } d^{\mathrm{FTF}}_{ij,1} = 1 - \mathbb{1}_{\{z_{ij}>-13\}} \end{cases} && (3.28)
\end{aligned}$$

where $B(\cdot,\cdot)$ is the beta function. We are left with K' i.i.d. samples where $z^{(k)}_{ij} > -13$ (and $d^{\mathrm{FTF}}_{ij,1} = 1$), giving samples of the additive Gaussian noise

$$\left\{v^{(k')}_{ij}\right\}_{k'=1,\ldots,K'} = \left\{\varphi^{(k)}_{ij} - d^{\mathrm{FTF}(k)}_{ij,2} : z^{(k)}_{ij} > -13,\, d^{\mathrm{FTF}(k)}_{ij,1} = 1\right\}$$

used to give the ML estimation of the variance

$$\hat{\sigma}^2_{ij,\mathrm{FTF},ML} = \hat{\sigma}^2_{ij,\mathrm{FTF},ML}\left(\left\{v^{(k')}_{ij}\right\}_{k'=1,\ldots,K'}\right) = \frac{1}{K'}\sum_{k'=1}^{K'}\left(v^{(k')}_{ij}\right)^2. \tag{3.29}$$

As before, the ML estimator in Equation 3.29 is unbiased. This gives the likelihood model $p(\mathbf{d}^{\mathrm{FTF}}|\mathbf{m})$ as:

$$p\left(\mathbf{d}^{\mathrm{FTF}}\,|\,\mathbf{m}\right) = \prod_{(i,j)\in\mathbf{L}} p\left(d^{\mathrm{FTF}}_{ij,1}\,|\,\mathbf{m}_{ij}, \mathbf{D}_{ij}\right) p\left(d^{\mathrm{FTF}}_{ij,2}\,|\,d^{\mathrm{FTF}}_{ij,1}, \mathbf{m}_{ij}\,;\,\hat{\sigma}^2_{ij,\mathrm{FTF},ML}\right) \tag{3.30}$$

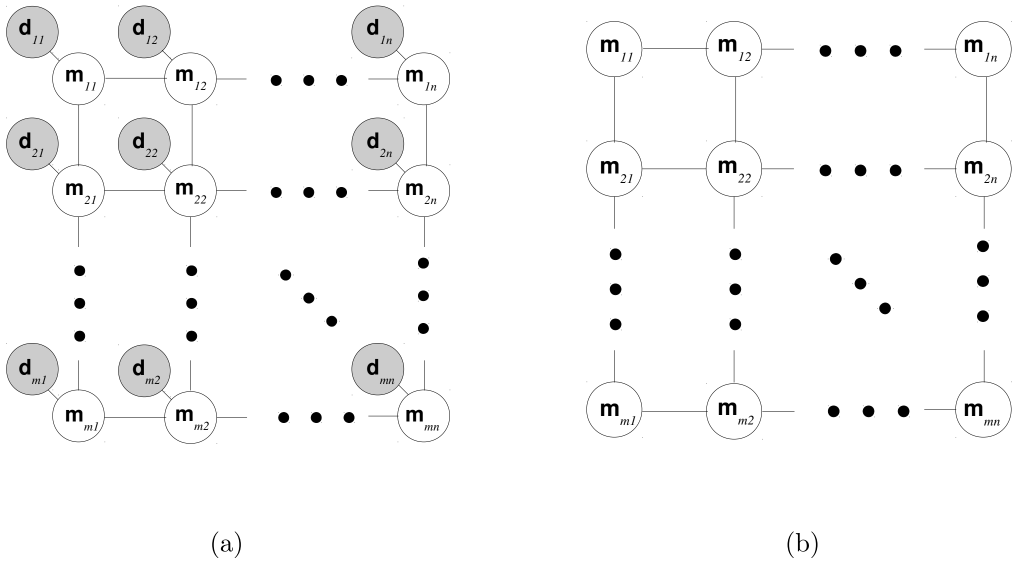


Figure 3-4: (a) Graphical model showing the Markovianity between the observations $\mathbf{d}$ and the fracture parameters $\mathbf{m}$. (b) Graphical model for the posterior distribution after removing the observed nodes.

3.4.3 Inference Algorithms

We return to the graphical model representation of the distribution, as this will play a key role in the inference algorithms used to obtain the posterior marginals and MAP estimate. Having defined the prior and likelihood models, the posterior distribution is given by Equation 3.1. We immediately notice that given the fracture properties of a particular grid node (i, j), the observations d_{ij} at that grid node are conditionally independent of the remaining fracture properties and observations, that is, with m_{-ij} and d_{-ij} defined as in Equation 3.3, $\mathsf{d}_{ij} \perp\!\!\!\perp \{\mathsf{m}_{-ij}, \mathsf{d}_{-ij}\} \,|\, \mathsf{m}_{ij}$; this is depicted in Figure 3-4(a).

Having observed the data $\mathsf{d} = \mathbf{d}$, the data are no longer random and hence separate nodes for the data are not included in the graphical model for the posterior distribution. Hence, we can write the posterior distribution in terms of the node and edge potentials of the graph, ψ_{ij} and $\psi_{ij,kl}$, respectively, where the node potentials capture the effect of the data and the edge potentials capture the prior distribution. These potentials are given by:

$$\psi_{ij}(\mathbf{m}_{ij}) = p(\mathbf{d}_{ij}^{\mathrm{AvAz}} \,|\, \mathbf{m}_{ij}) p(\mathbf{d}_{ij}^{\mathrm{FTF}} \,|\, \mathbf{m}_{ij}) \tag{3.31}$$

and

$$\psi_{ij,kl}(\mathbf{m}_{ij}, \mathbf{m}_{kl}) = \exp\{-\beta_{z_{ij,kl}}(z_{ij} - z_{kl})^2 - \beta_{\varphi_{ij,kl}}(\varphi_{ij} - \varphi_{kl})^2\}, \tag{3.32}$$

and the posterior distribution is then given as:

$$p(\mathbf{m}|\mathbf{d}) \propto \prod_{(i,j)\in\mathcal{V}} \psi_{ij}(\mathbf{m}_{ij}) \prod_{(ij,kl)\in\mathcal{E}} \psi_{ij,kl}(\mathbf{m}_{ij}, \mathbf{m}_{kl}). \tag{3.33}$$

Having fully described the posterior distribution in terms of its graphical model and node and edge potentials, we are able to apply belief propagation algorithms to perform approximate inference of the fracture properties m.

Loopy Belief Propagation

Belief propagation (BP) is a technique for performing inference on graphical models which has recently enjoyed much popularity for use amongst a wide-range of applications [48, 55, 56]. Originally formulated for tree graphs (i.e. graphs having no cycles), BP refers to message-passing algorithms for computing either marginal distributions (called the sum-product algorithm) or MAP configurations (called the max-product algorithm). In particular, for an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, the BP algorithm computes messages (denoted by $\mu_{ij\to kl}(m_{kl})$) from each node $(i,j) \in \mathcal{V}$ to every node (k,l) which shares an edge with (i,j) (called a 'neighbor' of (i,j)); these messages capture the beliefs node (i,j) carries about its neighbors. The messages are iteratively propagated from each node to its neighbors, hence the name 'belief propagation.'

The sum-product variant of the BP algorithm [35] with node and edge potentials ψ_{ij} and $\psi_{ij,kl}$ is given by the update equations

$$\mu^{(0)}_{ij\to kl}(m_{kl}) \propto 1 \tag{3.34}$$

$$\mu^{(t+1)}_{ij\to kl}(m_{kl}) \propto \sum_{m_{ij}} \psi_{ij}(m_{ij})\psi_{ij,kl}(m_{ij}, m_{kl}) \prod_{uv\in\mathbf{Nb}(ij)\setminus\{kl\}} \mu^{(t)}_{uv\to ij}(m_{ij}) \tag{3.35}$$

$$\hat{p}^{(t)}_{ij}(m_{ij}) \propto \psi_{ij}(m_{ij}) \prod_{uv\in\mathbf{Nb}(ij)} \mu^{(t)}_{uv\to ij}(m_{ij}) \tag{3.36}$$

$\forall (i,j) \in \mathcal{V}$, $((i,j),(k,l)) \in \mathcal{E}$, where $\mathbf{Nb}(ij)$ denotes the set of neighbors of node (i,j) in $\mathcal{G}$ and $\hat{p}^{(t)}_{ij}(m_{ij})$ is the estimate of the marginal for node (i,j) at iteration t. One can verify that if the underlying graphical model $\mathcal{G}$ is a tree, then the sum-product algorithm converges to the true marginal distributions in a number of iterations equal to the diameter of the tree [35]. The max-product algorithm is similarly defined by replacing summations with maximizations

$$\mu^{(0)}_{ij\to kl}(m_{kl}) \propto 1 \tag{3.37}$$

$$\mu^{(t+1)}_{ij\to kl}(m_{kl}) \propto \max_{m_{ij}} \psi_{ij}(m_{ij})\psi_{ij,kl}(m_{ij}, m_{kl}) \prod_{uv\in\mathbf{Nb}(ij)\setminus\{kl\}} \mu^{(t)}_{uv\to ij}(m_{ij}) \tag{3.38}$$

$$\hat{\bar{p}}^{(t)}_{ij}(m_{ij}) \propto \psi_{ij}(m_{ij}) \prod_{uv\in\mathbf{Nb}(ij)} \mu^{(t)}_{uv\to ij}(m_{ij}) \tag{3.39}$$

$\forall(i,j) \in \mathcal{V}$, $((i,j),(k,l)) \in \mathcal{E}$, where $\hat{\bar{p}}^{(t)}_{ij}(m_{ij})$ is the estimate of the node max-marginal $\bar{p}_{ij}(m_{ij})$ for node (i,j) at iteration t. The node max-marginal (at node (i,j)) is defined to be the function of m_{ij} one would obtain by fixing the random variable at node (i,j) to the value m_{ij} and then maximizing the joint distribution $p(\mathbf{m})$ over all other random variables. That is:

$$\bar{p}_{ij}(m_{ij}) \triangleq \max_{\mathbf{m}_{-ij}} p(m_{ij}, \mathbf{m}_{-ij}). \tag{3.40}$$

It is important to note that the node max-marginals are *not* the marginal distributions, and in fact they are not even probability distributions. However, they can be used to readily obtain the MAP estimate of $\mathbf{m}$. In particular, if the node max-marginals $\bar{p}_{ij}(m_{ij})$ have unique maximizers m^*_{ij}, then the MAP estimate is simply the vector of these unique maximizing values for each node:

$$\hat{\mathbf{m}}_{\text{MAP}} = \left[m^*_{ij}\right]_{ij\in\mathcal{V}} = \left[\arg\max_{m_{ij}} \bar{p}_{ij}(m_{ij})\right]_{ij\in\mathcal{V}}. \tag{3.41}$$

Thus, the estimated node max-marginals $\hat{\bar{p}}^{(t)}_{ij}(m_{ij})$ obtained from the max-product algorithm can be used to approximate the MAP estimate. Again, if $\mathcal{G}$ is a tree, then it can be shown that the max-product algorithm converges to the true node max-marginals and will hence produce the exact MAP estimate.

While BP was originally intended for tree graphs (and indeed converges to the correct result on trees), it can still be applied to graphs which are not trees, such as the grid graph for our posterior distribution. Applying BP to perform inference on a graph with loops is referred to as loopy belief propagation. While LBP is an approximate algorithm, as it does not, in general, converge to the correct answer, it has nonetheless been used extensively in various settings and found to often give very good approximations, particularly on graphs with a relatively sparse edge set (such as our 2-D grid graph) and when the node potentials are strong relative to the edge potentials [48]. With this in mind, we apply loopy belief propagation on the posterior distribution for the fracture parameters to approximate the MAP configuration and marginal distributions.

3.5 Results

3.5.1 Synthetic Data

We validate our methodology by performing inference on a synthetic data set. The synthetic data are obtained from a 3-D elastic finite-difference simulation of the seismic wavefield on reservoir models having topology as shown in Figure 3-1, and is the same data referenced in the fracture transfer function study by Fang et al. [22]. Each of the models consists of five flat homogeneous layers, with fractures in the third layer; the remaining layers are isotropic. Following the methodology of Coates and Schoenberg [14] to simulate discrete fractures, the finite-difference grid cells intersecting individual fractures are modeled as anisotropic, and the individual fractures are spaced uniformly within an isotropic background medium (note the finite-difference grid cells are distinct from and on a much smaller scale than the grid nodes on which the random variables are defined). The isotropic background parameters are given in Table 3.1. It is important to note that we are *not* modeling the entire fractured layer as anisotropic, but only the individual fractures; hence the validity of this model does not depend on the fracture spacing.

The fractured layer in the models contains a single set of discrete parallel fractures with strike 0° and individual normal and tangential fracture compliances of 10^{-9} m/Pa (note that the fracture compliance is distinct from the excess fracture compliance,

Layer	Thickness (m)	α (m/s)	β (m/s)	ρ (g/cm^3)
1	200	3000	1765	2.20
2	200	3500	2060	2.25
3	200	4000	2353	2.30
4	200	3500	2060	2.25
5	200	4000	2353	2.30

Table 3.1: Isotropic background parameters for the finite-difference synthetic data.

which is the ratio of fracture compliance to fracture spacing). The models differ from one another by fracture spacing, where synthetic data have been obtained from models having fracture spacings of 12 m, 20 m, 40 m, 60 m, 80 m, and 100 m, and where the fracture parameters in a particular model are constant over the entire layer. The synthetic seismic trace dataset for each model is obtained from a 2-D array of surface seismic receivers spaced 4 m apart responding to a single Ricker source wavelet having central frequency of 40 Hz. AvAz data are computed from the seismic trace dataset by using ray-tracing to compute the arrival time of the P-P arrival reflected from the top of Layer 3 for each receiver and taking the amplitude of this arrival in the seismic trace, where the source-receiver offsets and acquisition azimuths are known. Since the layers are flat and homogeneous, the reflection from all points on the horizontal grid are equivalent on average. Hence, the data are treated as the generic AvAz data for a single grid node or common depth point (CDP). Similarly, the FTF data are computed according to the procedure in Section 3.4.2 and treated as the generic FTF data corresponding to a single node or CDP. Prior to processing the synthetic data, we perturb the raw seismic traces with zero-mean Gaussian noise, with a standard deviation of 5% of the peak amplitude of the data, where a different realization of the noise is used for each CDP gather.

To obtain measurements for the entire grid $\mathbf{L}$, we predetermine the fracture parameters $\mathbf{m}$ over $\mathbf{L}$. For each node $(i, j) \in \mathbf{L}$, we are free to choose any fracture strike in the full range of azimuths $[0^\circ, 180^\circ)$, as we can simply rotate the synthetic data from 0° to any desired azimuth φ_{ij}. For the log excess compliance z_{ij}, we are able to use any value obtained from the models corresponding to excess compliances that can be achieved using fracture compliance of 10^{-9} m/Pa divided by any of the spacings for which synthetic data have been obtained. Having set the desired fracture properties across the grid, we map the noisy synthetic data to specific values across the grid $\mathbf{L}$. Processing the noisy data according to the procedure described in Section 3.4.2 results in our vector of noisy measurements $\mathbf{d}$ corresponding to the known fracture properties $\mathbf{m}$ over the entire grid. A grid spacing of $\ell = 200$ m is used for $\mathbf{L}$.

3.5.2 Results of Inference Procedure

We perform the inference on synthetic data arising from two scenarios. The first scenario is given by a 20-by-20 node grid of a single fracture set (so that the fracture properties are constant along the grid), for each of the available fracture spacing models. The second scenario is given by a 20-by-40 node grid of two fracture sets with distinct excess compliances and orientations, separated spatially by a linear discontinuity (such as that which may arise from a vertical, planar fault). The effect of the smoothness parameter β_c on the inference result is investigated by performing the inference for different choices of β_c.

LBP is applied in each case to obtain the approximate MAP configuration and posterior marginal distributions of the fracture properties, the latter of which are used to compute the approximate BLS estimate of the fracture properties. LBP converged for all models in less than 200 iterations, when the smoothness parameter β_c was taken to be less than or equal to 0.1. Choices of the smoothness parameter greater than 0.1 resulted in LBP not converging for some realizations of the noisy data.

We must take care to interpret the results correctly, as we have taken $z_{ij} = -13$ to mean that no fractures are present at node (i, j), which would render φ_{ij} meaningless and arbitrary. Thus, we compute the posterior marginals for φ_{ij} conditioned on the event $\{\mathsf{z}_{ij} > -13\}$, and likewise compute the BLS estimates for these random variables conditioned on the same event. The results of the inference procedure on the single fracture system are plotted in Figures 3-5–3-7. The resulting residuals between the estimates and true values are given in Table 3.2 in terms of the root mean squared (RMS) residuals over all nodes.

We observe that the inference procedure performs very well at fracture spacings smaller than 40 m. This is to be expected as the forward model for the AvAz data relies on the assumption that fractures cause the medium to behave as an equivalent anisotropic medium, but this assumption breaks down as the fracture spacing becomes comparable to the dominant seismic wavelength (which is 100 m in the fractured layer and 87.5 m in the layer above the fractures). Furthermore, while the scattering

Fracture Spacing	$\epsilon_{\text{rms},\hat{\boldsymbol{z}}_{\text{BLS}}}$	$\epsilon_{\text{rms},\hat{\boldsymbol{\varphi}}_{\text{BLS}}}$	$\epsilon_{\text{rms},\hat{\boldsymbol{z}}_{\text{MAP}}}$	$\epsilon_{\text{rms},\hat{\boldsymbol{\varphi}}_{\text{MAP}}}$
12 m	0.132 $\log_{10}\text{Pa}^{-1}$	1.27 °	0.123 $\log_{10}\text{Pa}^{-1}$	1.00°
20 m	0.102 $\log_{10}\text{Pa}^{-1}$	2.27 °	0.083 $\log_{10}\text{Pa}^{-1}$	2.00°
40 m	0.488 $\log_{10}\text{Pa}^{-1}$	10.53 °	0.235 $\log_{10}\text{Pa}^{-1}$	18.63°
60 m	0.670 $\log_{10}\text{Pa}^{-1}$	0.05 °	0.530 $\log_{10}\text{Pa}^{-1}$	0°
80 m	0.555 $\log_{10}\text{Pa}^{-1}$	7.98 °	0.756 $\log_{10}\text{Pa}^{-1}$	0°
100 m	0.358 $\log_{10}\text{Pa}^{-1}$	0.12 °	0.187 $\log_{10}\text{Pa}^{-1}$	0°

Table 3.2: Root mean square of residuals between estimates and ground truth (mean taken over all nodes) when estimating a single fracture set with fracture compliance 10^{-9} m/Pa, fracture strike $\varphi_{ij} = 60°$, and varying fracture spacing, and with smoothness parameter set to $\beta_c = 0.1$. For comparison, note that azimuth φ is discretized into 20° bins and log compliance z has been discretized into bins of size 0.1 $\log_{10}\text{Pa}^{-1}$.

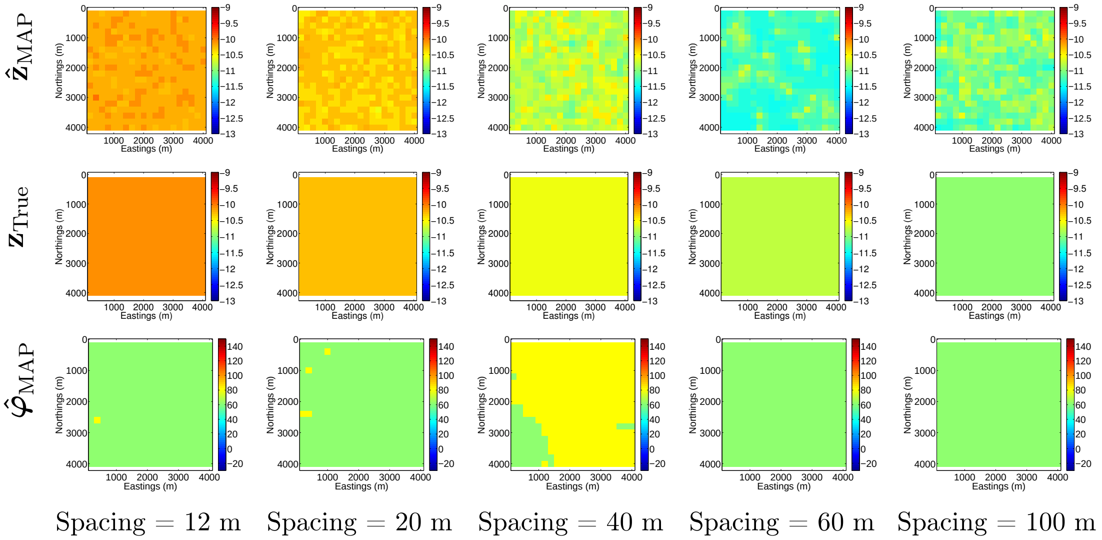


Figure 3-5: Approximate MAP estimates of the fracture properties computed for models of a single fracture set, with fracture compliance 10^{-9} m/Pa, fracture strike $\varphi_{ij} = 60°$, and varying fracture spacing, and with smoothness parameter set to $\beta_c = 0.1$.

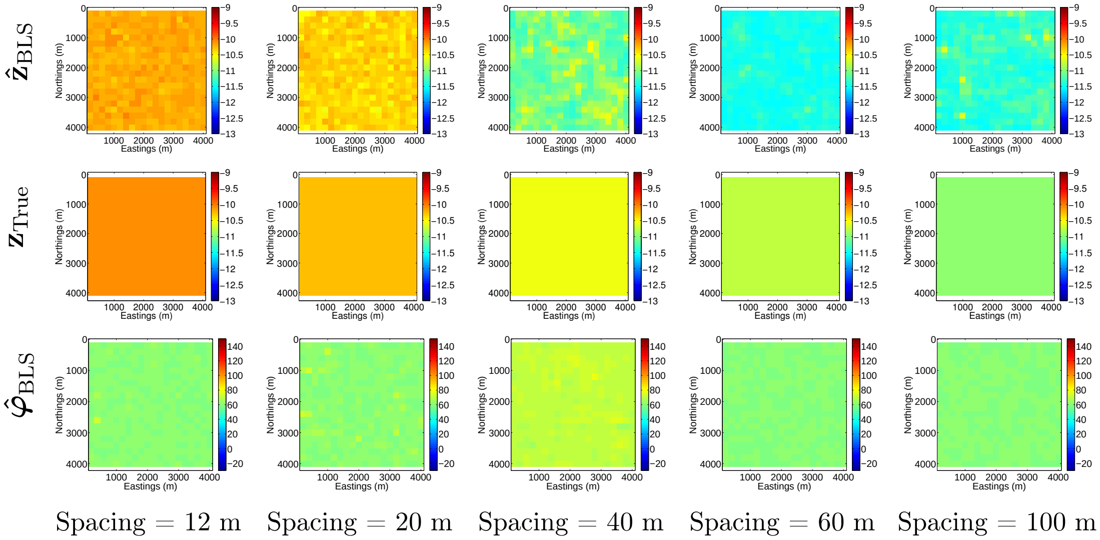
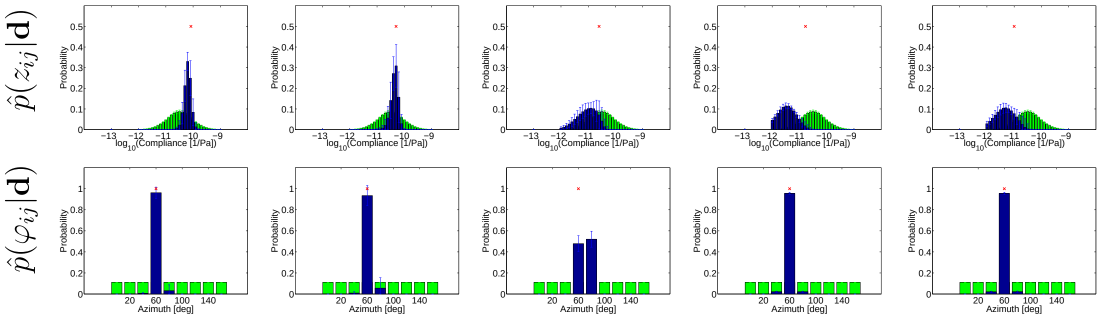


Figure 3-6: Approximate BLS estimates of the fracture properties computed for models of a single fracture set, with fracture compliance 10^{-9} m/Pa, fracture strike $\varphi_{ij} = 60°$, and varying fracture spacing, and with smoothness parameter set to $\beta_c = 0.1$.



Spacing = 12 m Spacing = 20 m Spacing = 40 m Spacing = 60 m Spacing = 100 m

Figure 3-7: Approximate posterior marginal distributions (blue) of the fracture properties at a single node plotted along with the prior distributions (green). Results are given as mean ± 1 S.D. over all grid nodes. True value is plotted with a red 'x'. Computed for models of a single fracture set, with fracture compliance 10^{-9} m/Pa, fracture strike $\varphi_{ij} = 60°$, and varying fracture spacing, and with smoothness parameter set to $\beta_c = 0.1$.

assumptions underlying the forward model for the FTF data are valid for fracture spacings on the order of the seismic wavelength, the FTF data in this study only contributes to fracture detection and strike estimation; the actual excess compliance value (given fractures are present) has no bearing on our model for the FTF data. In particular, we notice from Table 3.2 that the RMS residuals for estimating log excess compliance grows to multiple bin sizes at fracture spacings of 40 m and larger. The marginal distributions in Figure 3-7 convey well what happens at spacings of 40 m and larger. We see that the least best fit to fracture orientation is obtained at 40 m spacing; this is likely because at this mid-range fracture spacing, we see a weaker response from both the AvAz data and the FTF data. However, at larger fracture spacings, the marginal distributions for fracture orientation remain concentrated around the true orientation of 60°, as expected due to both the very simple model for FTF in terms of fracture orientation as well as the assumptions underlying the FTF model remaining strong at larger spacings. We observe that excess compliance tends to be underestimated at fracture spacing values of 40 m and above. This observation is consistent with our intuition for AvAz data; at larger spacings, the AvAz response becomes weaker, and a better fit to the data is found with smaller excess compliances than those resulting from the true fracture compliance and spacing.

In order to investigate the effect of the smoothness parameter β_c on the inference we apply our procedure over a range of choices for β_c on the 80 m spacing model. The effect of the smoothness parameter is most easily seen in the MAP estimates, which are plotted in Figure 3-8. Observing the changes in the estimate with increasing smoothness parameter β_c, we see that the a higher value of β_c has the effect of denoising the estimates. When β_c is 0, this is identical to performing the inference on a fully disconnected graph, as the edge potentials will all be identically equal to 1. As such, the estimate at each node will fit only the noisy data corresponding to itself. Increasing β_c strengthens the links between adjacent nodes, and can cause an incorrect fit to noisy data to be less probable; this is particularly true for fracture azimuth, which is estimated correctly at 80 m fracture spacing. Increasing β_c still smooths the estimates for excess compliance, however towards the underestimated value found at

80 m (again due to the weaker AvAz response at high fracture spacings). At $\beta_c = 0.1$, we see a considerable improvement in the estimates for fracture azimuth.

We now turn to the second scenario arising from two sets of fractures with different azimuths and fracture spacings separated by a discontinuity on the grid such as that which may arise from a geological fault. We investigate the effect of *a priori* knowledge of the fault by performing the inference both with and without knowledge of the fault location encoded in the prior with the smoothness parameter set to $\beta_{ij,kl} = 0$ at the fault and $\beta_{ij,kl} = \beta_c = 0.1$ elsewhere. The results are plotted in Figures 3-9 and 3-10 for fracture azimuth and log excess compliance, respectively. As evidenced by the figures, when the fault is unknown *a priori*, the estimates for fracture properties are smoothed across the fault. This is particularly undesirable for the estimation of fracture azimuth, where we would otherwise be able to obtain good estimates in both regions. Specifying the location of the fault *a priori* sets the smoothness parameter at the corresponding edges to 0, and hence we no longer observe this behavior.

3.6 Conclusions and Future Work

A methodology for estimation of fracture properties from AvAz and FTF data under a Bayesian inference framework has been presented. The inference is performed by running loopy belief propagation on the 2-D Markov random field of fracture variables. LBP converged relatively quickly on the synthetic data, in under 200 iterations for all models, when using a smoothness parameter less than or equal to 0.1. We have demonstrated that the approximate inference results perform well for both fracture azimuth and excess compliance at low spacings of 12 m and 20 m, and continue to give good estimates for fracture azimuth up to 100 m spacing. This is significant, as we are able to estimate the fracture properties in a rigorous manner at a greater range of spacings than would otherwise be attainable.

We further showed that our use of the spatial smoothness prior has the effect of denoising estimates that would otherwise be incorrect. We also demonstrated the capability of this framework to handle prior information about geological features,

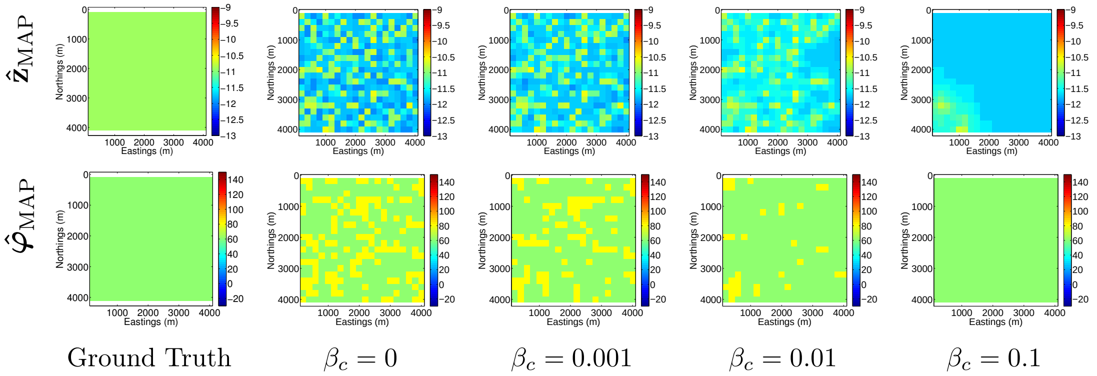
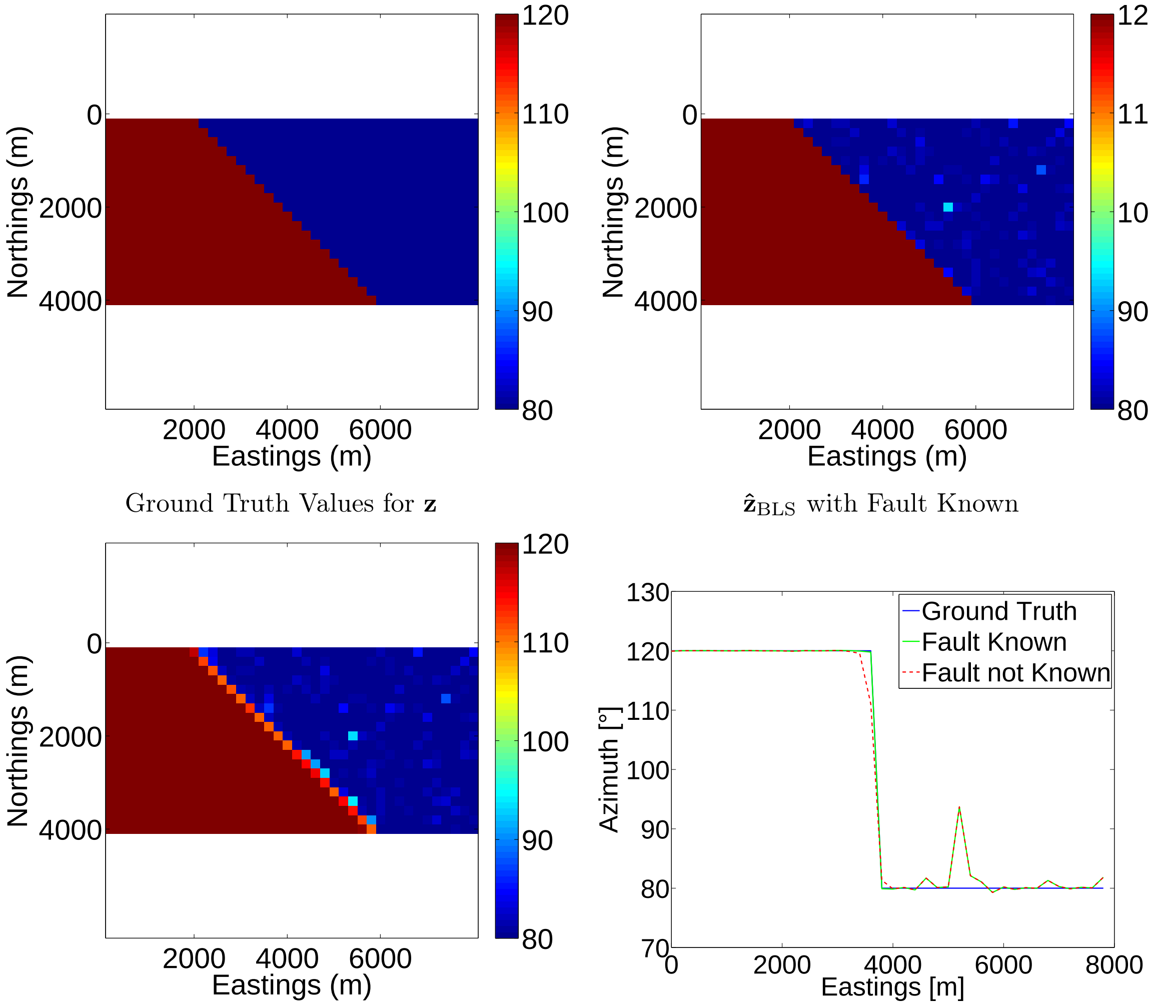


Figure 3-8: Approximate MAP estimates of the fracture properties computed on a model containing a single set of fractures with fracture compliance 10^{-9} m/Pa, fracture spacing 80 m, and fracture orientation 60°. Ground truth is plotted along with the estimates using various values for the smoothness parameter β_c.



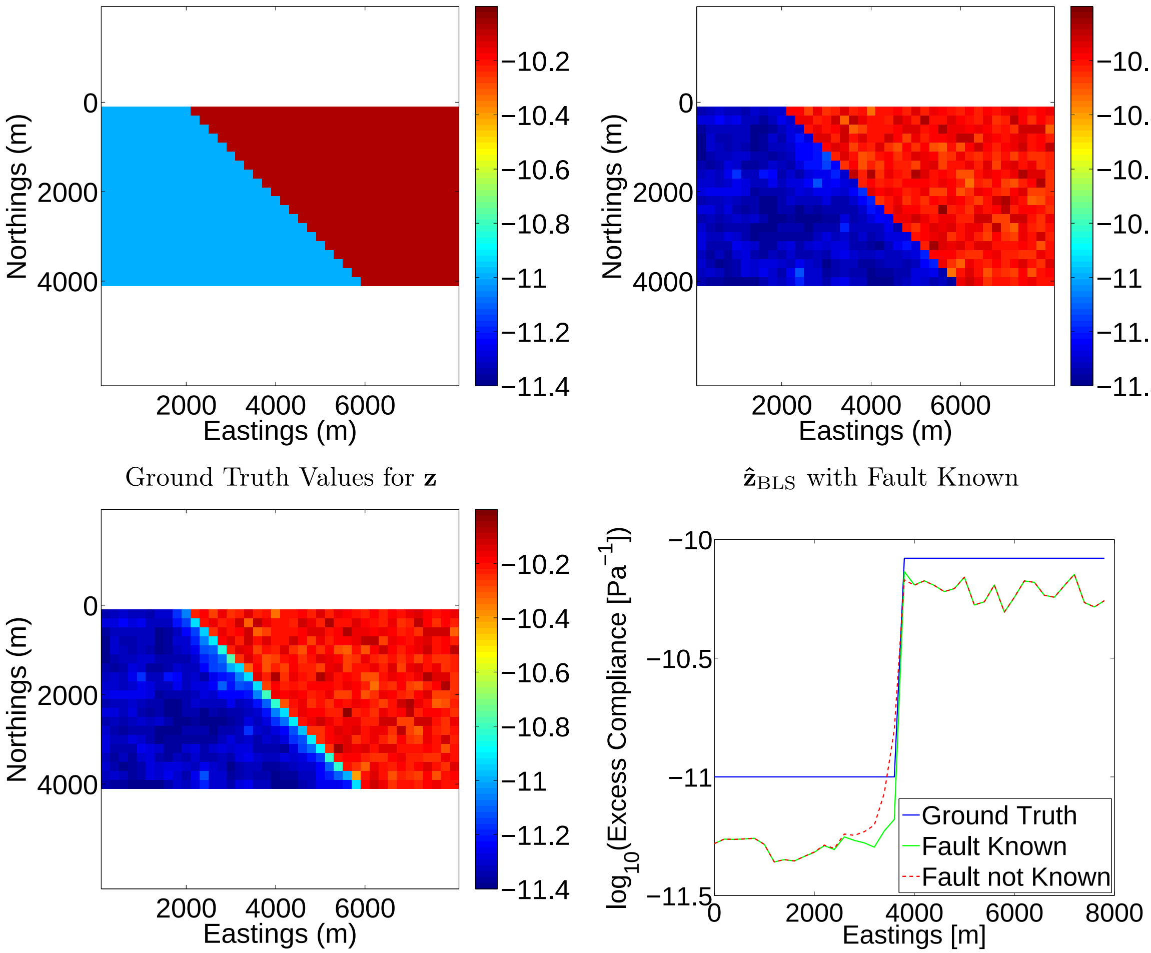
Ground Truth Values for $\mathbf{z}$

$\hat{\mathbf{z}}_{\mathrm{BLS}}$ with Fault Known

$\hat{\mathbf{z}}_{\mathrm{BLS}}$ with Fault not Known

Comparison of Estimates

Figure 3-9: Effect of *a priori* knowledge of the fault on approximate BLS estimates of the fracture azimuth of a model containing two fracture sets. The fractures on the left are at azimuth 120°, spacing 100 m, and fracture compliance 10^{-9} m/Pa. The fractures on the right are at azimuth 80°, spacing 12 m, and fracture compliance 10^{-9} m/Pa. Ground truth is plotted along with the estimates with the smoothness parameter $\beta_c = 0.1$. The fourth pane shows a comparison of the estimates at the horizontal slice North=2000 m.



Ground Truth Values for $\mathbf{z}$

$\hat{\mathbf{z}}_{\mathrm{BLS}}$ with Fault Known

$\hat{\mathbf{z}}_{\mathrm{BLS}}$ with Fault not Known

Comparison of Estimates

Figure 3-10: Effect of *a priori* knowledge of the fault on approximate BLS estimates of the fracture excess compliance of a model containing two fracture sets. The fractures on the left are at azimuth 120°, spacing 100 m, and fracture compliance 10^{-9} m/Pa. The fractures on the right are at azimuth 80°, spacing 12 m, and fracture compliance 10^{-9} m/Pa. Ground truth is plotted along with the estimates with the smoothness parameter $\beta_c = 0.1$. The fourth pane shows a comparison of the estimates at the horizontal slice North=2000 m.

such as the discontinuity shown in the previous figures. While we presented a very simple case of this, it is not difficult to extend this to more complicated scenarios. Having validated our procedure on synthetic data, the next natural step will be to obtain field data and apply the inference procedure to estimate the desired fracture properties.

One future direction to improve the inference is to relate fracture spacing and compliance to the FTF data. While the data we chose depended only on the presence and orientation of fractures, Fang et al. [22] showed both theoretically and in laboratory experiments that FTF also contains information about fracture spacing. However, even when using synthetic data, the precise physical relationship between FTF and fracture spacing has been difficult to determine, but a geophysical basis remains for exploring this avenue further. If a reliable forward model can be determined to relate FTF to fracture spacing, then we will be able to move beyond estimating excess fracture compliance to estimation of individual fracture compliances and fracture spacing. A related future direction is to incorporate additional features of the seismic data in the inference procedure. In particular, Zheng et al. [82] describe a theory for using 3-D beam interference to determine fracture properties of a reservoir from reflected seismic P-wave data.

Chapter 4

Least-Squares Migration with a Hierarchical Bayesian Framework

4.1 Summary

In many geophysical inverse problems, smoothness assumptions on the underlying geology are utilized to mitigate the effects of poor data coverage and noise in the data and to improve the quality of the inferred model parameters. Within a Bayesian inference framework, *a priori* assumptions about the probabilistic structure of the model parameters impose such a smoothness constraint (also known as regularization). We consider the particular problem of inverting seismic data for the subsurface reflectivity of a 2-D medium, where we assume a known velocity field. In particular, we consider a hierarchical Bayesian generalization of the Kirchhoff-based least-squares migration (LSM) method. We present here a novel methodology for estimation of both the reflectivity model and regularization parameters, using a Bayesian statistical framework that treats both of these as random variables to be inferred from the data. Hence rather than fixing the regularization parameters prior to inverting for the image, we allow the data to dictate where to regularize. In order to construct our prior of the subsurface and regularization parameters, we define an undirected graphical model (or Markov random field) on the image, where the vertices of the graph represent subsurface reflectivity values and the regularization parameters are

defined to parameterize the edges of the graph. Estimating these regularization parameters (which we refer to as edge strengths) gives us information about the degree of conditional correlation (or lack thereof) between neighboring image parameters, and subsequently incorporating this information in the final model produces more clearly visible discontinuities in the estimated image. The inference framework is verified on a 2-D synthetic dataset, where the hierarchical Bayesian imaging results significantly outperform standard LSM images. We note that while this method is presented within the context of seismic imaging, it is in fact a general methodology which can be applied to any linear inverse problem in which there are spatially-varying correlations in the model parameter space.

4.2 Introduction

Seismic imaging (also known as migration) refers to the process of creating an image of the Earth's subsurface reflectivity from seismograms generated by sources and recorded by receivers located, typically, at or near the surface. Traditional migration methods for constructing the image generally involve operating on the seismic data with the adjoint of an assumed forward modeling operator [11], possibly along with a modifying function which attempts to correct for amplitude loss due to geometric spreading, transmission, absorption, etc. [5, 29]. In recent years, attempts have been made to cast the imaging problem as a least-squares inverse problem [19, 49]. This approach to imaging is conventionally referred to as *least-squares migration* (LSM). Early treatments of this approach can be found in LeBras and Clayton [39] and Lambare et al. [37]. This chapter will deal mainly with Kirchhoff-based LSM , which uses a ray-theoretic based forward modeling operator; its derivation and application is discussed in Nemeth et al. [49] and Duquet et al. [19]. LSM can also be applied with wave-equation-based forward modeling, as shown by Kühl and Sacchi [36].

In solving the least-squares inverse problem, it is common to include some form of regularization in the LSM cost function in order to penalize less smooth images. For example, Clapp [12] describes two regularization schemes for LSM in which the image is constrained to be smooth either along geological features predetermined by a seismic interpreter or along the ray-parameter axis. In these and other applications of LSM, the regularization is chosen independently of the seismic data, i.e. it is a fixed input to the inversion procedure (as it is in the vast majority of geophysical applications of inversion). This, however, may result in sub-optimal inversion results; overly strong regularization may result in over-smoothing the image, whereas weak regularization may not adequately penalize roughness in the image due to noise. Even if an appropriate regularization strength is determined, the true smoothness structure of the model need not be spatially uniform or even isotropic; for example, the true earth may typically contain many sharp discontinuities where any form of smoothing would be undesirable.

In this chapter, we propose a more general approach to LSM which solves for parameters defining the image regularization in conjunction with the optimal image itself. The approach is formulated within the framework of Bayesian inference, in which regularization is accomplished with a prior probability distribution on the image parameters. We define a spatially-varying smoothness prior and seek to jointly estimate its parameters along with the image. In particular, we utilize a variant of Bayesian inference known as hierarchical Bayes, which provides a rigorous mathematical framework for addressing the joint estimation of the image and regularization parameters. This should allow for preserving sharpness in the image at the true discontinuities while still smoothing the effects of noise.

Previous applications of hierarchical Bayesian inference in geophysics include Malinverno and Briggs [45], who applied it to 1-D traveltime tomography, Malinverno [43, 44], who applied Bayesian model selection to find optimal parameterizations of 1-D density and resistivity models, Buland and Omre [9], who applied hierarchical Bayesian methods in amplitude versus offset (AVO) inversion, and Bodin et al. [7], who applied Bayesian model selection to determine group velocities for the Australian continent.

In the next sections, we review Kirchhoff-based LSM and proceed to develop the hierarchical Bayesian framework and algorithms used to solve the inference problem.

4.3 Methodology

4.3.1 Standard Kirchhoff-based LSM framework

Kirchhoff Modeling

The Kirchhoff modeling operator is a ray-based forward modeling operator that gives the seismic data as a linear function of the reflectivity model. In particular, to simulate the seismogram $d_{sr}(t)$ recorded at a seismic receiver r from a seismic source s, Kirchhoff forward modeling first generates a source-to-reflector-to-receiver travel time (or two-way travel time) field $\tau_{sr}(\mathbf{x})$ by utilizing what is known as the "exploding reflector" concept. This concept refers to the treatment of each point in the reflectivity model as a point source. The two-way travel time can be computed as the sum of the source-to-reflector and reflector-to-receiver travel times, as determined by ray-tracing through a specified background velocity model of the subsurface. The ray tracer also computes the field of ray-path lengths $R_s(\mathbf{x})$ and $R_r(\mathbf{x})$ and opening angles between the source and receiver rays at each reflection point $\theta_{sr}(\mathbf{x})$. Once these quantities have been computed, the synthetic data $\hat{d}_{sr}(t)$ are computed by superposition over reflector locations $\mathbf{x}$ of scaled and shifted versions of the source wavelet $w_s(t)$ (after applying a 90-degree phase-shift to simulate the effects of 2-D propagation). For each $\mathbf{x}$, the phase-shifted wavelet $\tilde{w}_s(t)$ is delayed by $\tau_{sr}(\mathbf{x})$ and scaled by the reflectivity value $m(\mathbf{x})$, an obliquity correction factor $\cos(\theta_{sr}(\mathbf{x})/2)$, and a geometric spreading correction (in 2-D, $1/\sqrt{R_s(\mathbf{x})R_r(\mathbf{x})}$). Thus,

$$\hat{d}_{sr}(t) = \int_{\mathcal{X}} m(\mathbf{x}) \frac{\tilde{w}_s\left(t - \tau_{sr}\left(\mathbf{x}\right)\right) \cos(\theta_{sr}(\mathbf{x})/2)}{\sqrt{R_s(\mathbf{x})R_r(\mathbf{x})}} \, \mathrm{d}\mathbf{x}, \tag{4.1}$$

where $\mathcal{X} \subset \mathbb{R}^2$ is the model domain. We note that the above Kirchhoff modeling operator is precisely the adjoint operator to the Kirchhoff migration operator, given by:

$$\hat{m}(\mathbf{x}) = \sum_s \sum_r \int_t d_{sr}(t) \frac{\tilde{w}_s\left(t - \tau_{sr}\left(\mathbf{x}\right)\right) \cos(\theta_{sr}(\mathbf{x})/2)}{\sqrt{R_s(\mathbf{x})R_r(\mathbf{x})}} \, \mathrm{d}t. \tag{4.2}$$

If we discretize time and space, we can represent our data and image as finite-dimensional vectors $\mathbf{d}$ and $\mathbf{m}$, where the dimension of $\mathbf{d}$ is the number of source-receiver pairs times the number of time samples, and where the dimension of $\mathbf{m}$ is the number of points in a spatial grid sampling the model domain. Then, replacing the integral in (4.1) with a summation, we can express the Kirchhoff modeling operator in matrix form:

$$\hat{\mathbf{d}} = A\mathbf{m}. \tag{4.3}$$

In particular, the ith column of A, corresponding to a point x_i in the model grid, will contain a sampled version of the source wavelet for each source-receiver pair, appropriately scaled or shifted, giving (in 2-D):

$$A_{srt,i} = \frac{\tilde{w}_s\left(t - \tau_{sr}\left(x_i\right)\right)\cos(\theta_{sr}(x_i)/2)}{\sqrt{R_s(x_i)R_r(x_i)}}\,\ell^2, \tag{4.4}$$

where ℓ is the spatial discretization interval.

Standard LSM Framework

Least-squares migration attempts to solve the imaging problem by seeking the image $\mathbf{m}_{\text{LS}}$ that minimizes the ℓ^2-norm of the residual (the difference between the observed data $\mathbf{d}$ and the modeled data $\hat{\mathbf{d}} = A\mathbf{m}$). Without regularization, the LSM image is given by

$$\mathbf{m}_{\text{LS}} = \arg\min_{\mathbf{m}} \|\mathbf{d} - A\mathbf{m}\|_2^2, \tag{4.5}$$

where $\|\cdot\|_2$ denotes the ℓ^2-norm in the (discretized) data-space given by

$$\|\mathbf{d}\|_2^2 = \sum_s \sum_r \sum_t d_{sr}(t)^2. \tag{4.6}$$

To ensure well-posedness of the LSM solution, regularization is often introduced by augmenting the LSM cost function with a term which penalizes differences between model parameters and an additional term which penalizes the magnitude of the image.

This gives the regularized LSM image as

$$\mathbf{m}_{\mathrm{RLS}} = \arg\min_{\mathbf{m}} \|\mathbf{d} - A\mathbf{m}\|_2^2 + \lambda \left(\sum_{(i,j)\in\mathcal{E}} \beta_{ij}(m_i - m_j)^2 + \epsilon \sum_i m_i^2 \right) \tag{4.7}$$

$$= \arg\min_{\mathbf{m}} \|\mathbf{d} - A\mathbf{m}\|_2^2 + \lambda \mathbf{m}^T \left(D\left(\boldsymbol{\beta}\right) + \epsilon I\right) \mathbf{m}, \tag{4.8}$$

where $\beta_{ij} \in [0, 1]$ indicates how strongly to penalize the difference between m_i and m_j, $\mathcal{E}$ is the set of all pairs of image parameter indices whose difference we decide to potentially penalize, $\lambda > 0$ assigns the maximal weight given to penalizing these differences, and $\epsilon > 0$ weights the penalty on parameter magnitudes. Equation (4.8) is simply (4.7) rewritten in compact matrix-vector notation, where D is a differencing operator defined by the vector $\boldsymbol{\beta} = \{\beta_{ij} : (i, j) \in \mathcal{E}\}$. Taking the derivative of the right-hand side of (4.8) and setting it to zero yields the solution to the regularized LSM problem:

$$\mathbf{m}_{\mathrm{RLS}} = \left(A^T A + \lambda(D\left(\boldsymbol{\beta}\right) + \epsilon I)\right)^{-1} A^T \mathbf{d}. \tag{4.9}$$

Note that $\epsilon > 0$ ensures that the regularized LSM cost function is a positive-definite quadratic function of the image $\mathbf{m}$, and hence its minimizer is unique.

4.3.2 Bayesian Framework

Standard Bayesian Formulation

The same solution to LSM can be derived from a Bayesian formulation of the imaging problem, wherein the image $\mathbf{m}$ and the data $\mathbf{d}$ are taken to be random vectors. In particular, we take m *a priori* to be Gaussian with zero mean and some covariance matrix C (i.e. $\mathsf{m} \sim \mathcal{N}(0, C)$) , so that the prior distribution $p(\mathbf{m})$ for m is given by

$$p(\mathbf{m}) \propto \exp\left\{ -\frac{1}{2}\mathbf{m}^T C^{-1} \mathbf{m} \right\}. \tag{4.10}$$

We model the seismic data as $\mathsf{d} = A\mathsf{m} + \mathsf{n}$ where A is our Kirchhoff modeling operator and n is zero-mean Gaussian noise with some covariance matrix Σ (i.e. $\mathsf{n} \sim \mathcal{N}(0, \Sigma)$).

Thus the conditional distribution for the data $\mathbf{d}$ given the model $\mathbf{m}$ will be

$$p(\mathbf{d}|\mathbf{m}) \propto \exp\left\{-\frac{1}{2}\left(\mathbf{d} - A\mathbf{m}\right)^T \Sigma^{-1}\left(\mathbf{d} - A\mathbf{m}\right)\right\}, \tag{4.11}$$

i.e. $\mathbf{d}|\mathbf{m} \sim \mathcal{N}(A\mathbf{m}, \Sigma)$.

Applying Bayes' rule gives the posterior distribution for the model $\mathbf{m}$ conditioned on the data $\mathbf{d}$ as

$$p(\mathbf{m}|\mathbf{d}) = \frac{p(\mathbf{m})p(\mathbf{d}|\mathbf{m})}{p(\mathbf{d})} \tag{4.12}$$

$$\propto \frac{1}{p(\mathbf{d})} \exp\left\{-\frac{1}{2}\left[\mathbf{m}^T C^{-1}\mathbf{m} + \left(\mathbf{d} - A\mathbf{m}\right)^T \Sigma^{-1}\left(\mathbf{d} - A\mathbf{m}\right)\right]\right\}. \tag{4.13}$$

Rearranging terms in (4.13) and dropping any multiplicative factors that do not depend on $\mathbf{m}$, we obtain

$$p(\mathbf{m}|\mathbf{d}) \propto \exp\left\{-\frac{1}{2}\left(\mathbf{m} - \boldsymbol{\mu}_{\text{post}}\right)\Lambda_{\text{post}}^{-1}\left(\mathbf{m} - \boldsymbol{\mu}_{\text{post}}\right)^T\right\}. \tag{4.14}$$

where $\boldsymbol{\mu}_{\text{post}}$ is the posterior mean given by

$$\boldsymbol{\mu}_{\text{post}} = \left(A^T\Sigma^{-1}A + C^{-1}\right)^{-1} A^T\Sigma^{-1}\mathbf{d} \tag{4.15}$$

and Λ_{post} is the posterior covariance matrix given by

$$\Lambda_{\text{post}} = \left(A^T\Sigma^{-1}A + C^{-1}\right)^{-1}. \tag{4.16}$$

That is, the posterior distribution for $\mathbf{m}$ conditioned on $\mathbf{d}$ is itself Gaussian: $\mathbf{m}|\mathbf{d} \sim \mathcal{N}(\boldsymbol{\mu}_{\text{post}}, \Lambda_{\text{post}})$.

The Bayesian *maximum a posteriori* (MAP) estimate $\mathbf{m}_{\text{MAP}}$ is the image that maximizes the posterior distribution (4.13). It is clear from (4.14) that $\mathbf{m}_{\text{MAP}} = \boldsymbol{\mu}_{\text{post}}$. Comparing to Equation (4.9), we also see that $\mathbf{m}_{\text{MAP}} = \mathbf{m}_{\text{RLS}}$ when we set the prior and noise covariance matrices as

$$C = \left(\lambda(D\left(\boldsymbol{\beta}\right) + \epsilon I)\right)^{-1} \tag{4.17}$$

and

$$\Sigma = I. \tag{4.18}$$

Constructing the Prior via a Graphical Model

The choice of $\boldsymbol{\beta}$ plays a key role in determining the spatial smoothness properties of the prior on the model. This is perhaps best seen through the expressive formalism of probabilistic graphical models. In particular, we define an undirected graphical model $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, with a set of vertices (or nodes) $\mathcal{V}$, which index the random variables m_i comprising the random vector $\mathbf{m}$, and a set of edges $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$, represented as pairs of vertices in $\mathcal{V}$ that encode dependencies between the random variables. Recall from Chapter 2 that for a Gaussian random vector $\mathbf{m}$ with precision matrix (or inverse covariance matrix) $Q = C^{-1}$, it can be shown that $\mathbf{m}$ forms an MRF over $\mathcal{G}$ if and only if Q is at least as sparse as the edge set $\mathcal{E}$ (meaning that no edge between nodes i and j in $\mathcal{G}$ implies that $Q_{ij} = Q_{ji} = 0$) [35].

Defining the prior precision matrix as $Q = \lambda(D(\boldsymbol{\beta}) + \epsilon I)$, as in (4.17), allows the β_{ij} to determine the "strength" of each edge in $\mathcal{G}$. In probabilistic terms, $\boldsymbol{\beta}$ captures the prior conditional dependence structure of the image $\mathbf{m}$, such that $\beta_{ij} = 0$ implies that, prior to observing $\mathbf{d}$, m_i is conditionally independent of m_j when $\{\mathsf{m}_k : k \neq i, j\}$ is given. For this reason, we sometimes refer to the elements of $\boldsymbol{\beta}$ as the *edge strengths* of $\mathcal{G}$ and to $D(\boldsymbol{\beta})$ as the weighted graph Laplacian of $\mathcal{G}$ (weighted by $\boldsymbol{\beta}$). This is depicted in Figure 4-1 for a simple nine pixel image. Note that although Figure 4-1 shows edges connecting only nearest neighbors horizontally and vertically, this need not be the case. We can consider a situation where each node shares an edge with all other nodes within a specified radius; the graphical model depicted in the figure results from using a radius of 1 node.

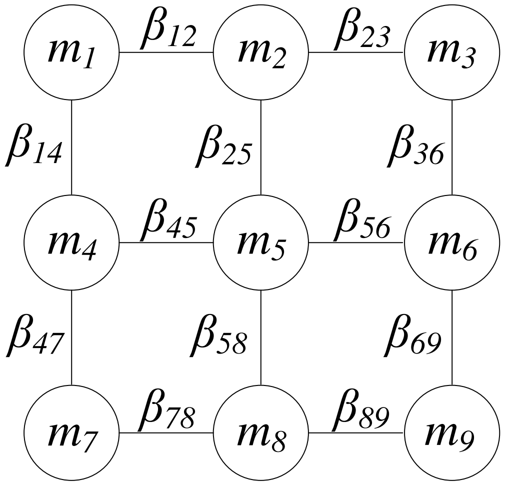


Figure 4-1: The Markov random field imposed on $\mathbf{m}$ by fixing $\boldsymbol{\beta}$ prior to observing the data $\mathbf{d}$, for a simple nine pixel image.

Hierarchical Bayesian Formulation

Thus far we have assumed that the parameters λ, ϵ, and $\boldsymbol{\beta}$, which determine the regularization in the LSM framework and the prior model covariance structure in the Bayesian framework, are known. We now describe how we can expand the Bayesian formulation to the problem of estimating these regularization parameters from the data $\mathbf{d}$, in addition to the image $\mathbf{m}$. We focus on the estimation of the edge strengths $\boldsymbol{\beta}$, which capture our belief about where we think the image should be smooth. This is a reasonable approach since the edge strengths $\boldsymbol{\beta}$ give us prior information about our model $\mathbf{m}$, and $\mathbf{m}$ gives us information about our data $\mathbf{d}$, hence we should be able to infer something about $\boldsymbol{\beta}$ from $\mathbf{d}$. This is depicted in the directed graphical model of Figure 4-2, which also illustrates the induced Markov chain structure between $\boldsymbol{\beta}$, $\mathbf{m}$, and $\mathbf{d}$.

In order to estimate $\boldsymbol{\beta}$ from $\mathbf{d}$, we consider $\boldsymbol{\beta}$ to be a random vector endowed with its own prior $p(\boldsymbol{\beta})$. Accordingly, all probability distributions in the previous sections can be considered as conditional on $\boldsymbol{\beta}$. In particular, we now write the prior on $\mathbf{m}|\boldsymbol{\beta}$ as

$$p\left(\mathbf{m}\,|\,\boldsymbol{\beta}\right) = \frac{\left|\lambda\left(D\left(\boldsymbol{\beta}\right)+\epsilon I\right)\right|^{1/2}\exp\left\{-\frac{1}{2}\mathbf{m}^T(\lambda\left(D\left(\boldsymbol{\beta}\right)+\epsilon I\right))\mathbf{m}\right\}}{\left(2\pi\right)^{N/2}} \tag{4.19}$$

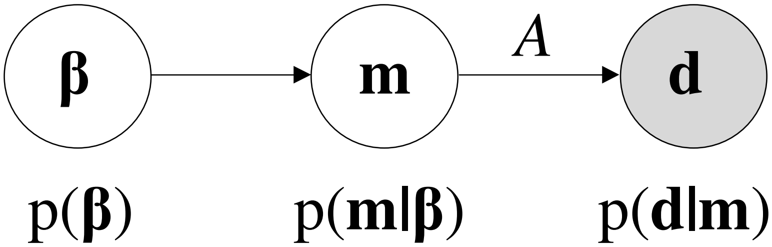


Figure 4-2: The directed graphical model capturing the Markov chain structure between $\boldsymbol{\beta}$, $\mathbf{m}$, and $\mathbf{d}$. The node for $\mathbf{d}$ is shaded to indicate that $\mathbf{d}$ is an observed quantity that the posterior distributions of $\boldsymbol{\beta}$ and $\mathbf{m}$ are conditioned upon.

and the conditional distribution for $\mathbf{d}|\mathbf{m},\boldsymbol{\beta}$ as

$$p\left(\mathbf{d}\,|\,\mathbf{m},\boldsymbol{\beta}\right) = p\left(\mathbf{d}\,|\,\mathbf{m}\right) \tag{4.20}$$

$$= \frac{\exp\left\{-\frac{1}{2}\left(\mathbf{d}-A\mathbf{m}\right)^T \Sigma^{-1}\left(\mathbf{d}-A\mathbf{m}\right)\right\}}{\left(2\pi\right)^{K/2}\Sigma^{1/2}}, \tag{4.21}$$

where N is the number of model parameters (i.e. the dimension of $\mathbf{m}$) and K is the number of data points (the dimension of $\mathbf{d}$). We again apply Bayes' rule to obtain the joint posterior distribution for $\mathbf{m}$ *and* $\boldsymbol{\beta}$ given the data $\mathbf{d}$:

$$p\left(\mathbf{m},\boldsymbol{\beta}\,|\,\mathbf{d}\right) = \frac{p\left(\boldsymbol{\beta}\right)p\left(\mathbf{m}\,|\,\boldsymbol{\beta}\right)p\left(\mathbf{d}\,|\,\mathbf{m},\boldsymbol{\beta}\right)}{p\left(\mathbf{d}\right)} \tag{4.22}$$

$$= \frac{p\left(\boldsymbol{\beta}\right)}{p\left(\mathbf{d}\right)}\frac{\left|\lambda\left(D\left(\boldsymbol{\beta}\right)+\epsilon I\right)\right|^{1/2}}{\left(2\pi\right)^{(N+K)/2}\Sigma^{1/2}}$$
$$\exp\left\{-\frac{1}{2}\Big(\left(\mathbf{d}-A\mathbf{m}\right)^T\Sigma^{-1}\left(\mathbf{d}-A\mathbf{m}\right)+\mathbf{m}^T(\lambda\left(D\left(\boldsymbol{\beta}\right)+\epsilon I\right))\mathbf{m}\Big)\right\}. \tag{4.23}$$

To define $p(\boldsymbol{\beta})$, we endow each β_{ij} with a uniform prior on the set $[0,1]$ and let the β_{ij} be mutually independent random variables, so that

$$p(\boldsymbol{\beta}) = \prod_{(i,j)\in\mathcal{E}} \mathbb{1}_{[0,1]}\left(\beta_{ij}\right), \tag{4.24}$$

where

$$\mathbb{1}_S(x) = \begin{cases} 1 & \text{if } x \in S \\ 0 & \text{if } x \notin S \end{cases} .$$

We note that (4.23) is very similar to the posterior distribution in the non-hierarchical Bayesian setting (where $\boldsymbol{\beta}$ is fixed) with some important differences: firstly, (4.23) is now a function of both $\mathbf{m}$ *and* $\boldsymbol{\beta}$, and secondly, outside the exponential of (4.23) is the determinant of $\mathbf{m}$'s prior precision matrix Q (which can no longer be dropped as a proportionality constant, since it depends on $\boldsymbol{\beta}$). Computing this determinant is expensive, with time complexity $\mathcal{O}\left(N^2\right)$ (since Q is a sparse matrix with bandwidth $N^{1/2}$), and reflects the additional computational cost of the hierarchical Bayesian approach.

Having obtained the joint posterior distribution $p\left(\mathbf{m}, \boldsymbol{\beta} \,|\, \mathbf{d}\right)$, the task of estimating the best image remains. Here, we explore two estimation methodologies within the hierarchical Bayesian framework: the *hierarchical Bayes* solution and the *empirical Bayes* solution [45]. What is strictly known as the hierarchical Bayes solution is the full marginal posterior distribution of the image $p\left(\mathbf{m} \,|\, \mathbf{d}\right)$ (marginalizing out $\boldsymbol{\beta}$ from the joint posterior distribution $p\left(\mathbf{m}, \boldsymbol{\beta} \,|\, \mathbf{d}\right)$). Hence, we have for the hierarchical Bayes solution

$$p\left(\mathbf{m} \,|\, \mathbf{d}\right) = \int_{\mathcal{B}} p\left(\mathbf{m}, \boldsymbol{\beta} \,|\, \mathbf{d}\right) \mathrm{d}\boldsymbol{\beta} \tag{4.25}$$

where $\mathcal{B}$ is the domain of admissible vectors $\boldsymbol{\beta}$. Unfortunately, the marginalization operation cannot be performed analytically and must be computed numerically. We may also consider the MAP estimates for the image that can be derived within the hierarchical Bayesian setting. The hierarchical Bayes MAP estimate $\mathbf{m}_{\text{HB}}$ is the MAP estimate of $\mathbf{m}$ based on its marginal posterior distribution $p\left(\mathbf{m} \,|\, \mathbf{d}\right)$:

$$\mathbf{m}_{\text{HB}} = \arg\max_{\mathbf{m}} \int_{\mathcal{B}} p\left(\mathbf{m}, \boldsymbol{\beta} \,|\, \mathbf{d}\right) \mathrm{d}\boldsymbol{\beta} \tag{4.26}$$

One can think of $\mathbf{m}_{\text{HB}}$ as the single *best* image $\mathbf{m}$ over *all* choices of edge strengths $\boldsymbol{\beta}$. While the posterior marginal distribution for the image (4.25) is the complete solution to the Bayesian inference problem, a number of computational issues prevent its use in practice. Firstly, due to both the high-dimensionality of $\mathcal{B}$ and the cost of evaluating the joint posterior distribution (4.23), both stochastic sampling from and direct marginalization of the joint posterior distribution are computationally intractable. Furthermore, even if we were able to evaluate the marginal posterior (4.25), the high-dimension of $\mathbf{m}$ would make it difficult to explore.

A somewhat different solution for estimating the image is known as the *empirical Bayes* solution, which first looks for the *best* choice for $\boldsymbol{\beta}$, then, using that choice, finds the best image $\mathbf{m}_{\mathrm{EB}}$. If one takes the MAP estimate for $\boldsymbol{\beta}$ then we would have

$$\boldsymbol{\beta}_{\mathrm{MAP}} = \arg\max_{\boldsymbol{\beta}} \int_{\mathcal{M}} p\left(\mathbf{m}, \boldsymbol{\beta} | \mathbf{d}\right) \mathrm{d}\mathbf{m} \tag{4.27}$$

where, it turns out, the marginalization over $\mathbf{m}$ can be performed analytically but the maximization over $\boldsymbol{\beta}$ must still be performed numerically. Given $\boldsymbol{\beta}_{\mathrm{MAP}}$, the empirical Bayes solution is taken as the MAP estimate with respect to $p(\mathbf{m} \,|\, \mathbf{d}, \boldsymbol{\beta}_{\mathrm{MAP}})$. The results of the previous sections then imply

$$\mathbf{m}_{\mathrm{EB}} = \left(A^T \Sigma^{-1} A + \lambda \left(D\left(\boldsymbol{\beta}_{\mathrm{MAP}}\right) + \epsilon I\right)\right)^{-1} A^T \Sigma^{-1} \mathbf{d}. \tag{4.28}$$

The empirical Bayes solution is within reach as long as we are able to compute $\boldsymbol{\beta}_{\mathrm{MAP}}$ by solving the marginal MAP problem of (4.27). In order to do so, we turn to the expectation-maximization (E-M) algorithm, which has direct application in solving such marginal MAP problems.

4.3.3 The Expectation-Maximization (E-M) Algorithm

The E-M algorithm [17, 46] is a powerful and versatile algorithm for solving maximum likelihood and MAP parameter estimation problems when a subset of the variables relevant to the parameter estimation is unobserved (referred to as latent variables). In the context of the seismic imaging problem we consider here, we view the image $\mathbf{m}$ as the latent variables. In the empirical Bayes approach, these variables must be marginalized from the joint posterior distribution on $\mathbf{m}$ and $\boldsymbol{\beta}$ when attempting to estimate the edge strengths $\boldsymbol{\beta}$. For our purposes, E-M can be thought of as a coordinate ascent algorithm for solving the marginal MAP optimization problem (4.27), whereby subsequent estimations are performed between the latent variables ($\mathbf{m}$) and the parameters to be estimated ($\boldsymbol{\beta}$).

In what follows of this section, we give a derivation of the E-M algorithm; similar derivations and a more thorough treatment of E-M can be found in Bishop [4] or McLachlan and Krishnan [46]. To derive the E-M algorithm, we note that maximizing a probability distribution is equivalent to maximizing its logarithm, and define our objective function as the log marginal posterior

$$\ell(\boldsymbol{\beta}) = \log p(\boldsymbol{\beta} \,|\, \mathbf{d}). \tag{4.29}$$

Rearranging terms in the joint posterior distribution, we can rewrite the MAP objective function as:

$$\begin{aligned}
\ell(\boldsymbol{\beta}) &= \log \int_{\mathcal{M}} p(\mathbf{m}, \boldsymbol{\beta} \,|\, \mathbf{d}) \, \mathrm{d}\mathbf{m} && (4.30)\\
&= \log \int_{\mathcal{M}} \frac{p(\mathbf{m}, \boldsymbol{\beta}, \mathbf{d})}{p(\mathbf{d})} \, \mathrm{d}\mathbf{m} && (4.31)\\
&= \log \int_{\mathcal{M}} \frac{p(\boldsymbol{\beta}) p(\mathbf{m}, \mathbf{d} \,|\, \boldsymbol{\beta})}{p(\mathbf{d})} \, \mathrm{d}\mathbf{m} && (4.32)\\
&= \log \int_{\mathcal{M}} p(\mathbf{m}, \mathbf{d} \,|\, \boldsymbol{\beta}) \, \mathrm{d}\mathbf{m} + \log p(\boldsymbol{\beta}) - \log p(\mathbf{d}). && (4.33)
\end{aligned}$$

Here we introduce a proxy distribution on the image, $q(\mathbf{m} \,|\, \mathbf{d})$, where we can choose q to be any probability distribution we like as long as it has the same support as $p(\mathbf{m})$ and where we have made explicit that q can depend on the data $\mathbf{d}$. Dividing and multiplying by q, we have:

$$\ell(\boldsymbol{\beta}) = \log \int_{\mathcal{M}} \frac{q(\mathbf{m} \,|\, \mathbf{d})}{q(\mathbf{m} \,|\, \mathbf{d})} p(\mathbf{m}, \mathbf{d} \,|\, \boldsymbol{\beta}) \, \mathrm{d}\mathbf{m} + \log p(\boldsymbol{\beta}) - \log p(\mathbf{d}) \tag{4.34}$$

$$= \log \mathbb{E}_{q(\mathbf{m}|\mathbf{d})} \left[\frac{p(\mathbf{m}, \mathbf{d} \,|\, \boldsymbol{\beta})}{q(\mathbf{m} \,|\, \mathbf{d})} \right] + \log p(\boldsymbol{\beta}) - \log p(\mathbf{d}), \tag{4.35}$$

where the integral in (4.34) has been recognized as the expected value with respect to q (denoted by $\mathbb{E}_q$) to arrive at (4.35). Now, by Jensen's inequality [46] and the concavity of the log function, we have

$$\ell(\boldsymbol{\beta}) \geq \mathbb{E}_{q(\mathbf{m}|\mathbf{d})}\left[\log\left(\frac{p(\mathbf{m},\mathbf{d}\,|\,\boldsymbol{\beta})}{q(\mathbf{m}\,|\,\mathbf{d})}\right)\right] + \log p(\boldsymbol{\beta}) - \log p(\mathbf{d}) \tag{4.36}$$

$$= \hat{\ell}(q,\boldsymbol{\beta}). \tag{4.37}$$

We see that the function $\hat{\ell}(q,\boldsymbol{\beta})$ is a lower bound on the original objective function $\ell(\boldsymbol{\beta})$. The E-M algorithm maximizes this lower-bound according to the following coordinate ascent scheme, starting with an initial guess $\hat{\boldsymbol{\beta}}^{(0)}$ and iterated for $t = 0, 1, 2, \dots :$

$$\textbf{E-Step:}\;\; \hat{q}^{(t+1)} = \arg\max_{q} \hat{\ell}(q, \hat{\boldsymbol{\beta}}^{(t)}) \tag{4.38}$$

$$\textbf{M-Step:}\;\; \hat{\boldsymbol{\beta}}^{(t+1)} = \arg\max_{\boldsymbol{\beta}} \hat{\ell}(\hat{q}^{(t+1)}, \boldsymbol{\beta}). \tag{4.39}$$

It turns out the E-step can be solved analytically. Let us propose a candidate solution $\tilde{q}$ as the Bayesian posterior of **m** conditioned on **d** and the last iterate $\hat{\boldsymbol{\beta}}^{(t)}$ of $\boldsymbol{\beta}$:

$$\tilde{q}(\mathbf{m}\,|\,\mathbf{d}) = p\left(\mathbf{m}\,|\,\mathbf{d}, \hat{\boldsymbol{\beta}}^{(t)}\right). \tag{4.40}$$

Then if we plug the $\tilde{q}$ into the E-step objective function (4.38), we have:

$$\hat{\ell}(\tilde{q}, \hat{\boldsymbol{\beta}}^{(t)}) = \mathbb{E}_{p\left(\mathbf{m}\,|\,\mathbf{d},\hat{\boldsymbol{\beta}}^{(t)}\right)}\left[\log\left(\frac{p\left(\mathbf{m},\mathbf{d}\,|\,\hat{\boldsymbol{\beta}}^{(t)}\right)}{p\left(\mathbf{m}\,|\,\mathbf{d},\hat{\boldsymbol{\beta}}^{(t)}\right)}\right)\right] + \log p\left(\hat{\boldsymbol{\beta}}^{(t)}\right) - \log p(\mathbf{d}). \tag{4.41}$$

Recognizing the quotient in (4.41) as $p(\mathbf{d}\,|\,\hat{\boldsymbol{\beta}}^{(t)})$, and since the expectation of $p(\mathbf{d}\,|\,\hat{\boldsymbol{\beta}}^{(t)})$

is just itself, we have

$$\begin{aligned}
\hat{\ell}\left(\tilde{q}, \hat{\boldsymbol{\beta}}^{(t)}\right) &= \log p\left(\mathbf{d} \,|\, \hat{\boldsymbol{\beta}}^{(t)}\right) + \log p\left(\hat{\boldsymbol{\beta}}^{(t)}\right) - \log p(\mathbf{d}) && (4.42)\\
&= \log \frac{p\left(\mathbf{d} \,|\, \hat{\boldsymbol{\beta}}^{(t)}\right) p\left(\hat{\boldsymbol{\beta}}^{(t)}\right)}{p(\mathbf{d})} = \log p\left(\hat{\boldsymbol{\beta}}^{(t)} \,|\, \mathbf{d}\right) && (4.43)\\
&= \ell\left(\hat{\boldsymbol{\beta}}^{(t)}\right) && (4.44)\\
\text{(by Eq. 4.37)} \quad &\geq \hat{\ell}\left(q, \hat{\boldsymbol{\beta}}^{(t)}\right) \quad \forall q. && (4.45)
\end{aligned}$$

Since $\hat{\ell}(q, \hat{\boldsymbol{\beta}}^{(t)}) \leq \ell(\hat{\boldsymbol{\beta}}^{(t)})$ for any q, it is clear that the candidate solution $\tilde{q}$ solves the E-step, i.e.

$$q^{(t+1)} = p\left(\mathbf{m} \,|\, \mathbf{d}, \hat{\boldsymbol{\beta}}^{(t)}\right). \qquad (4.46)$$

Now, coming to the M-step, we can simplify its objective function by dropping all terms which do not depend on $\boldsymbol{\beta}$. Thus, plugging into (4.39) and employing (4.46), we can write:

$$\hat{\boldsymbol{\beta}}^{(t+1)} = \arg\max_{\boldsymbol{\beta}} \left\{ \log p(\boldsymbol{\beta}) + \mathbb{E}_{p\left(\mathbf{m} \,|\, \mathbf{d}, \hat{\boldsymbol{\beta}}^{(t)}\right)} \left[\log p(\mathbf{m}, \mathbf{d} \,|\, \boldsymbol{\beta})\right] \right\}. \qquad (4.47)$$

Because we were able to solve the E-step analytically, the E-M algorithm reduces to iterating the single step given by (4.47). We do not actually need to compute the Bayesian posterior in the E-step, but need only take the *expectation with respect to it* (which is why the E-step is so named). It can be shown that an iteration of the E-M algorithm (via Equation (4.47)) will never decrease the marginal posterior distribution for $\boldsymbol{\beta}$ (which is maximized by the marginal MAP solution), and, under very general conditions, the E-M algorithm does indeed converge to a (local) maximum of the original marginal MAP problem of (4.27) [46].

4.3.4 Application of E-M to LSM

We now proceed to apply the E-M algorithm to our LSM problem. For notational convenience, we can rewrite the E-M algorithm of (4.47) in terms of the E-M objective function $\phi^{(t)}(\boldsymbol{\beta})$ given by

$$\phi^{(t)}(\boldsymbol{\beta}) = \log p(\boldsymbol{\beta}) + \mathbb{E}_{p\left(\mathbf{m}\,|\,\mathbf{d},\hat{\boldsymbol{\beta}}^{(t)}\right)}\left[\log p(\mathbf{m},\mathbf{d}\,|\,\boldsymbol{\beta})\right], \tag{4.48}$$

so the E-M iteration becomes

$$\hat{\boldsymbol{\beta}}^{(t+1)} = \arg\max_{\boldsymbol{\beta}} \phi^{(t)}(\boldsymbol{\beta}). \tag{4.49}$$

For every iteration of E-M, we perform the maximization of $\phi^{(t)}$ via a gradient ascent scheme, for which we must compute the gradient of $\phi^{(t)}$.

To derive the exact form of $\phi^{(t)}$ and its gradient, we substitute our distributions into the E-M objective function. From (4.24), we have

$$\log p\left(\boldsymbol{\beta}\right) = \begin{cases} 0 & \text{if } \beta_{ij} \in [0,1], \quad \forall (i,j) \in \mathcal{E} \\ -\infty & \text{otherwise} \end{cases}, \tag{4.50}$$

which simply means the prior on $\boldsymbol{\beta}$ restricts us to consider only $\beta_{ij} \in [0,1]$. From (4.19) and (4.20), we have

$$\begin{aligned}\log p\left(\mathbf{m},\mathbf{d}\,|\,\boldsymbol{\beta}\right) = \frac{1}{2}\big(&\log\det\left(\lambda\left(D\left(\boldsymbol{\beta}\right)+\epsilon I\right)\right) - \mathbf{m}^T\left(\lambda\left(D\left(\boldsymbol{\beta}\right)+\epsilon I\right)\right)\mathbf{m} \\ &-(\mathbf{d}-A\mathbf{m})^T\Sigma^{-1}(\mathbf{d}-A\mathbf{m})\big) - Z,\end{aligned} \tag{4.51}$$

where Z is a normalization constant given by

$$Z = \frac{(N+K)\log 2\pi + \log\Sigma}{2}. \tag{4.52}$$

Inserting these into (4.48) yields (when every $\beta_{ij} \in [0,1]$):

$$\begin{aligned}\phi^{(t)}(\boldsymbol{\beta}) = \frac{1}{2}\,\mathbb{E}_{p\left(\mathbf{m}\,|\,\mathbf{d},\hat{\boldsymbol{\beta}}^{(t)}\right)}\big[&\log\det\left(\lambda\left(D\left(\boldsymbol{\beta}\right)+\epsilon I\right)\right) - \mathbf{m}^T\left(\lambda\left(D\left(\boldsymbol{\beta}\right)+\epsilon I\right)\right)\mathbf{m} \\ &-(\mathbf{d}-A\mathbf{m})^T\Sigma^{-1}(\mathbf{d}-A\mathbf{m})\big] - Z.\end{aligned} \tag{4.53}$$

The log determinant term in (4.53) only depends on $\boldsymbol{\beta}$ and is not affected by the expectation with respect to $\mathbf{m}$. Now, we can rewrite the second term in the above expectation as

$$\mathbf{m}^T \left(\lambda \left(D\left(\boldsymbol{\beta}\right) + \epsilon I\right)\right) \mathbf{m} = \lambda \operatorname{tr}\left(\left(D\left(\boldsymbol{\beta}\right) + \epsilon I\right) \mathbf{m}\mathbf{m}^T\right), \tag{4.54}$$

so

$$\mathbb{E}_{p\left(\mathbf{m}\,|\,\mathbf{d},\hat{\boldsymbol{\beta}}^{(t)}\right)}\left[\mathbf{m}^T \left(\lambda \left(D\left(\boldsymbol{\beta}\right) + \epsilon I\right)\right) \mathbf{m}\right] = \lambda \operatorname{tr}\left(\left(D\left(\boldsymbol{\beta}\right) + \epsilon I\right) \mathbb{E}_{p\left(\mathbf{m}\,|\,\mathbf{d},\hat{\boldsymbol{\beta}}^{(t)}\right)}\left[\mathbf{m}\mathbf{m}^T\right]\right). \tag{4.55}$$

The expected value on the right-hand side of (4.55) is just the non-central second moment matrix of $\mathbf{m}$, as determined by the posterior distribution $p(\mathbf{m}\,|\,\mathbf{d}, \hat{\boldsymbol{\beta}}^{(t)})$, given by

$$\mathbb{E}_{p\left(\mathbf{m}\,|\,\mathbf{d},\hat{\boldsymbol{\beta}}^{(t)}\right)}\left[\mathbf{m}\mathbf{m}^T\right] = \Lambda^{(t)} + \mu^{(t)}{\mu^{(t)}}^T, \tag{4.56}$$

and where $\mu^{(t)}$ and $\Lambda^{(t)}$ are the posterior mean and covariance matrix, respectively, when conditioning on $\mathbf{d}$ and $\hat{\boldsymbol{\beta}}^{(t)}$, given by

$$\mu^{(t)} = \left(A^T\Sigma^{-1}A + \lambda\left(D(\hat{\boldsymbol{\beta}}^{(t)}) + \epsilon I\right)\right)^{-1} A^T\Sigma^{-1}\mathbf{d} \tag{4.57}$$

and

$$\Lambda^{(t)} = \left(A^T\Sigma^{-1}A + \lambda\left(D(\hat{\boldsymbol{\beta}}^{(t)}) + \epsilon I\right)\right)^{-1}. \tag{4.58}$$

We further note that the $\epsilon I\, \mathbb{E}[\mathbf{m}\mathbf{m}^T]$ term in (4.55) does not depend on the variable $\boldsymbol{\beta}$ which is being optimized and hence can be dropped from the E-M objective function $\phi^{(t)}(\boldsymbol{\beta})$. Similarly, the third and fourth terms in (4.53) also do not depend on $\boldsymbol{\beta}$ and can be neglected. Combining the above and rearranging terms, we can rewrite the E-M objective function as

$$\phi^{(t)}(\boldsymbol{\beta}) = \frac{1}{2}\left(\log\det\left(\lambda\left(D\left(\boldsymbol{\beta}\right) + \epsilon I\right)\right) - \lambda \operatorname{tr}\left(D\left(\boldsymbol{\beta}\right)\Lambda^{(t)}\right) - \lambda\, {\mu^{(t)}}^T D\left(\boldsymbol{\beta}\right)\mu^{(t)}\right). \tag{4.59}$$

In order to compute $\nabla\phi^{(t)}$, the gradient of $\phi^{(t)}$ with respect to $\boldsymbol{\beta}$, we first note that the $\boldsymbol{\beta}$-weighted graph Laplacian matrix $D(\boldsymbol{\beta})$ is a linear function of $\boldsymbol{\beta}$, particularly:

$$D(\boldsymbol{\beta}) = \sum_{(i,j)\in\mathcal{E}} \beta_{ij} P^{ij} \tag{4.60}$$

where the entries of P^{ij} are

$$P^{ij}_{kl} = \begin{cases} 1 & \text{if } kl = ii \text{ or } jj \\ -1 & \text{if } kl = ij \text{ or } ji \\ 0 & \text{otherwise} \end{cases} . \tag{4.61}$$

We also note that

$$\frac{\partial}{\partial \beta_{ij}} \log\det\left(\lambda(D(\boldsymbol{\beta}) + \epsilon I)\right) = \operatorname{tr}\left((\lambda(D(\boldsymbol{\beta}) + \epsilon I))^{-1} \frac{\partial(\lambda(D(\boldsymbol{\beta}) + \epsilon I))}{\partial \beta ij} \right).$$

Letting $C(\boldsymbol{\beta}) = (\lambda(D(\boldsymbol{\beta}) + \epsilon I))^{-1}$ denote the prior covariance matrix of the image (when conditioning on $\boldsymbol{\beta}$), to compute $\nabla\phi^{(t)}$, we have:

$$\frac{\partial}{\partial \beta_{ij}} \phi^{(t)}(\boldsymbol{\beta}) = \frac{\lambda}{2}\left(\operatorname{tr}\left(C(\boldsymbol{\beta})P^{ij}\right) - \operatorname{tr}\left(\Lambda^{(t)}P^{ij}\right) - {\mu^{(t)}}^T P^{ij} \mu^{(t)} \right) \tag{4.62}$$

$$= \frac{\lambda}{2}\left(C(\boldsymbol{\beta})_{ii} + C(\boldsymbol{\beta})_{jj} - 2C(\boldsymbol{\beta})_{ij} - \left(\Lambda^{(t)}_{ii} + \Lambda^{(t)}_{jj} - 2\Lambda^{(t)}_{ij} \right) - \left(\mu^{(t)}_i - \mu^{(t)}_j \right)^2 \right). \tag{4.63}$$

We constrain each β_{ij} to the interval $[0,1]$ by introducing proxy variables γ_{ij} which we map to the β_{ij} using a sigmoidal function. In particular we set

$$\beta_{ij} = \frac{\arctan(\gamma_{ij})}{\pi} + \frac{1}{2}, \tag{4.64}$$

so that while γ_{ij} is free to take any value in $\mathbb{R}$, β_{ij} remains within $[0,1]$. We can then compute $\nabla\phi^{(t)}(\boldsymbol{\gamma})$, the gradient of $\phi^{(t)}$ with respect to $\boldsymbol{\gamma}$, by

$$\frac{\partial}{\partial\gamma_{ij}}\phi^{(t)}(\boldsymbol{\gamma}) = \frac{\partial\phi^{(t)}}{\partial\beta_{ij}}\frac{\partial\beta_{ij}}{\partial\gamma_{ij}} \tag{4.65}$$

$$= \frac{1}{\pi(1+\gamma_{ij}^2)}\frac{\partial\phi^{(t)}}{\partial\beta_{ij}}. \tag{4.66}$$

Unfortunately, direct computation of the gradient would require matrix inversions to compute both the prior and posterior covariance matrices. To avoid this, we note that we only need the node and edge-wise elements of these covariance matrices, and we could instead sample from their associated Gaussian probability distributions and approximate these elements from the samples. Thus, to approximate the prior covariance matrix $C(\boldsymbol{\beta})$, we generate L samples $\mathbf{m}^{(1)},\ldots,\mathbf{m}^{(L)}$, of the underlying Gaussian prior distribution of $\mathbf{m}|\boldsymbol{\beta}$ and approximate C as:

$$C(\boldsymbol{\beta}) \approx \frac{1}{L}\sum_{\ell=1}^{L}\mathbf{m}^{(\ell)}\mathbf{m}^{(\ell)T}. \tag{4.67}$$

We describe the sampling algorithm we use to approximate the elements of $C(\boldsymbol{\beta})$ in the following section.

Perturbation-Optimization Sampling of Gaussian Distributions

To sample from $\mathcal{N}(\mathbf{0}, C)$, we first note that the precision matrix $Q = \lambda(D(\boldsymbol{\beta}) + \epsilon I)$ can be rewritten as

$$Q = \lambda(F^T B(\boldsymbol{\beta})F + \epsilon I) \tag{4.68}$$

where F is a first-differencing matrix (having number of rows equal to $|\mathcal{E}|$, the number of edges in $\mathcal{E}$, and and number of columns equal to N, the number of image parameters) and $B(\boldsymbol{\beta})$ is an $|\mathcal{E}|$-by-$|\mathcal{E}|$ diagonal matrix, with the β_{ij} on its diagonal. Referred to as Perturbation-Optimization (P-O) sampling by Orieux et al. [53], a straight-forward sampling algorithm (that avoids the need for Cholesky factorization of the precision matrix) is available when the precision matrix can be expressed in the form

$$Q = \sum_{t=1}^{T} M_t^T R_t^{-1} M_t \tag{4.69}$$

and sampling from $\mathcal{N}(\mathbf{0}, R_t)$ is feasible (which is certainly true in our case, as we have diagonal R_t matrices). The sampling algorithm is then as follows:

Algorithm 4.1 Perturbation-Optimization algorithm for sampling from $\mathcal{N}(\mathbf{0}, Q^{-1})$ [53]

```
1. Perturbation step: Generate independent vectors
     $\boldsymbol{\eta}_t \sim \mathcal{N}(\mathbf{0}, R_t) \qquad \text{for } t = 1, \ldots, T$
2. Optimization step: Compute $\hat{\mathbf{m}}$ as the minimizer of
     $J(\mathbf{m}) = \sum_{t=1}^{T} (\boldsymbol{\eta}_t - M_t \mathbf{m})^T R_t^{-1} (\boldsymbol{\eta}_t - M_t \mathbf{m})$
Return $\hat{\mathbf{m}}$ as the sample from $\mathcal{N}(\mathbf{0}, Q^{-1})$.
```

The proof that $\hat{\mathbf{m}}$ is a sample from $\mathcal{N}(\mathbf{0}, Q^{-1})$ is straight-forward and given in Orieux et al. [53]. The optimization step simply requires solving the linear system

$$Q\hat{\mathbf{m}} = \sum_{t=1}^{T} M_t^T R_t^{-1} \boldsymbol{\eta}_t, \tag{4.70}$$

which, in our case, is very fast ($\mathcal{O}(kN)$ using an iterative solver with k steps) due to the sparsity of F.

Block Diagonal Approximations

While this sampling approach can also be used to approximate the elements of the posterior covariance matrix $\Lambda^{(t)}$, in practice generating a reasonably large number of samples from $\mathcal{N}(\mathbf{0}, \Lambda^{(t)})$ is not feasible due to the increased cost of solving a system involving the posterior precision matrix $A^T \Sigma^{-1} A + \lambda(D(\hat{\boldsymbol{\beta}}^{(t)}) + \epsilon I)$ (we would need to perform a regularized LSM inversion for *each sample* when using the P-O approach).

In order to estimate the node and edge-wise elements of $\Lambda^{(t)}$, we note that when using the Kirchhoff operator A, there is a closed form expression for the elements of the posterior precision matrix ${\Lambda^{(t)}}^{-1}$ (combining (4.58) and (4.4)). With this in mind, we can estimate elements of $\Lambda^{(t)}$ by considering a block diagonal approximation to the precision matrix. In particular, we can construct an M-by-M partition of the posterior precision matrix corresponding to an image point and its $M-1$ nearest neighbors in space within some radius (we used a 49-pixel neighborhood to perform this approximation), then approximate the covariance matrix at image point i, $\Lambda_{ii}^{(t)}$, from the inverse of this M-by-M partition matrix. The off-diagonal elements $\Lambda_{ij}^{(t)}$ (for each edge $(i, j) \in \mathcal{E}$) are similarly estimated from the same matrix inverse by taking the elements corresponding to covariance between m_i and m_j (however, care must be taken to ensure that the M-by-M partition of the precision matrix is large enough to sufficiently "surround" both the image point i and all its neighbors j with which it shares an edge). This approximation will perform reasonably well as long as the posterior precision matrix decays spatially (in the model domain) as we move away from the diagonal (as is the case here).

Summary of E-M Algorithm

To implement the above approximations to calculate $\nabla\phi^{(t)}$, we need to approximate the entries of $\Lambda^{(t)}$ only once per E-M iteration (as $\Lambda^{(t)}$ does not vary with $\boldsymbol{\beta}$). However we would need to reapproximate the entries of $C(\boldsymbol{\beta})$ with the sampling algorithm in each iteration of the first-order gradient-ascent method (which must be re-run in each iteration of the E-M algorithm). We now summarize our above developments for applying the E-M algorithm to obtain the empirical Bayesian estimate of the image in LSM in the algorithm below:

Algorithm 4.2 Expectation-Maximization Algorithm for LSM

```
Initialize each $\gamma_{ij}^{(0)} = 0$, so that $\hat{\beta}_{ij}^{(0)} = 0.5$.
Specify step-size $\alpha$ for gradient ascent.
Set $t = 0$. Iterate on $t$:
1. Compute $\boldsymbol{\mu}^{(t)}$ via (4.57).
2. Compute $\Lambda_{ii}^{(t)}$ $(\forall i \in \mathcal{V})$ and $\Lambda_{ij}^{(t)}$ $(\forall (i,j) \in \mathcal{E})$ via block diagonal approximation.
3. Initialize $\tilde{\boldsymbol{\gamma}}^{(0)} = \boldsymbol{\gamma}^{(t)}$. Set $s = 0$ and iterate on $s$ to perform gradient ascent on $\boldsymbol{\gamma}$:
    a. Generate samples from $\mathcal{N}(0, C(\boldsymbol{\beta}(\tilde{\boldsymbol{\gamma}}^{(s)})))$ via P-O sampling.
    b. Estimate $C(\boldsymbol{\beta}(\tilde{\boldsymbol{\gamma}}^{(s)}))_{ii}$ $(\forall i \in \mathcal{V})$ and $C(\boldsymbol{\beta}(\tilde{\boldsymbol{\gamma}}^{(s)}))_{ij}$ $(\forall (i,j) \in \mathcal{E})$ via (4.67).
    c. Compute $\nabla \phi^{(t)}(\tilde{\boldsymbol{\gamma}}^{(s)})$ via (4.66).
    d. Update $\tilde{\boldsymbol{\gamma}}^{(s+1)} = \tilde{\boldsymbol{\gamma}}^{(s)} + \alpha \nabla \phi^{(t)}(\tilde{\boldsymbol{\gamma}}^{(s)})$
4. Update $\boldsymbol{\gamma}^{(t+1)} = \tilde{\boldsymbol{\gamma}}^{(s+1)}$.
5. Update $\hat{\boldsymbol{\beta}}^{(t+1)}$ via (4.64) using $\boldsymbol{\gamma}^{(t+1)}$.
Upon termination, return:
    $\boldsymbol{\beta}_{\text{MAP}} = \hat{\boldsymbol{\beta}}^{(t+1)}$,
    $\mathbf{m}_{\text{EB}} = \left(A^T \Sigma^{-1} A + \lambda \left(D\left(\boldsymbol{\beta}_{\text{MAP}}\right) + \epsilon I\right)\right)^{-1} A^T \Sigma^{-1} \mathbf{d}$.
```

4.4 Results

In order to validate our approach, we ran our inference algorithm on synthetic datasets. We present two test cases: the first case being a simple example where the data arise from a small image consisting of three dipping reflectors separated by a weakly reflective fault and the second case being data simulated from the Marmousi model. Synthetic data were created using the same Kirchhoff modeling operator A that is used in the inference algorithms. Hence, these test cases are what are known as *inverse crime* tests. The purpose of using the same forward modeling operator to create the synthetic data as is used in the inference is to isolate the *inversion* problem from the *modeling* problem. In order to somewhat avoid the inverse crime, we add zero-mean white Gaussian noise to the data (with a standard deviation equal to

10% of the maximum amplitude of the data, assumed to be known by our inversion procedure).

We first describe the example of the three dipping reflectors. The data are created from a single surface seismic source (at the center) and 50 equally spaced surface seismic receivers (with spacing of 50 m) using a homogeneous background velocity model (of 4000 m/s). The source wavelet is a 20 Hz Ricker wavelet; hence the dominant wavelength is 200 m. The seismic traces are sampled at 1 ms, and the medium is sampled spatially at 50 m in both the lateral and vertical directions. The entire medium has spatial dimensions of 2500 m by 2500 m, hence $N_x = N_z = 50$ and the number of image parameters is $N = N_x N_z = 2500$. The purpose on testing our algorithm on such a small model is so that we can verify the performance of our algorithm in the absence of any approximations (i.e. in this case, we can directly compute the elements of C and Λ without the need of P-O sampling or block diagonal approximations). For this example, we used an MRF in which each node shares an edge with its four nearest neighbors. Here, we ran 10 iterations of the E-M algorithm to obtain the MAP estimate of the edge strengths and the empirical Bayes image, where each iteration of the E-M algorithm ran in approximately 1 minute on a quad core Intel$^{\text{TM}}$ Xeon W3550 3.0GHz processor.

For the case of the Marmousi model, we use a smoothed version of the true Marmousi velocity model (sampled at 24 m spacing) for our background velocity model in conjunction with the true (unsmoothed) reflectivity model to simulate the data. The data are created from a set of 20 collocated surface sources and receivers (resulting in 400 traces), with a 480 m spacing between stations, where the source wavelet is a 25 Hz Ricker wavelet. For this case, we must resort to the approximate methods outlined above (P-O sampling and block diagonal approximations) to compute the gradient $\nabla\phi^{(t)}$. Here, in order to capture the more complex dipping structures of the Marmousi model, we defined the MRF so that each node shares an edge with all nodes within a radius of $\sqrt{2}$ nodes (i.e. a node shares an edge with its four diagonal neighbors in addition to its four nearest neighbors). In this example, the MAP estimate of the edge strengths and the empirical Bayes image were obtained with

3 iterations of the E-M algorithm, where each iteration of the E-M algorithm took approximately 33 minutes on a quad core Intel$^{\text{TM}}$ Xeon W3550 3.0GHz processor. We note that each iteration of the E-M algorithm requires performing a standard least-squares migration in addition to the computation required to obtain the block diagonal approximation to the posterior covariance matrix and perform the optimizations required for P-O sampling from the prior. The LSM images are computed using the method of conjugate gradients (CG), which is run for 200 iterations per LSM. Each iteration of CG requires performing a single Kirchhoff modeling followed by a single Kirchhoff migration, and, for this example, takes about 0.7 seconds per CG iteration (on the same machine).

Figures 4-3–4-9 show the results from the test case of the three dipping layer model, where the edge strengths obtained using our algorithm are shown in Figure 4-9 and the resulting image obtained using these edge strengths is shown in Figure 4-8. Performing a Kirchhoff migration on the data results in the image of Figure 4-5; here the reflectors are imaged somewhat, but we also see heavy imaging artifacts (i.e. the migration smiles) due to the limited source-receiver geometry (where only a single source is being used). We observe that in the case of the unregularized LSM image (Figure 4-6), the reflectors are imaged, but unfortunately, the noise in the data is also imaged so strongly that the reflectors are nearly impossible to distinguish from the noise. We can improve on the unregularized image by using a uniform regularization scheme (setting each $\beta_{ij} = 1$) to obtain the regularized LSM image of Figure 4-7; here, the use of regularization has filtered out the noise, but as a side effect has also smoothed out the reflectors. The empirical Bayesian MAP image (Figure 4-8) obtained by using our estimate of the edge strengths significantly improves upon this result. This is clear from a qualitative comparison between the images; we can see the reflectors imaged quite strongly with sharpness preserved at the reflectors, while the noise is filtered out elsewhere in the image. Additionally, the weakly reflective fault is also slightly imaged in the empirical Bayesian MAP image, whereas it cannot be seen in the other images. We further note that the correlation of the empirical Bayesian MAP image with the true image is significantly higher than the correlations of the other images with the true image. Examining the estimate of the edge strengths in Figure 4-9, we see that the edge strengths take on a pattern similar to our expectations: they are high where the image is constant, but close to 0 where there are differences in the image (surrounding the reflectors).

Figures 4-10–4-17 show the results from the test case with the Marmousi model. We again observe the same features in the images as seen in the 3 layer test case. The unregularized LSM image (Figure 4-14) shows the reflectors along with a very strong noise component. Regularizing in a uniform fashion (by setting each $\beta_{ij} = 1$) results in the regularized image of Figure 4-15 in which the noise has been filtered out, but the image is also overly smooth in some areas. Once again, using our algorithm to estimate the edge strengths (which are shown in Figure 4-17) results in the empirical Bayesian MAP image of Figure 4-16. We notice the same qualitative improvements in the image as seen previously: the image remains sharp near the reflectors while smoothing out the noise away from the reflectors. And, as before, the correlation of the empirical Bayesian MAP image with the true image is significantly higher than the correlations of the other images with the true image.

4.5 Conclusions and Future Work

Our study shows that the Bayesian framework provides a flexible methodology for estimating both the image and smoothness parameters (or edge strengths) in a least-squares migration setting. By estimating the edge strengths, we are able to remove the effects of noise while, by and large, preserving sharpness at the reflectors in the image. The expectation-maximization algorithm, in particular, allowed us to solve the marginal MAP problem for estimating the edge strengths $\boldsymbol{\beta}$ (without having to explicitly compute the marginal posterior distribution for $\boldsymbol{\beta}$).

We note that while our algorithm was presented within the context of the seismic imaging problem, the methodology we have developed is broadly applicable to many linear inverse problems where the model parameters may exhibit spatially (or temporally) varying smoothness properties. The operator A (or, more generally, the

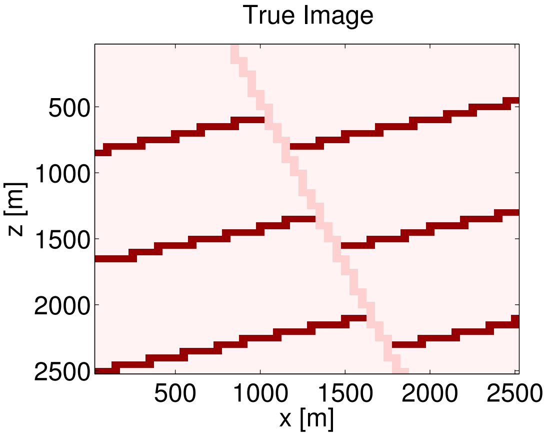


Figure 4-3: True image for the three layer test case.

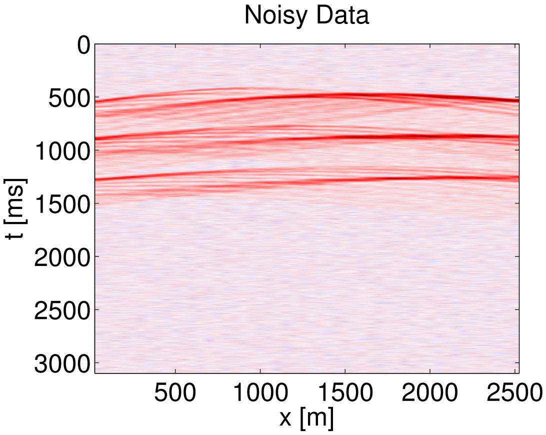


Figure 4-4: Noisy synthetic data for the three layer test case.

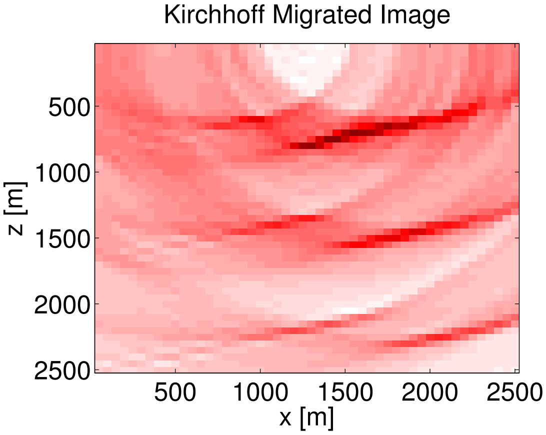


Figure 4-5: Kirchhoff migrated image for the three layer test case. Correlation with true image = 0.4705.

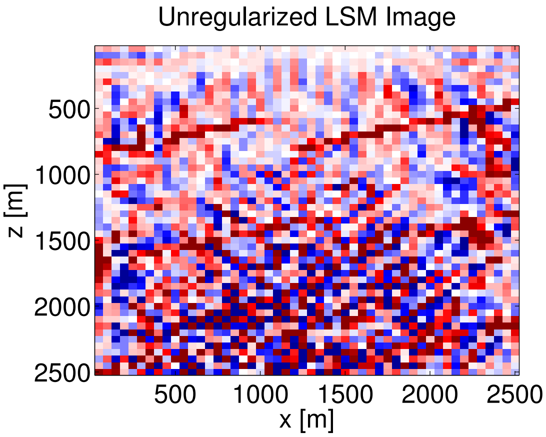


Figure 4-6: Unregularized LSM image (each $\beta_{ij} = 0$) for the three layer test case. Correlation with true image $= 0.3649$.

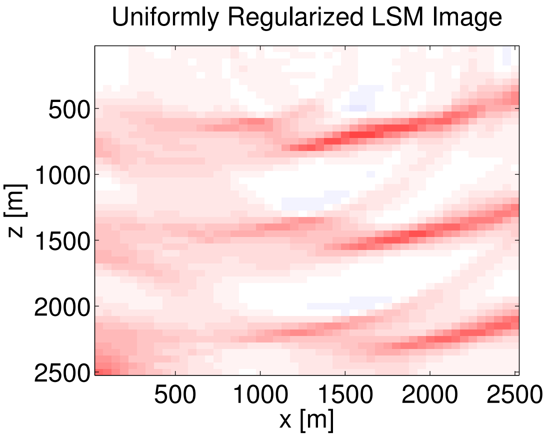


Figure 4-7: Uniformly regularized LSM image (each $\beta_{ij} = 1$) for the three layer test case. Correlation with true image $= 0.5879$.

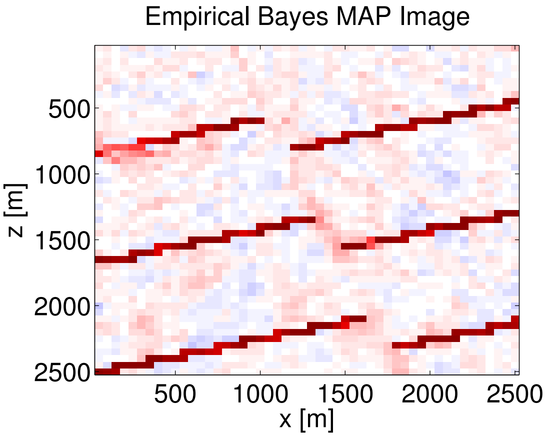


Figure 4-8: Empirical Bayesian MAP image (computed after estimating $\boldsymbol{\beta}$) for the three layer test case. Correlation with true image $= 0.9607$.

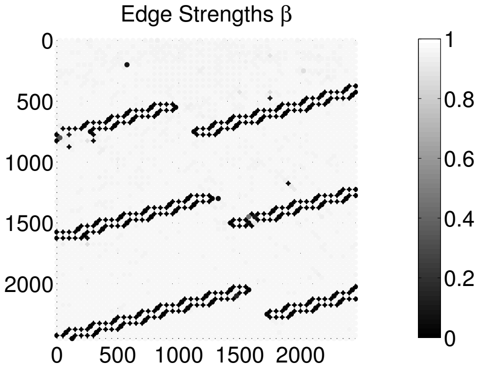


Figure 4-9: Edge strengths $\boldsymbol{\beta}$ estimated with E-M algorithm for the three layer test case.

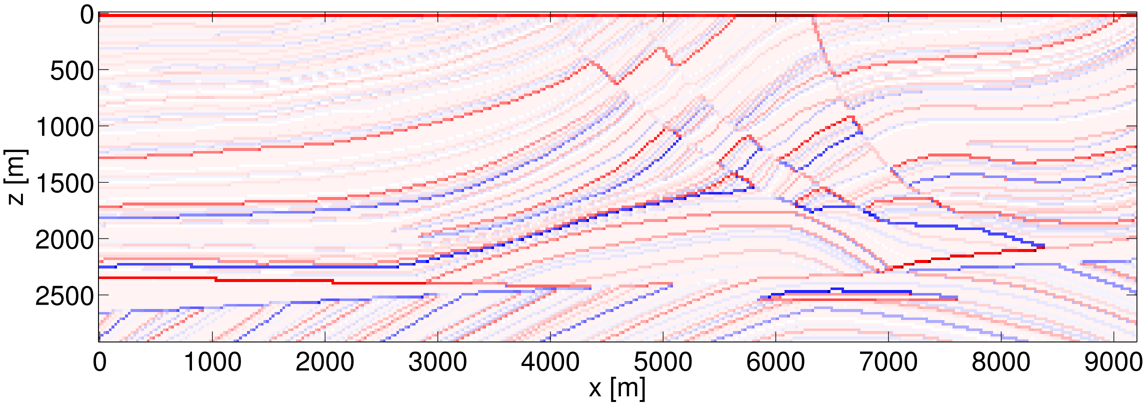


Figure 4-10: True image for the Marmousi test case.

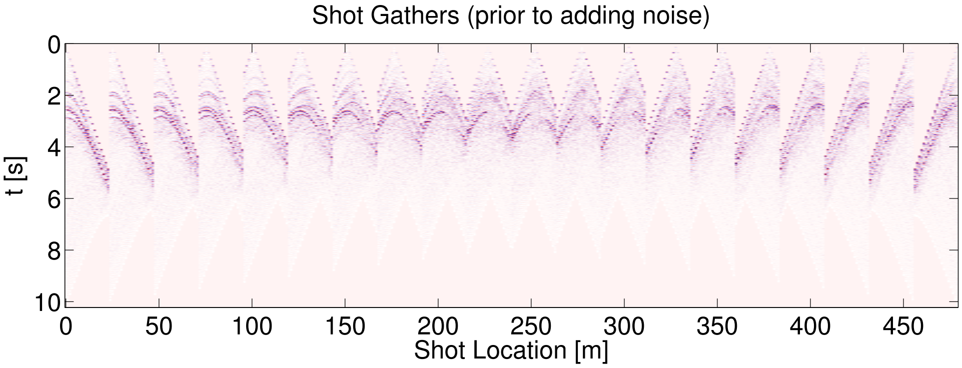


Figure 4-11: Synthetic data for the Marmousi test case prior to adding noise.

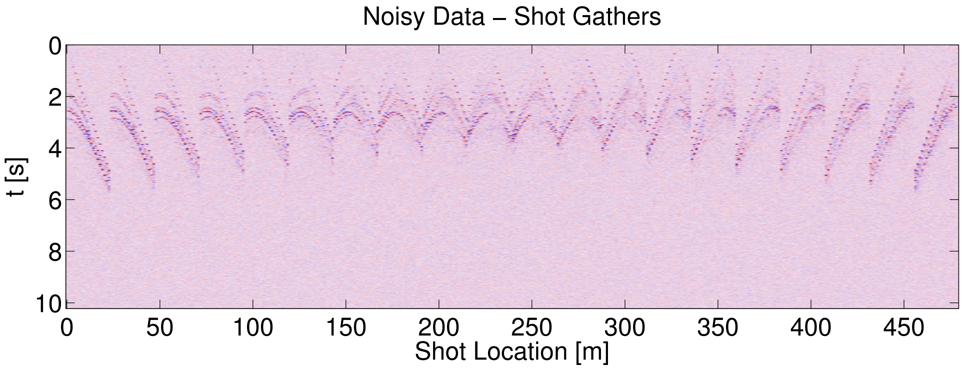


Figure 4-12: Noisy synthetic data for the Marmousi test case.

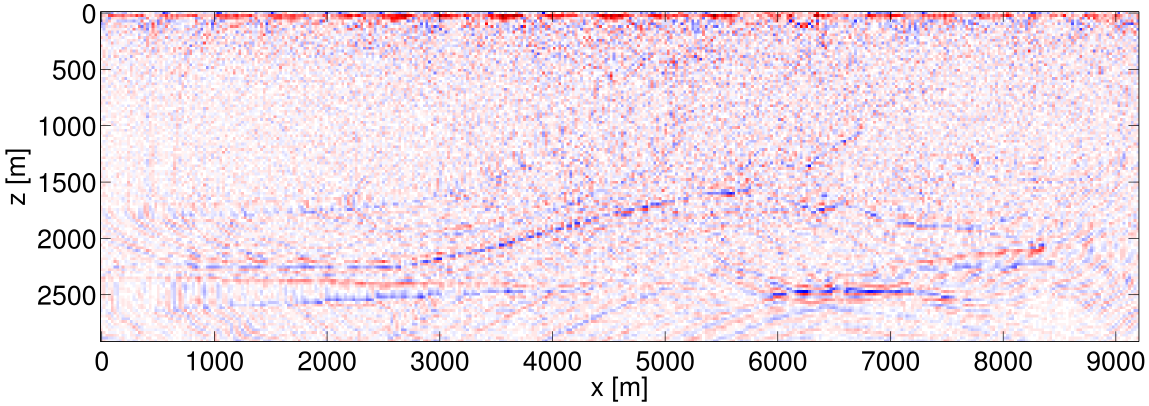


Figure 4-13: Kirchhoff migrated image for the Marmousi test case. Correlation with true image = 0.4371.

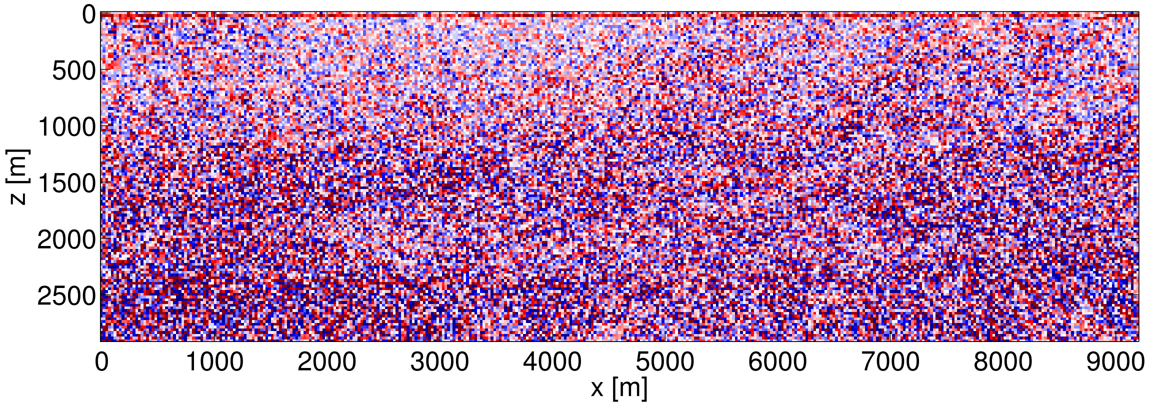


Figure 4-14: Unregularized LSM image (each $\beta_{ij} = 0$) for the Marmousi test case. Correlation with true image = 0.3682.

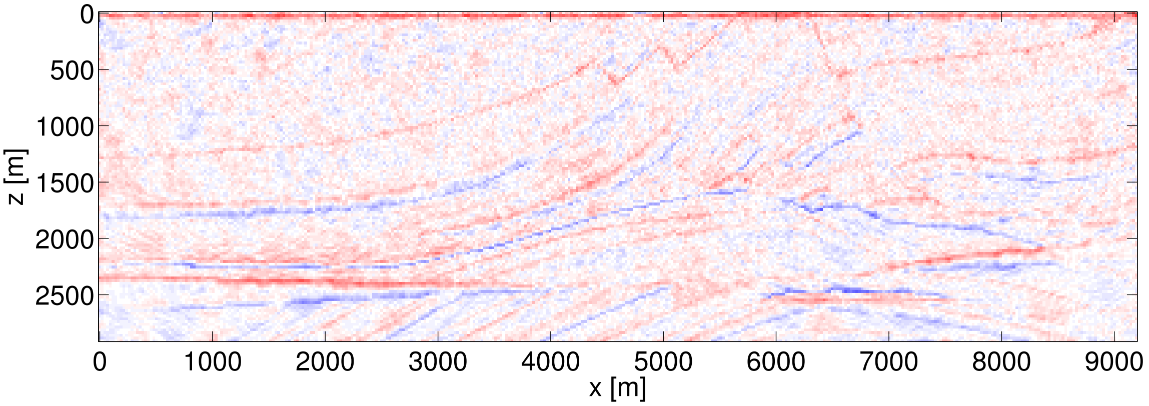


Figure 4-15: Uniformly regularized LSM image (each $\beta_{ij} = 1$) for the Marmousi test case. Correlation with true image = 0.6369.

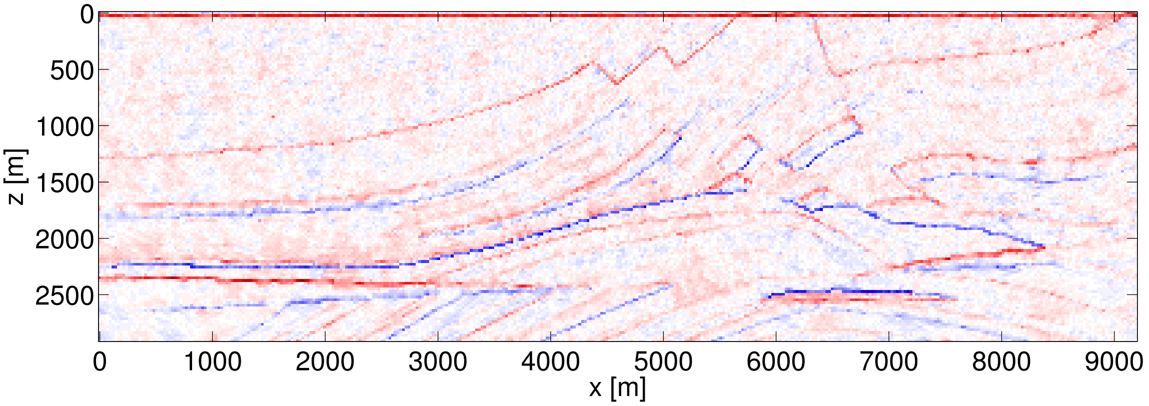


Figure 4-16: Empirical Bayesian MAP image (computed after estimating $\boldsymbol{\beta}$) for the Marmousi test case. Correlation with true image = 0.7973.

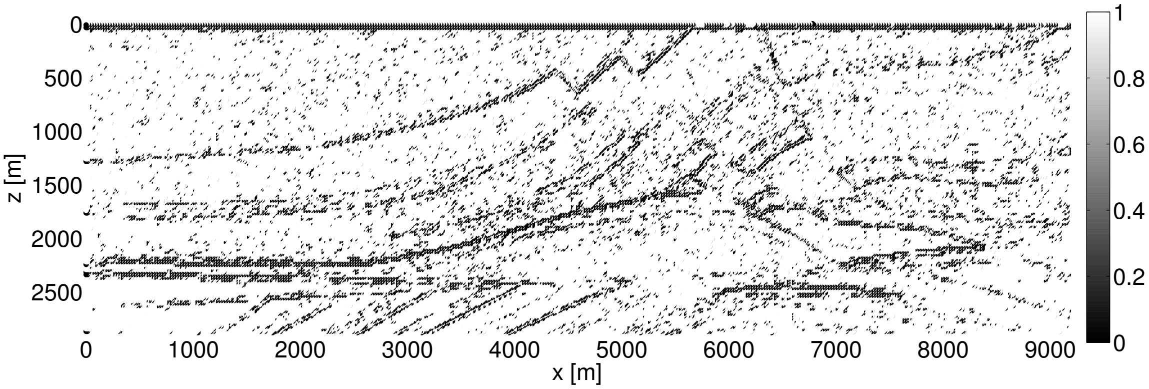


Figure 4-17: Edge strengths $\boldsymbol{\beta}$ estimated with E-M algorithm for the Marmousi test case.

conditional distribution for the data given the model $p(\mathbf{d}|\mathbf{m})$) would change if we were solving a different problem, but the methodology and algorithm described in this chapter would still apply.

While we have developed our algorithm in the setting of solving a linear inverse problem, an interesting direction for future work is to generalize this methodology to non-linear inverse problems. A second direction for future work is to explore alternative ways to parameterize the prior on the image within the hierarchical Bayesian setting. We say more about these two future directions in Chapter 7. Additionally, with the parameterization of the prior presented in this chapter, one may wish to explore inferring other parameters than just the edge strengths. For example, rigorously picking the regularization parameter λ remains an open question in the field of inverse problems. Another natural future direction is application of this methodology to a more realistic synthetic dataset (or to a field dataset), where we expect similar improvements in quality of the resulting image. The latter two directions are explored in the following chapter.

Chapter 5

Interpretation and Estimation of Regularization Parameters

In this chapter, we undertake a more rigorous investigation into the regularization parameters used in Chapter 4. In the first part of this chapter, we describe how these parameters characterize the prior covariance of the model by exploring the connections between these parameters and the covariance function that arises in the limiting case when the model is treated as a random function. In the second part of this chapter, we generalize the methodology of Chapter 4 to estimate these parameters within the hierarchical Bayesian framework.

Recall the probabilistic model for the edge strengths $\boldsymbol{\beta}$, image $\mathbf{m}$, and data $\mathbf{d}$ described in Chapter 4 (where image $\mathbf{m}$ is defined on the graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$) :

$$\boldsymbol{\beta} \sim \text{Uniform}([0,1]^{|\mathcal{E}|}), \tag{5.1}$$

$$\mathbf{m} \,|\, \boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, C(\boldsymbol{\beta})), \tag{5.2}$$

$$\mathbf{d} \,|\, \mathbf{m}, \boldsymbol{\beta} \sim \mathcal{N}(A\mathbf{m}, \Sigma), \tag{5.3}$$

where the prior precision matrix $Q(\boldsymbol{\beta}) = C^{-1}(\boldsymbol{\beta})$ for $\mathbf{m} \,|\, \boldsymbol{\beta}$ is given by

$$Q(\boldsymbol{\beta}) = C^{-1}(\boldsymbol{\beta}) = \lambda(D(\boldsymbol{\beta}) + \epsilon I), \tag{5.4}$$

and where $\lambda > 0$, $\epsilon > 0$, and $D(\boldsymbol{\beta})$ is the $\boldsymbol{\beta}$-weighted graph Laplacian on $\mathcal{G}$ defined by the quadratic form

$$\mathbf{m}^T D(\boldsymbol{\beta})\mathbf{m} = \sum_{(i,j)\in\mathcal{E}} \beta_{ij}(m_i - m_j)^2. \tag{5.5}$$

5.1 Connecting Regularization Parameters and Covariance Functions

In order to better understand the significance of the parameters λ, $\boldsymbol{\beta}$, and ϵ, we investigate the connections between these parameters and the resulting prior model covariance. In particular, we examine the impact of these parameters in the continuous case, where the model $\mathsf{m}(\mathbf{x})$ is now treated as a zero-mean Gaussian random field with model covariance specified by a covariance function $C(\mathbf{x}, \mathbf{x}')$ such that

$$\mathbb{E}[\mathsf{m}(\mathbf{x})\mathsf{m}(\mathbf{x}')] = C(\mathbf{x}, \mathbf{x}'). \tag{5.6}$$

Following the development in Rodi and Myers [59] and Simpson et al. [68], we recognize that the covariance function $C(\mathbf{x}, \mathbf{x}')$ can be taken as the Green's function of a differential operator L, so that, subject to appropriate boundary conditions, we have

$$LC(\mathbf{x}, \mathbf{x}') = \delta(\mathbf{x} - \mathbf{x}'), \tag{5.7}$$

where $C(\mathbf{x}, \mathbf{x}')$ will be a covariance function when L is self-adjoint and positive-definite [59]. To connect our finite-dimensional model of Equation (5.2) to this Gaussian random field setting, we note that the covariance matrix C is a discrete approximation to the covariance function $C(\mathbf{x}, \mathbf{x}')$ when the precision matrix $Q = C^{-1}$ is a finite-difference approximation to the differential operator L [59]. Hence, we want to choose the differential operator L for which $Q(\boldsymbol{\beta})$ is a finite-difference approximation, i.e. we want L such that

$$\int_V m(\mathbf{x}) L m(\mathbf{x}) \, \mathrm{d}V(\mathbf{x}) \approx \mathbf{m}^T Q \mathbf{m} \tag{5.8}$$

$$= \lambda \left(\sum_{(i,j)\in\mathcal{E}} \beta (m(x_i) - m(x_j))^2 + \epsilon \sum_{i\in\mathcal{V}} m(x_i)^2 \right), \tag{5.9}$$

where V denotes the model domain and $\mathrm{d}V(\mathbf{x})$ is the volume measure at $\mathbf{x}$ and where we have considered the special case of Q where the β_{ij} are all set to a common value $\beta_{ij} = \beta > 0$, so that the underlying model covariance will be stationary far from the boundaries. For the 2-D problem we are considering here, we take V to be a box in $\mathbb{R}^2$ and $\mathrm{d}V(\mathbf{x}) = \mathrm{d}x \, \mathrm{d}z$. Letting ℓ denote the spacing of the uniform grid on which the finite-dimensional model $\mathbf{m}$ is defined (so a grid cell has area ℓ^2), we consider the following candidate for L:

$$L = \lambda\beta \left(\frac{\epsilon}{\beta\ell^2} - \Delta \right), \tag{5.10}$$

where Δ denotes the Laplacian operator in 2-D ($\Delta = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial z^2}$) and use the Dirichlet boundary condition $m(\mathbf{x}) = 0$ on the boundary ∂V. Then, we have:

$$\int_V m(\mathbf{x}) L m(\mathbf{x}) \, \mathrm{d}V = \lambda\beta \left(\frac{\epsilon}{\beta\ell^2} \int_V m^2 \, \mathrm{d}V - \int_V m\Delta m \, \mathrm{d}V \right). \tag{5.11}$$

Now noting that

$$\nabla \cdot (m\nabla m) = \|\nabla m\|_2^2 + m\Delta m, \tag{5.12}$$

we can rewrite (5.11) as

$$\int_V m(\mathbf{x}) L m(\mathbf{x}) \, \mathrm{d}V = \lambda\beta \left(\frac{\epsilon}{\beta\ell^2} \int_V m^2 \, \mathrm{d}V + \int_V \|\nabla m\|_2^2 \, \mathrm{d}V - \int_V \nabla \cdot (m\nabla m) \, \mathrm{d}V \right). \tag{5.13}$$

Applying the divergence theorem to the last integral in (5.13),

$$\int_V \nabla \cdot (m\nabla m) \, \mathrm{d}V = \int_{\partial V} (m\nabla m) \cdot \mathbf{n} \, \mathrm{d}S = 0, \tag{5.14}$$

where the last equality holds due to the boundary condition, and (5.13) becomes

$$\int_V m(\mathbf{x}) L m(\mathbf{x})\, \mathrm{d}V = \lambda\beta \left(\frac{\epsilon}{\beta\ell^2} \int_V m^2 \, \mathrm{d}V + \int_V \|\nabla m\|_2^2 \, \mathrm{d}V \right). \tag{5.15}$$

From (5.15), we see that L is indeed positive-definite: $\int_V m(\mathbf{x}) L m(\mathbf{x})\, \mathrm{d}V > 0$ for all $m \neq 0$ satisfying the boundary condition, and it can similarly be shown that L is self-adjoint. Now we proceed to approximate the integrals and derivatives in (5.15) with summations and differences:

$$\begin{aligned}
\int_V m(\mathbf{x}) L m(\mathbf{x})\, \mathrm{d}V &= \lambda\beta \left(\frac{\epsilon}{\beta\ell^2} \int_V m^2 \, \mathrm{d}V + \int_V \|\nabla m\|_2^2 \, \mathrm{d}V \right) && (5.16)\\
&\approx \lambda\beta \Bigg(\frac{\epsilon}{\beta\ell^2} \sum_{i=1}^{N_x} \sum_{j=1}^{N_z} m^2\left((x_i, z_j)\right) \ell^2 \\
&\qquad + \sum_{i=1}^{N_z} \sum_{j=1}^{N_x - 1} \frac{\left(m\left((x_j, z_i)\right) - m\left((x_{j+1}, z_i)\right)\right)^2}{\ell^2} \ell^2 && (5.17)\\
&\qquad + \sum_{i=1}^{N_x} \sum_{j=1}^{N_z - 1} \frac{\left(m\left((x_i, z_j)\right) - m\left((x_i, z_{j+1})\right)\right)^2}{\ell^2} \ell^2 \Bigg) \\
&= \lambda \left(\epsilon \sum_{i \in \mathcal{V}} m^2(\mathbf{x}_i) + \sum_{(i,j) \in \mathcal{E}} \beta (m(\mathbf{x}_i) - m(\mathbf{x}_j))^2 \right) && (5.18)\\
&= \mathbf{m}^T Q \mathbf{m}. && (5.19)
\end{aligned}$$

Hence our candidate L from Equation (5.10) can be viewed as a continuous extension of the differencing matrix Q.

We can now proceed to solve Equation 5.7 to obtain $C(\mathbf{x}, \mathbf{x}')$ as the Green's function of L. To facilitate this, we take the Green's function of L in the whole space, which results in a stationary covariance function $C(\mathbf{x}, \mathbf{x}') = C(\mathbf{x} - \mathbf{x}')$. Making a change of variables $\mathbf{y} = \mathbf{x} - \mathbf{x}' = (y_x, y_z)$, and defining for notational convenience $\rho = \frac{1}{\lambda\beta}$ and $\kappa^2 = \frac{\epsilon}{\beta\ell^2}$ we have

$$\frac{1}{\rho}(\kappa^2 - \Delta) C(\mathbf{y}) = \delta(\mathbf{y}). \tag{5.20}$$

Taking the 2-D spatial Fourier transform of both sides, this becomes

$$\frac{1}{\rho}(\kappa^2 + k_x^2 + k_z^2)\hat{C}(\mathbf{k}) = 1, \tag{5.21}$$

where $\mathbf{k} = (k_x, k_z)$ is the wavenumber (or spatial frequency) and $\hat{C}(\mathbf{k})$ is the Fourier transform of $C(\mathbf{y})$, i.e. $\hat{C}(\mathbf{k})$ is the power spectral density of $\mathsf{m}(\mathbf{x})$. Rearranging and applying the inverse Fourier transform we have

$$C(\mathbf{y}) = \frac{\rho}{(2\pi)^2} \iint_{\mathbb{R}^2} \frac{1}{\kappa^2 + k_x^2 + k_z^2} e^{i\mathbf{k}\cdot\mathbf{y}} \,\mathrm{d}k_x \,\mathrm{d}k_z. \tag{5.22}$$

Applying a change of variables to polar coordinates

$$\begin{aligned} k_x &= k_r \cos\theta & \qquad\qquad y_x &= y_r \cos\phi \\ k_z &= k_r \sin\theta & \qquad\qquad y_z &= y_r \sin\phi \end{aligned}$$

we have

$$C(\mathbf{y}) = \frac{\rho}{(2\pi)^2} \int_0^\infty \frac{1}{\kappa^2 + k_r^2} \int_{-\pi}^{\pi} e^{ik_r y_r (\cos\theta\cos\phi + \sin\theta\sin\phi)} \,\mathrm{d}\theta \, k_r \,\mathrm{d}k_r \tag{5.23}$$

$$= \frac{\rho}{2\pi} \int_0^\infty \frac{1}{\kappa^2 + k_r^2} \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{ik_r y_r \cos(\theta - \phi)} \,\mathrm{d}\theta \, k_r \,\mathrm{d}k_r. \tag{5.24}$$

Making a final change of variables $\psi = \theta - \phi - \pi/2$, we recognize the inner integral as a Bessel function:

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} e^{ik_r y_r \cos(\theta - \phi)} \,\mathrm{d}\theta = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-ik_r y_r \sin\psi} \,\mathrm{d}\psi \tag{5.25}$$

$$= J_0(y_r k_r), \tag{5.26}$$

where J_0 denotes the zeroth-order Bessel function of the first kind. Then (5.24) becomes

$$C(\mathbf{y}) = \frac{\rho}{2\pi}\int_0^\infty \frac{1}{\kappa^2 + k_r^2} J_0(y_r k_r) k_r \,\mathrm{d}k_r \tag{5.27}$$

$$= \frac{\rho}{2\pi} K_0(\kappa y_r), \tag{5.28}$$

where we have recognized the integral in (5.27) as the zeroth-order Hankel transform of $\frac{1}{\kappa^2+k_r^2}$ and K_0 denotes the zeroth-order modified Bessel function of the second kind. Reverting to the initial choice of variables, we then have for the covariance function

$$C(\mathbf{x}, \mathbf{x}') = \frac{1}{2\pi\lambda\beta} K_0\left(\frac{\|\mathbf{x} - \mathbf{x}'\|_2}{\ell\sqrt{\beta/\epsilon}}\right). \tag{5.29}$$

This covariance function will be a valid approximation of the model covariance in the finite-dimensional setting when considering points far from the boundary of the grid (to avoid boundary effects, since the Green's function is taken in the whole-space) and when $\mathbf{x} \neq \mathbf{x}'$, since $K_0(t) \to \infty$ as $t \to 0$.

From (5.29), we observe that the correlation length ξ, defined here as the parameter governing the rate of decay of the covariance function, is determined by

$$\xi = \ell\sqrt{\beta/\epsilon}. \tag{5.30}$$

This can also be observed from the power spectral density of $\mathsf{m}(\mathbf{x})$ in (5.21): the modulus of the wavenumber at which the magnitude of the power spectrum drops to half of its peak value ($\hat{C}(0)$) should be proportional to $1/\xi$. We observe this at $\|\mathbf{k}\|_2 = \kappa = \sqrt{\epsilon/\beta\ell^2}$:

$$\hat{C}(\mathbf{k})\Big|_{\|\mathbf{k}\|_2=\kappa} = \frac{1}{2}\hat{C}(0). \tag{5.31}$$

Although the continuous model has infinite variance (since $C(\mathbf{x}, \mathbf{x}') \to \infty$ as $\mathbf{x} \to \mathbf{x}'$), we can still gain insights about the (finite) variance of the discretized model at a given grid cell from the power spectrum of the continuous model. From (5.21), we can see the value of $\hat{C}(0)$, the power spectrum at D.C., which is

$$\hat{C}(0) = \frac{\rho}{\kappa^2} = \frac{\ell^2}{\epsilon\lambda}. \tag{5.32}$$

We note that in the limit as $\beta \to 0$, the correlation length $\xi \to 0$ and the power spectrum becomes flat, corresponding to white noise with power spectrum $\frac{\ell^2}{\epsilon\lambda}$. In this case, the variance $\mathrm{var}(\mathsf{m}_i)$ of the corresponding discretized model at a grid cell (of length ℓ) would then be given by $\mathrm{var}(\mathsf{m}_i) = \frac{1}{\ell^2}\hat{C}(0) = \frac{1}{\epsilon\lambda}$. From (5.29), we also see that the parameter λ scales only the inverse variance (whereas β and ϵ play a role in determining both the variance and correlation length). Hence, in determining an appropriate choice for these parameters, one may choose to first set β and ϵ to achieve a desired correlation length, then pick λ to appropriately scale the model variance $\mathrm{var}(\mathsf{m}_i)$. The exact value of the model variance $\mathrm{var}(\mathsf{m}_i)$ can be determined numerically by computing $(Q^{-1})_{ii}$.

As a numerical verification of the validity of (5.29), in Figure 5-1, we compare the covariances given by the covariance function for the continuous model $\mathsf{m}(\mathbf{x})$ to the numerically computed covariances of the discretized model $\mathbf{m}$ on a 101-by-101 node grid of grid cell length $\ell = 1$ m. We set the parameters $\lambda = 1$, $\epsilon = 10^{-3}$, and varied β between 10^{-3} and 1. We see strong agreement between the covariances for the discrete and continuous cases when $\mathbf{x} \neq \mathbf{x}'$ for all values of β plotted other than $\beta = 1$. At $\beta = 1$, the correlation length of the covariance function is large enough that the boundary effects in the discrete case can no longer be neglected, and hence (5.29) becomes a less accurate approximation of the covariance of the discretized model. At $\mathbf{x} = \mathbf{x}'$, the covariance function goes to infinity, but the variance of the discretized model remains finite.

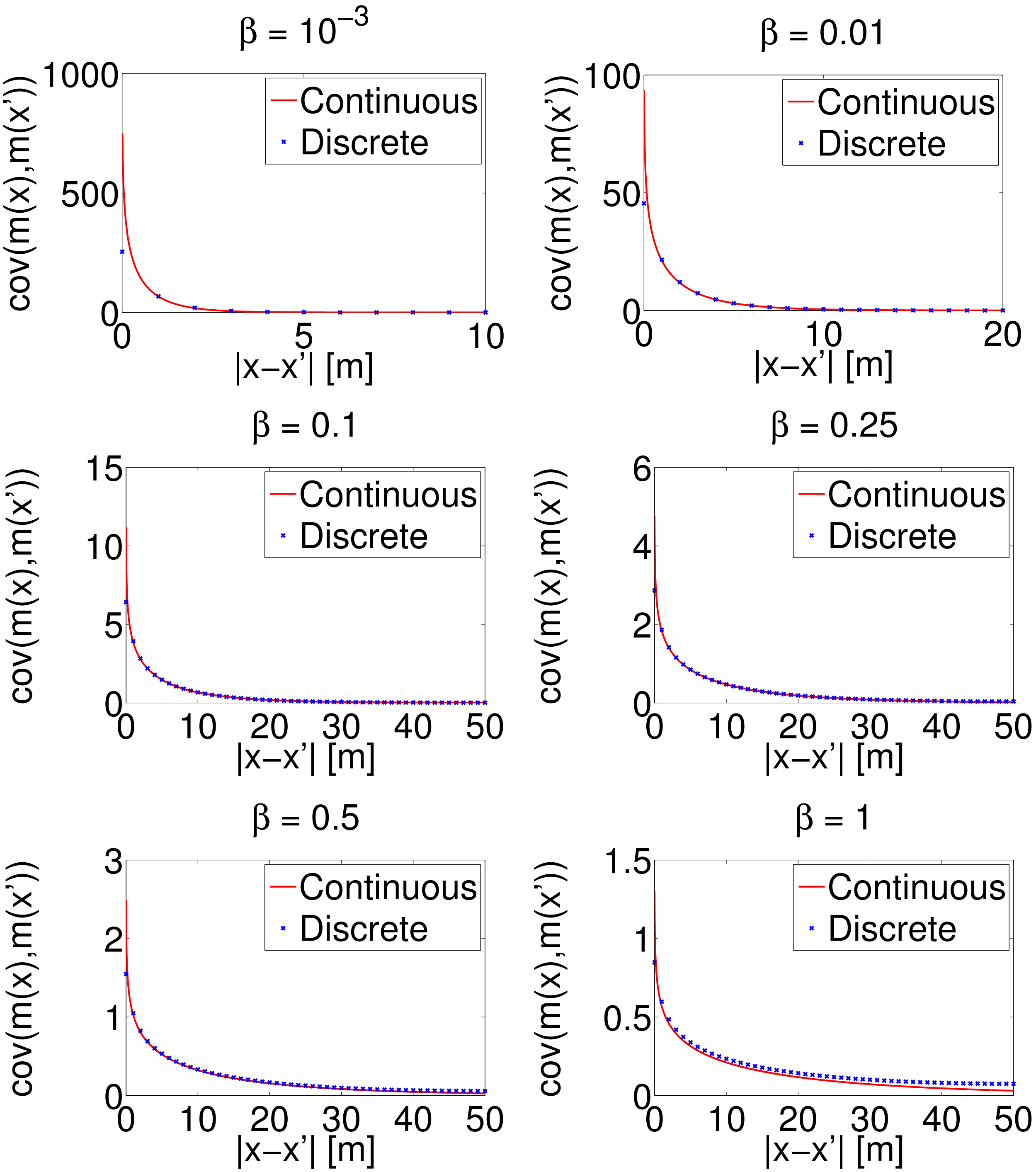


Figure 5-1: Comparison of numerically computed covariances in the discrete case to the covariance function in the continuous limit, with $\lambda = 1$, $\epsilon = 10^{-3}$, and for different values of β. Covariances for the discrete case were computed with the central node on a 101-by-101 node grid having grid cell length $\ell = 1$ m. (Note that the range of the x-axis is reduced for small values of β to make the plots more visible.)

5.2 Variational Bayesian Estimation of Regularization Parameters

Having described how the parameters λ, β, and ϵ characterize the prior model covariance, we now return to the problem of estimating the regularization parameters in a Bayesian inference setting. We note that there is some redundancy in our parameterization in regards to the effect of the parameters on the prior model covariance. In particular, the covariance function can be parameterized by two characteristic parameters, the correlation length $\xi = \ell\sqrt{\frac{\beta}{\epsilon}}$ and the power spectrum at D.C. $\hat{C}(0) = \frac{\ell^2}{\lambda\epsilon}$, whereas our parameterization provides three parameters that can be tuned. In order to remove this redundancy, and additionally recognizing the role ϵ plays in keeping the prior precision matrix Q positive-definite, we keep ϵ fixed while estimating the remaining regularization parameters. From (5.30), we observe that fixing ϵ (while restricting β to $[0,1]$) determines the maximum correlation length as $\xi_{\max} = \frac{\ell}{\sqrt{\epsilon}}$, hence we can set ϵ in accordance with a desired $\xi_{\max}$.

In the previous section, we set all the β_{ij} to a common value β in order to obtain a stationary covariance function. However, we will remove this restriction when attempting to estimate the β_{ij}, as in Chapter 4. In addition to estimating the parameters $\boldsymbol{\beta}$ and λ governing the prior distribution, we also estimate the inverse variance of the noise, which we denote by ζ.

5.2.1 Hierarchical Bayesian Formulation

We slightly redefine the probabilistic model of (5.1)-(5.3) to include λ and ζ as random variables in the hierarchical Bayesian framework. We define the prior distribution for the model $\mathbf{m}$ given the regularization parameters as before so that

$$p\left(\mathbf{m}|\boldsymbol{\beta},\lambda\right) \propto \frac{\lambda^{N/2}\exp\left\{-\frac{1}{2}\lambda\mathbf{m}^T(D\left(\boldsymbol{\beta}\right)+\epsilon I)\mathbf{m}\right\}}{\left|D\left(\boldsymbol{\beta}\right)+\epsilon I\right|^{-1/2}}, \tag{5.33}$$

where N is the number of image parameters and $|\cdot|$ denotes the matrix determinant. And we again let the data $\mathbf{d}$ be defined by

$$\mathbf{d} = A\mathbf{m} + \mathbf{n}, \tag{5.34}$$

where A is the forward modeling operator (in our case we take A to be the Kirchhoff modeling operator, so that A^T is the Kirchhoff migration operator, as defined in Chapter 4) and $\mathbf{n}$ is additive noise which we model as white Gaussian noise with zero mean and covariance matrix $\Sigma = \zeta^{-1}I$. Under these assumptions, the conditional distribution for the data is given by

$$p(\mathbf{d}|\mathbf{m}, \zeta, \boldsymbol{\beta}, \lambda) = p(\mathbf{d}|\mathbf{m}, \zeta) \propto \zeta^{\,K/2} \exp\left\{-\frac{1}{2}\zeta \left\|\mathbf{d} - A\mathbf{m}\right\|^2\right\}, \tag{5.35}$$

where K is the number of data points.

Letting $\boldsymbol{\theta} = (\boldsymbol{\beta}, \lambda, \zeta)$ denote the vector of regularization parameters (in which we include the inverse noise variance), we model $\boldsymbol{\theta}$ as a random vector with its own prior distribution $p(\boldsymbol{\theta})$. Since λ and ζ scale the inverse model and noise variances, respectively, we are able to introduce conjugate priors for these parameters; in particular, we model these parameters as Gamma random variables *a priori*, as the conjugate prior for the inverse variance parameter of a Gaussian is the Gamma distribution [24]. Hence, we have

$$p(\lambda) \propto \lambda^{a_\lambda - 1} e^{-b_\lambda \lambda} \qquad \lambda \geq 0 \tag{5.36}$$

and

$$p(\zeta) \propto \zeta^{a_\zeta - 1} e^{-b_\zeta \zeta} \qquad \zeta \geq 0, \tag{5.37}$$

where $a_\lambda, b_\lambda, a_\zeta, b_\zeta > 0$ are the shape and rate parameters of the Gamma distributions. These are called conjugate priors because the conditional posterior distributions for λ and ζ will remain Gamma distributions, only with updated shape and rate parameters. (We note that setting $a = 1$ and taking the limit as $b \to 0$ results in an (improper) flat prior on λ (or ζ) ≥ 0, so the Gamma priors are quite general.) Furthermore, as before, we endow each β_{ij} with a uniform prior on the set $[0, 1]$ and let

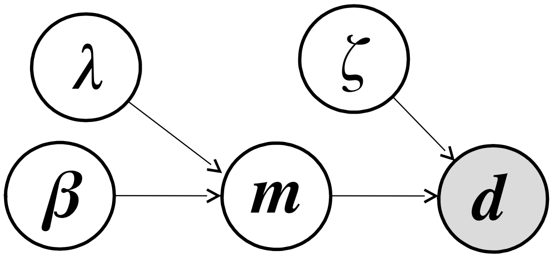


Figure 5-2: The directed graphical model capturing the Markov structure between $\boldsymbol{\beta}$, λ, ζ, $\mathbf{m}$, and $\mathbf{d}$. The node for $\mathbf{d}$ is shaded to indicate that $\mathbf{d}$ is an observed quantity that the posterior distribution for the regularization parameters and model is conditioned upon.

λ, ζ, and the β_{ij} be mutually independent random variables. Combining the above defines our prior distribution for $\boldsymbol{\theta}$. The directed graphical model in Figure 5-2 depicts the Markov structure between the regularization parameters, the model, and data. Having fully specified our probabilistic model, we can now apply Bayes' rule to obtain the joint posterior distribution for $\boldsymbol{\theta}$ and $\mathbf{m}$ given the data $\mathbf{d}$:

$$p\left(\mathbf{m},\boldsymbol{\theta}|\mathbf{d}\right) \propto p\left(\boldsymbol{\theta}\right) p\left(\mathbf{m}|\boldsymbol{\theta}\right) p\left(\mathbf{d}|\mathbf{m},\boldsymbol{\theta}\right) \tag{5.38}$$

$$\begin{aligned} &\propto \lambda^{a_\lambda+N/2-1}\,\zeta^{a_\zeta+K/2-1}\left|D\left(\boldsymbol{\beta}\right)+\epsilon I\right|^{1/2} \\ &\quad\exp\left\{-\lambda\left(b_\lambda+\frac{1}{2}\mathbf{m}^T(D\left(\boldsymbol{\beta}\right)+\epsilon I)\mathbf{m}\right)\right. \\ &\qquad\left.-\zeta\left(b_\zeta+\frac{1}{2}\left\|\mathbf{d}-A\mathbf{m}\right\|^2\right)\right\}. \end{aligned} \tag{5.39}$$

5.2.2 Variational Bayesian Methods

The complete solution to the hierarchical Bayesian problem would involve obtaining and tractably exploring, often via sampling techniques, the marginal posterior distributions for the parameters of interest $\mathbf{m}$. However, since the posterior distribution (5.39) is costly to evaluate (due to the determinant factor), sampling from the posterior quickly becomes infeasible for even moderately sized models ($N \sim 10^4$). To avert this problem, we turn to an approximate inference framework known as variational Bayes (VB) [3, 4], which can be viewed as a generalization of the E-M algorithm used in Chapter 4.

The idea behind the variational Bayesian method is to approximate an intractable posterior distribution $p(\mathbf{u}|\mathbf{d})$ (defined on the set of random variables $\mathbf{u} = (\mathbf{m}, \boldsymbol{\theta})$) by searching within a family of tractable distributions $q \in \mathcal{Q}$ for the distribution, q^*, which is closest to the posterior $p_{\mathbf{u}|\mathbf{d}}$. The approximate posterior q^* is found as the solution to the following variational problem (hence the name):

$$q^* = \operatorname*{arg\,min}_{q\in\mathcal{Q}}\ D(q\|p_{\mathbf{u}|\mathbf{d}}), \tag{5.40}$$

where $D(\cdot\|\cdot)$ is the KL divergence, a pseudo-metric on the space of probability distributions given by

$$D(q\|p_{\mathbf{u}|\mathbf{d}}) = \mathbb{E}_q \left[\log \left(\frac{q(\mathbf{u})}{p(\mathbf{u}|\mathbf{d})} \right) \right] , \tag{5.41}$$

and where $\mathbb{E}_q$ denotes the expectation operator with respect to the distribution q. As hinted at earlier, the expectation-maximization algorithm [17, 46] can be derived as a special case of VB when we restrict $\mathcal{Q}$ to the family of point distributions (i.e. Dirac delta distributions) on some subset of the random variables in $\mathbf{u}$ [3].

The most common variant of VB is known as the *mean field approximation*, in which $\mathcal{Q}$ comprises distributions which factorize over specified partitions on the set of random variables $\mathbf{u}$. In our problem, it is natural to take $\mathbf{m}$ and $\boldsymbol{\theta}$ as two partitions of the unknowns $\mathbf{u}$. Using this partitioning with the mean field approximation, VB will search within the family of distributions that, conditioned on $\mathbf{d}$, have $\mathbf{m}$ independent from $\boldsymbol{\theta}$ (which is not the case in the true posterior distribution (5.39)). We further specialize $\mathcal{Q}$ by restricting the class of distributions on $\boldsymbol{\beta}$ to point distributions (this will result in effectively estimating $\boldsymbol{\beta}$ via the E-M algorithm while estimating the remaining parameters in the more general mean field setting). We thus have

$$\begin{aligned} q(\mathbf{m}, \boldsymbol{\theta}) &= q_{\mathbf{m}}(\mathbf{m}) q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) && (5.42) \\ &= q_{\mathbf{m}}(\mathbf{m}) q_{\lambda,\zeta,\boldsymbol{\beta}}(\lambda, \zeta, \boldsymbol{\beta}) && (5.43) \\ &= q_{\mathbf{m}}(\mathbf{m}) q_{\lambda,\zeta}(\lambda, \zeta) \delta(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}), && (5.44) \end{aligned}$$

where $\bar{\boldsymbol{\beta}}$ is the point at which the delta distribution $q_{\boldsymbol{\beta}}(\boldsymbol{\beta})$ is centered (and hence the only parameter defining $q_{\boldsymbol{\beta}}(\boldsymbol{\beta})$).

Coordinate-Descent

To solve the variational problem (5.40) for $q \in \mathcal{Q}$, we substitute the above factorized form of q into the KL divergence and take a coordinate-descent approach, which alternates between minimizing the KL divergence with respect to $\bar{\beta}$ and $(q_{\mathbf{m}}, q_{\lambda,\zeta})$, giving the iterative procedure:

$$\bar{\boldsymbol{\beta}}^{(t+1)} = \arg\min_{\bar{\boldsymbol{\beta}}} D(q_{\mathbf{m}}^{(t)} q_{\lambda,\zeta}^{(t)} \delta_{\bar{\boldsymbol{\beta}}} \| p_{\lambda,\zeta,\boldsymbol{\beta},\mathbf{m}|\mathbf{d}}) \tag{5.45}$$

$$\left(q_{\mathbf{m}}^{(t+1)}, q_{\lambda,\zeta}^{(t+1)}\right) = \arg\min_{\left(q_{\mathbf{m}}, q_{\lambda,\zeta}\right)} D(q_{\mathbf{m}} q_{\lambda,\zeta} \delta_{\bar{\boldsymbol{\beta}}^{(t+1)}} \| p_{\lambda,\zeta,\boldsymbol{\beta},\mathbf{m}|\mathbf{d}}). \tag{5.46}$$

Examining the update equation for $\bar{\boldsymbol{\beta}}^{(t+1)}$ (5.45), and dropping terms that do not depend on $\bar{\beta}$, we see that this equation is essentially the M-step of the E-M algorithm:

$$\begin{aligned} \bar{\boldsymbol{\beta}}^{(t+1)} &= \arg\min_{\bar{\boldsymbol{\beta}}} D(q_{\mathbf{m}}^{(t)} q_{\lambda,\zeta}^{(t)} \delta_{\bar{\boldsymbol{\beta}}} \| p_{\lambda,\zeta,\boldsymbol{\beta},\mathbf{m}|\mathbf{d}}) & (5.47) \\ &= \arg\max_{\boldsymbol{\beta}} \mathbb{E}_{q_{\mathbf{m}}^{(t)} q_{\lambda,\zeta}^{(t)}} [\log p(\lambda, \zeta, \boldsymbol{\beta}, \mathbf{m}, \mathbf{d})]. & (5.48) \end{aligned}$$

And indeed, we can perform the above maximization using the same methodology as was used to solve the M-step in Chapter 4 via a gradient ascent method, where

$$\begin{aligned} \frac{\partial}{\partial \beta_{ij}} \mathbb{E}_{q_{\mathbf{m}}^{(t)} q_{\lambda,\zeta}^{(t)}} [\log p(\lambda, \zeta, \boldsymbol{\beta}, \mathbf{m}, \mathbf{d})] = \frac{\bar{\lambda}^{(t)}}{2} \Big(& C_{ii}(\bar{\lambda}^{(t)}, \boldsymbol{\beta}) + C_{jj}(\bar{\lambda}^{(t)}, \boldsymbol{\beta}) - 2C_{ij}(\bar{\lambda}^{(t)}, \boldsymbol{\beta}) \\ & -(\Lambda_{ii}^{(t)} + \Lambda_{jj}^{(t)} - 2\Lambda_{ij}^{(t)}) - (\mu_i^{(t)} - \mu_j^{(t)})^2 \Big), \end{aligned} \tag{5.49}$$

and where $\bar{\lambda}^{(t)} = \mathbb{E}_{q_\lambda^{(t)}}[\lambda]$ is the expected value of λ under $q_\lambda^{(t)}$, $\boldsymbol{\mu}^{(t)} = \mathbb{E}_{q_{\mathbf{m}}^{(t)}}[\mathbf{m}]$ and $\Lambda^{(t)} = \text{cov}_{q_{\mathbf{m}}^{(t)}}(\mathbf{m})$ are the mean vector and covariance matrix of $\mathbf{m}$ under $q_{\mathbf{m}}^{(t)}$, and $C(\bar{\lambda}^{(t)}, \boldsymbol{\beta}) = \text{cov}_{p_{\mathbf{m}|\bar{\lambda}^{(t)},\boldsymbol{\beta}}}(\mathbf{m})$ is the prior covariance matrix of $\mathbf{m}$ under $p_{\mathbf{m}|\bar{\lambda}^{(t)},\boldsymbol{\beta}}$:

$$C(\bar{\lambda}^{(t)}, \boldsymbol{\beta}) = \left(\bar{\lambda}^{(t)}(D(\boldsymbol{\beta}) + \epsilon I)\right)^{-1}. \tag{5.50}$$

Fixed-Point Updates for $q_{\lambda,\zeta}^{(t)}$ and $q_{\mathbf{m}}^{(t)}$

The update equation (5.46) for the distributions $q_{\lambda,\zeta}^{(t)}$ and $q_{\mathbf{m}}^{(t)}$ still requires finding the distributions that minimize the KL divergence (with $\bar{\boldsymbol{\beta}}$ now fixed to $\bar{\boldsymbol{\beta}}^{(t+1)}$). One can derive the stationarity conditions on $q_{\lambda,\zeta}^{(t+1)}$ and $q_{\mathbf{m}}^{(t+1)}$ by forming a Lagrangian $\boldsymbol{\mathcal{L}}$ (to account for normalization constraints) and setting the functional derivatives $\frac{\delta \mathcal{L}}{\delta q_{\mathbf{m}}(\mathbf{m})}$

and $\frac{\delta \mathcal{L}}{\delta q_{\lambda,\zeta}(\lambda,\zeta)}$ to 0. This gives the standard equations of the mean field approximation:

$$\log q_{\lambda,\zeta}^{(t+1)}(\lambda,\zeta) = \mathbb{E}_{q_{\mathbf{m}}^{(t+1)}}\left[\log p(\lambda,\zeta,\bar{\boldsymbol{\beta}}^{(t+1)},\mathbf{m},\mathbf{d})\right] - Z_{\lambda,\zeta}, \tag{5.51}$$

$$\log q_{\mathbf{m}}^{(t+1)}(\mathbf{m}) = \mathbb{E}_{q_{\lambda,\zeta}^{(t+1)}}\left[\log p(\lambda,\zeta,\bar{\boldsymbol{\beta}}^{(t+1)},\mathbf{m},\mathbf{d})\right] - Z_{\mathbf{m}}, \tag{5.52}$$

where $Z_{\mathbf{m}}$ and $Z_{\lambda,\zeta}$ are normalization constants. The cyclic dependence between Equations (5.51) and (5.52) induces a natural fixed-point algorithm in which we pick an initial guess $q_{\lambda,\zeta}^{(t+1,0)}(\lambda,\zeta)$ and $q_{\mathbf{m}}^{(t+1,0)}(\mathbf{m})$ and update these guesses by sequentially solving (5.51) and (5.52), repeatedly until convergence. Denoting the iteration number of this fixed-point algorithm by s, we have:

$$\log q_{\lambda,\zeta}^{(t+1,s+1)}(\lambda,\zeta) = \mathbb{E}_{q_{\mathbf{m}}^{(t+1,s)}}\left[\log p(\lambda,\zeta,\bar{\boldsymbol{\beta}}^{(t+1)},\mathbf{m},\mathbf{d})\right] - Z_{\lambda,\zeta} \tag{5.53}$$

$$\log q_{\mathbf{m}}^{(t+1,s+1)}(\mathbf{m}) = \mathbb{E}_{q_{\lambda,\zeta}^{(t+1,s+1)}}\left[\log p(\lambda,\zeta,\bar{\boldsymbol{\beta}}^{(t+1)},\mathbf{m},\mathbf{d})\right] - Z_{\mathbf{m}}. \tag{5.54}$$

In order to implement this fixed-point algorithm, we substitute the joint posterior distribution into (5.53) and (5.54). First substituting into (5.53) gives the form of the update for $q_{\lambda,\zeta}^{(t+1,s+1)}(\lambda,\zeta)$:

$$\begin{aligned}
\log q_{\lambda,\zeta}^{(t+1,s+1)}(\lambda,\zeta) &= (a_\lambda + N/2 - 1)\log\lambda + (a_\zeta + K/2 - 1)\log\zeta \\
&\quad - \lambda\left(b_\lambda + \tfrac{1}{2}\mathbb{E}_{q_{\mathbf{m}}^{(t+1,s)}}\left[\mathbf{m}^T(D(\bar{\boldsymbol{\beta}}^{(t+1)}) + \epsilon I)\mathbf{m}\right]\right) \qquad (5.55) \\
&\quad - \zeta\left(b_\zeta + \tfrac{1}{2}\mathbb{E}_{q_{\mathbf{m}}^{(t+1,s)}}\left[\|\mathbf{d} - A\mathbf{m}\|^2\right]\right) - Z'_{\lambda,\zeta} \\
&= (a_\lambda + N/2 - 1)\log\lambda \\
&\quad - \lambda\Big(b_\lambda + \tfrac{1}{2}\operatorname{tr}((D(\bar{\boldsymbol{\beta}}^{(t+1)}) + \epsilon I)\Lambda^{(t+1,s)}) \\
&\quad + \tfrac{1}{2}\boldsymbol{\mu}^{(t+1,s)^T}(D(\bar{\boldsymbol{\beta}}^{(t+1)}) + \epsilon I)\boldsymbol{\mu}^{(t+1,s)}\Big) - Z''_\lambda \qquad (5.56) \\
&\quad + (a_\zeta + K/2 - 1)\log\zeta \\
&\quad - \zeta\left(b_\zeta + \tfrac{1}{2}\operatorname{tr}\left(A^T A\Lambda^{(t+1,s)}\right) + \tfrac{1}{2}\|\mathbf{d} - A\boldsymbol{\mu}^{(t+1,s)}\|^2\right) - Z''_\zeta \\
&= \log q_\lambda^{(t+1,s+1)}(\lambda) + \log q_\zeta^{(t+1,s+1)}(\zeta), \qquad (5.57)
\end{aligned}$$

where we have recognized in the last equality that $q_{\lambda,\zeta}^{(t+1,s+1)}(\lambda,\zeta)$ factorizes (i.e.

$q_{\lambda,\zeta}^{(t+1,s+1)}$ models λ and ζ as independent). As before, $\boldsymbol{\mu}^{(t+1,s)}$ and $\Lambda^{(t+1,s)}$ denote the mean and covariance of $\mathbf{m}$ under $q_{\mathbf{m}}^{(t+1,s)}$, and $Z'_{\lambda,\zeta}$, Z''_{λ}, Z''_{ζ} are simply new normalization constants. From the form of Equation (5.56), we recognize $q_{\lambda}^{(t+1,s+1)}$ and $q_{\zeta}^{(t+1,s+1)}$ as Gamma distributions, and hence we only need to keep track of their shape and rate parameters, $a_{\lambda}^{(t+1,s+1)}$ and $b_{\lambda}^{(t+1,s+1)}$ for $q_{\lambda}^{(t+1,s+1)}$ and $a_{\zeta}^{(t+1,s+1)}$ and $b_{\zeta}^{(t+1,s+1)}$ for $q_{\zeta}^{(t+1,s+1)}$, which are given by:

$$a_{\lambda}^{(t+1,s+1)} = a_{\lambda} + N/2, \tag{5.58}$$

$$b_{\lambda}^{(t+1,s+1)} = b_{\lambda} + \tfrac{1}{2}\operatorname{tr}((D(\bar{\boldsymbol{\beta}}^{(t+1)}) + \epsilon I)\Lambda^{(t+1,s)}) + \tfrac{1}{2}\boldsymbol{\mu}^{(t+1,s)^T}(D(\bar{\boldsymbol{\beta}}^{(t+1)}) + \epsilon I)\boldsymbol{\mu}^{(t+1,s)}, \tag{5.59}$$

$$a_{\zeta}^{(t+1,s+1)} = a_{\zeta} + K/2, \tag{5.60}$$

$$b_{\zeta}^{(t+1,s+1)} = b_{\zeta} + \tfrac{1}{2}\operatorname{tr}\left(A^T A \Lambda^{(t+1,s)}\right) + \tfrac{1}{2}\|\mathbf{d} - A\boldsymbol{\mu}^{(t+1,s)}\|^2. \tag{5.61}$$

From these, we can compute the expected values of λ and ζ under $q_{\lambda,\zeta}^{(t+1,s+1)}$, denoted by $\bar{\lambda}^{(t+1,s+1)}$ and $\bar{\zeta}^{(t+1,s+1)}$, as

$$\bar{\lambda}^{(t+1,s+1)} = \frac{a_{\lambda}^{(t+1,s+1)}}{b_{\lambda}^{(t+1,s+1)}} \tag{5.62}$$

$$= \frac{a_{\lambda} + N/2}{b_{\lambda} + \frac{1}{2}\operatorname{tr}((D(\bar{\boldsymbol{\beta}}^{(t+1)}) + \epsilon I)\Lambda^{(t+1,s)}) + \frac{1}{2}\boldsymbol{\mu}^{(t+1,s)^T}(D(\bar{\boldsymbol{\beta}}^{(t+1)}) + \epsilon I)\boldsymbol{\mu}^{(t+1,s)}}, \tag{5.63}$$

$$\bar{\zeta}^{(t+1,s+1)} = \frac{a_{\zeta}^{(t+1,s+1)}}{b_{\zeta}^{(t+1,s+1)}} = \frac{a_{\zeta} + K/2}{b_{\zeta} + \frac{1}{2}\operatorname{tr}\left(A^T A \Lambda^{(t+1,s)}\right) + \frac{1}{2}\|\mathbf{d} - A\boldsymbol{\mu}^{(t+1,s)}\|^2}. \tag{5.64}$$

We similarly update the distribution $q_{\mathbf{m}}^{(t+1,s+1)}$ by substituting the joint posterior (5.39) into (5.54). This gives

$$\begin{aligned}\log q_{\mathbf{m}}^{(t+1,s+1)}(\mathbf{m}) = &-\frac{1}{2}\mathbf{m}^T\left(\bar{\zeta}^{(t+1,s+1)}A^T A + \bar{\lambda}^{(t+1,s+1)}(D(\bar{\boldsymbol{\beta}}^{(t+1)}) + \epsilon I)\right)\mathbf{m} \\ &+ \bar{\zeta}^{(t+1,s+1)}\mathbf{m}^T A^T \mathbf{d} - Z'_{\mathbf{m}}.\end{aligned} \tag{5.65}$$

From (5.65), we recognize $q_{\mathbf{m}}^{(t+1,s+1)}$ as a Gaussian distribution, and, as before, we need only keep track of the parameters defining the distribution: its mean vector $\boldsymbol{\mu}^{(t+1,s+1)}$ and covariance matrix $\Lambda^{(t+1,s+1)}$ which are given by

$$\mu^{(t+1,s+1)} = \left(\bar{\zeta}^{(t+1,s+1)} A^T A + \bar{\lambda}^{(t+1,s+1)} (D(\bar{\boldsymbol{\beta}}^{(t+1)}) + \epsilon I)\right)^{-1} \bar{\zeta}^{(t+1,s+1)} A^T \mathbf{d} \tag{5.66}$$

and

$$\Lambda^{(t+1,s+1)} = \left(\bar{\zeta}^{(t+1,s+1)} A^T A + \bar{\lambda}^{(t+1,s)} (D(\bar{\boldsymbol{\beta}}^{(t+1)}) + \epsilon I)\right)^{-1}. \tag{5.67}$$

Formulating our probabilistic model with conjugate priors, the Gamma prior for λ and ζ and the Gaussian prior for $\mathbf{m}$, caused the conditional posterior distributions for these parameters to remain Gamma and Gaussian, respectively. For this reason, the variational Bayesian approximations also remain as Gamma and Gaussian distributions and are hence tractable, as only the parameters defining these distributions need to be updated.

It is important to point out that we do *not* have to store or compute the entire covariance matrix $\Lambda^{(t,s)}$ to update $\bar{\boldsymbol{\beta}}^{(t)}$, $\lambda^{(t,s)}$, and $\zeta^{(t,s)}$. The partial derivatives in Equation (5.49) used to update $\bar{\boldsymbol{\beta}}^{(t)}$ only require elements of the covariance matrix corresponding to the edges of the graph $\mathcal{G}$, and these can be computed using the techniques outlined in Chapter 4. Updating $\lambda^{(t,s)}$ and $\zeta^{(t,s)}$ requires computing matrix traces of the product of $\Lambda^{(t,s)}$ and another matrix. These traces can be computed stochastically using a trace estimation algorithm due to Hutchinson [32]:

Algorithm 5.1 Hutchinson Trace Estimation Algorithm [32]

```
Let $H$ be an $N$-by-$N$ matrix. Specify number of samples $M$.
1. Draw $M$ i.i.d. white noise vector samples: $\mathbf{u}^{(i)} \sim \text{Uniform}(\{-1,1\}^N)$.
2. Estimate $\text{tr}(H) \approx \frac{1}{M}\sum_{i=1}^{M} {\mathbf{u}^{(i)}}^T H \mathbf{u}^{(i)}$.
```

It is straightforward to show that this estimate converges to the true matrix trace.

The estimator of $\mathrm{tr}(H)$ converges to the true mean of $\mathbf{u}^T H\mathbf{u}$ which is

$$\begin{aligned}\mathbb{E}[\mathbf{u}^T H\mathbf{u}] &= \mathbb{E}[\mathrm{tr}(\mathbf{u}^T H\mathbf{u})] && (5.68)\\ &= \mathrm{tr}(H\mathbb{E}[\mathbf{u}\mathbf{u}^T]) && (5.69)\\ &= \mathrm{tr}(H), && (5.70)\end{aligned}$$

since $\mathbb{E}[\mathbf{u}\mathbf{u}^T] = I$. Indeed, this algorithm will work with any zero-mean, unit-variance white noise vector $\mathbf{u}$ (such as, e.g. $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, I)$). It is important to note that we neither need to compute nor store the matrix H explicitly to estimate its trace with this algorithm; we only need to be able to apply H to a vector. For our case, H is the product of Λ and another matrix (which we generically denote by R), where Λ is the inverse of a posterior precision matrix. Then the action of $H = R\Lambda$ on a vector $\mathbf{u}$ can be computed by first solving the system $\Lambda^{-1}\mathbf{v} = \mathbf{u}$ for $\mathbf{v}$, then taking $R\Lambda\mathbf{u} = R\mathbf{v}$.

Summary of VB algorithm

To summarize our approach, we give our variational Bayesian algorithm for jointly estimating model and regularization parameters below:

Algorithm 5.2 Variational Bayesian algorithm for estimating $\boldsymbol{\beta}^*$ and approximate marginal posteriors q^*_λ, q^*_ζ, $q^*_{\mathbf{m}}$

```
Initialize $\bar{\boldsymbol{\beta}}^{(0)}$,$\bar{\lambda}^{(0,0)}$,$\bar{\zeta}^{(0,0)}$.
Set $t = 0$. Iterate on $t$:
1. Set $s = 0$. Iterate on $s$:
    a. Compute $\boldsymbol{\mu}^{(t,s)}$ via (5.66).
    b. Compute matrix traces in (5.63) and (5.64) via Hutchinson algorithm
        with $\Lambda^{(t,s)}$ as defined in (5.67).
    c. Compute $\bar{\lambda}^{(t,s+1)}$ via (5.63).
    d. Compute $\bar{\zeta}^{(t,s+1)}$ via (5.64).
2. Update $\bar{\lambda}^{(t+1)} = \bar{\lambda}^{(t,s+1)}$, $\bar{\zeta}^{(t+1)} = \bar{\zeta}^{(t,s+1)}$, $\boldsymbol{\mu}^{(t+1)} = \boldsymbol{\mu}^{(t,s)}$.
3. Update $\Lambda^{(t+1)}$ to be defined as $\Lambda^{(t,s)}$ in (5.67).
4. Update $\bar{\boldsymbol{\beta}}^{(t+1)}$ via the M-step of the E-M algorithm of Chapter 4
    (i.e. Steps 3-5 of Alg. 4.2 using $\bar{\lambda}^{(t+1)}$, $\bar{\zeta}^{(t+1)}$, $\boldsymbol{\mu}^{(t+1)}$, $\Lambda^{(t+1)}$).

Upon termination, return:
    $\bar{\lambda}^* = \bar{\lambda}^{(t+1)}$, $\bar{\zeta}^* = \bar{\zeta}^{(t+1)}$,
    $a^*_\lambda = a_\lambda + N/2$, $b^*_\lambda = \frac{a_\lambda + N/2}{\bar{\lambda}^*}$, $a^*_\zeta = a_\zeta + K/2$, $b^*_\zeta = \frac{a_\zeta + K/2}{\bar{\zeta}^*}$,
    $\boldsymbol{\beta}^* = \bar{\boldsymbol{\beta}}^{(t+1)}$, $\boldsymbol{\mu}^* = \boldsymbol{\mu}^{(t+1)}$, $\Lambda^* = \Lambda^{(t+1)}$,
    $q^*_\lambda = \text{Gamma}(a^*_\lambda, b^*_\lambda)$, $q^*_\zeta = \text{Gamma}(a^*_\zeta, b^*_\zeta)$, $q^*_{\mathbf{m}} = \mathcal{N}(\boldsymbol{\mu}^*, \Lambda^*)$.
```

5.2.3 Results

To obtain a single estimate of the image, we take the MAP estimate of the image from its approximate marginal posterior distribution $q^*_{\mathbf{m}}$, which we refer to as its variational Bayes MAP (or VB-MAP) estimate. We observe from (5.65) that $q^*_{\mathbf{m}}(\mathbf{m})$ happens to be equal to the conditional posterior $p(\mathbf{m}|\mathbf{d}, \boldsymbol{\beta}^*, \bar{\lambda}^*, \bar{\zeta}^*)$ using the $\boldsymbol{\beta}^*, \bar{\lambda}^*, \bar{\zeta}^*$ obtained from the VB algorithm. Hence the VB-MAP estimate of the image is equivalent to its empirical Bayes MAP estimate discussed in Chapter 4 when using the parameters $\boldsymbol{\beta}^*, \bar{\lambda}^*, \bar{\zeta}^*$.

We validate our approach by applying our inference algorithm to synthetic datasets arising from 2-D reflectivity models. We consider data simulated from two different physical models: Kirchhoff modeling and acoustic wave-equation modeling. As in Chapter 4, the purpose of using Kirchhoff modeled synthetic data is to test the performance of our algorithm when used with a consistent forward model. The wave-equation modeled data, on the other hand, additionally tests the ability of the Kirchhoff forward operator used in our algorithm to correctly model acoustic data. For both cases, we defined the MRF so that each node shares an edge with all other nodes within a radius of $\sqrt{2}$ nodes (to include diagonal connections in the MRF).

Kirchhoff modeled data

As in the previous chapter, the Kirchhoff data example is simulated from the Marmousi model. The data are acquired from a set of 20 collocated surface sources and receivers (resulting in 400 traces), with a 480 m spacing between receivers (and sources), where the source wavelet is a 25 Hz Ricker wavelet. As before, we add zero-mean white Gaussian noise to the data (with a standard deviation equal to 10% of the maximum amplitude of the data). Note that this *inverse crime* test example is the same as that in Chapter 4. In this example, the VB estimates of the marginal posteriors were obtained with 3 (outer) t-iterations (to update $\boldsymbol{\beta}$) and 10 (inner) s-iterations (to update q_λ,q_ζ,$q_{\mathbf{m}}$) of the VB algorithm (see Algorithm 5.2), where each outer t-iteration of VB took approximately 3 hours on a quad core Intel$^{\text{TM}}$ Xeon W3550 3.0GHz processor (hence the entire VB algorithm ran in approximately 9 hours). We note that each t-iteration of the VB algorithm performs all 10 s-iterations, where each s-iteration requires performing multiple LSM inversions to both compute the mean model parameters for updating $q_{\mathbf{m}}$ and to estimate the traces required to update q_λ and q_ζ via the Hutchinson algorithm. As in the previous chapter, the LSM inversions are computed using the method of conjugate gradients (CG), which is run for 200 iterations per LSM. Each iteration of CG requires performing a single Kirchhoff modeling followed by a single Kirchhoff migration, and, for this example, takes about 0.7 seconds per CG iteration (on the same machine).

The imaging results are shown in Figures 5-3–5-10, where the edge strengths obtained using our algorithm are shown in Figure 5-10 and the VB-MAP estimate of the image is shown in Figure 5-9. For comparison, we show the unregularized LSM image (Figure 5-4) along with the regularized LSM images obtained using a uniform regularization scheme (setting each $\beta_{ij} = 1$) and different values for λ (fixing ζ to $\bar{\zeta}^*$) (Figures 5-5–5-8).

As before, the unregularized image is heavily influenced by the noise in the data to the point that the reflectors in the image are largely obscured by the noise. The effects of noise are largely diminished in the uniformly regularized LSM images obtained with setting λ to $\bar{\lambda}^*$ or higher (Figures 5-7–5-8), but these same images significantly smooth out the reflectors. This is consistent with our expectations: $\bar{\lambda}^*$ is estimated in conjunction with the β_{ij}^*, which are always between 0 and 1. Hence, using this $\bar{\lambda}^*$ to regularize while setting all the β_{ij} to 1, should result in over-smoothing. Using values of λ smaller than $\bar{\lambda}^*$ (with $\beta_{ij} = 1$) prevents over-smoothing, but fails to significantly remove the effects of noise (Figures 5-5–5-6). The VB-MAP image (Figure 5-9), by contrast, remains sharp near the reflectors while smoothing out the noise away from the reflectors. For a quantitative comparison, we note that the correlation of the VB-MAP image with the true image is significantly higher than the correlations of the other images with the true image.

Figure 5-11 shows the VB approximations to the posterior distributions for λ and ζ. We can interpret these approximate posteriors in terms of the quantities λ and ζ represent. We find that the approximate posterior $q_\zeta^*(\zeta)$ slightly overestimates the true inverse noise variance (which is $\zeta = 5.61 \cdot 10^8$ Pa^{-2}). This is somewhat expected, since the VB-MAP model may provide a better fit to the data than the true model (and hence give a lower estimated noise variance). Indeed, the inverse variance of the data residual resulting from the VB-MAP model is $5.81 \cdot 10^8$ Pa^{-2}, which is much closer to $\bar{\zeta}^* = 5.79 \cdot 10^8$ Pa^{-2} (the mean of q_ζ^*). As discussed in the beginning of this chapter, λ scales the prior model variance. Hence, to interpret the value we obtain for $\bar{\lambda}^*$ (the mean of q_λ^*), we can measure the empirical variance of the true model and compare this to the average prior model variance predicted by $\bar{\lambda}^*$ (which can be

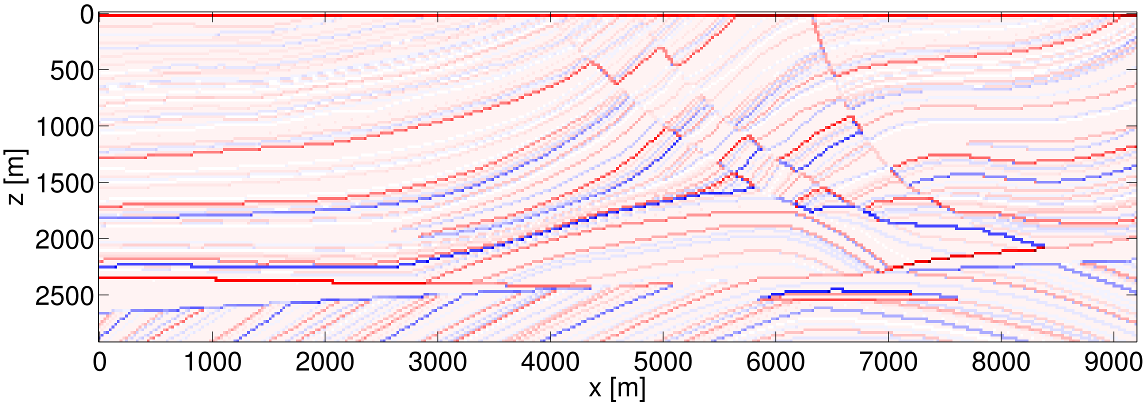


Figure 5-3: The true Marmousi reflectivity model.

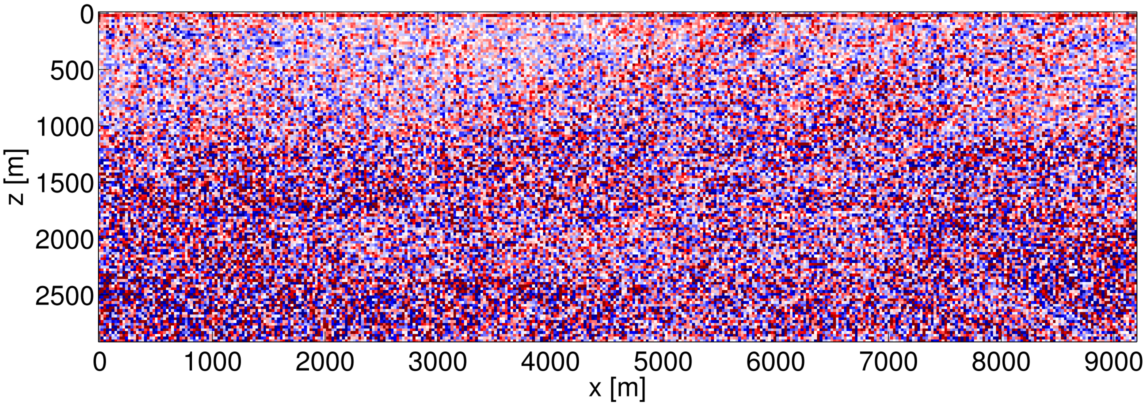


Figure 5-4: Unregularized LSM image (each $\beta_{ij} = 0$) using Kirchhoff modeled data. Correlation with true image = 0.3682.

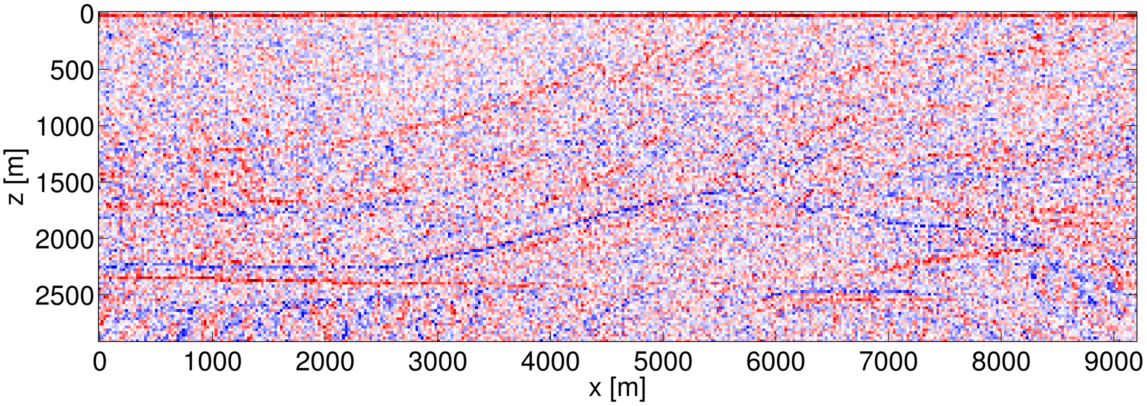


Figure 5-5: Uniformly regularized LSM image (each $\beta_{ij} = 1$) with $\lambda = 0.01\bar{\lambda}^*$ and $\zeta = \bar{\zeta}^*$ using Kirchhoff modeled data. Correlation with true image = 0.4364.

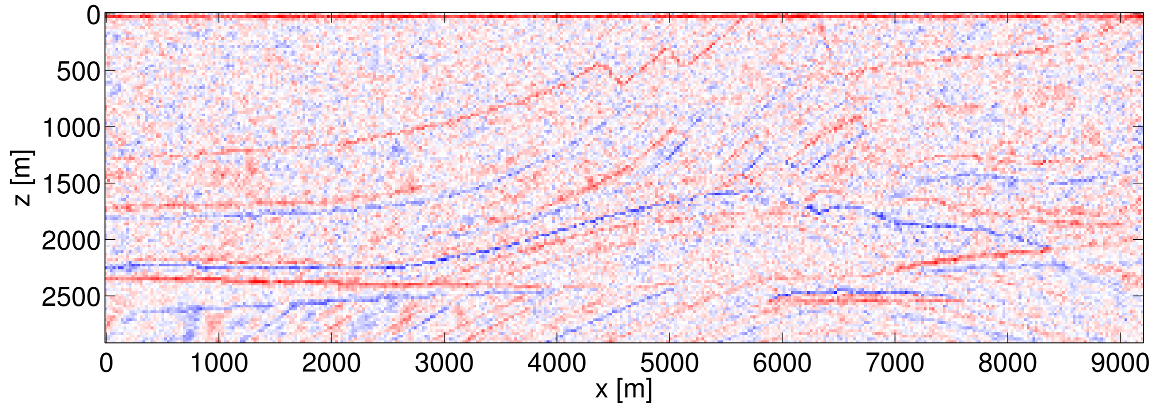


Figure 5-6: Uniformly regularized LSM image (each $\beta_{ij} = 1$) with $\lambda = 0.1\bar{\lambda}^*$ and $\zeta = \bar{\zeta}^*$ using Kirchhoff modeled data. Correlation with true image = 0.6282.

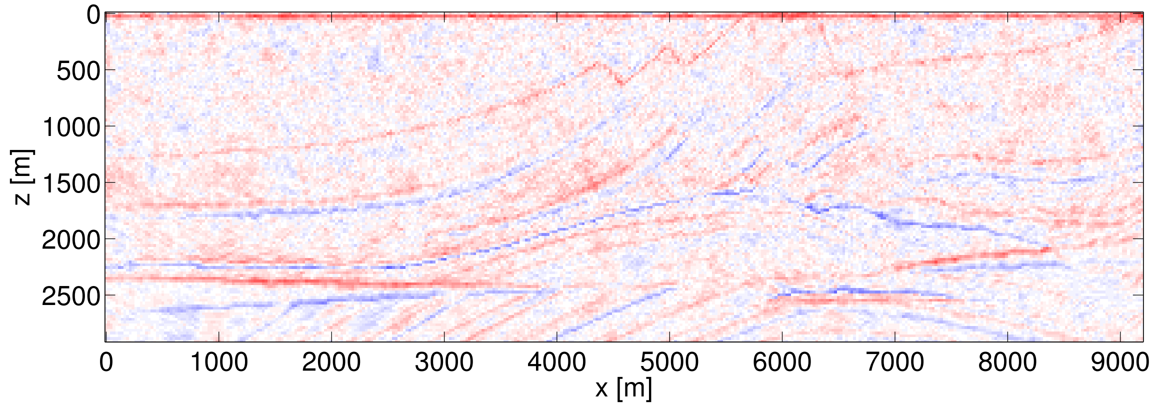


Figure 5-7: Uniformly regularized LSM image (each $\beta_{ij} = 1$) with $\lambda = \bar{\lambda}^*$ and $\zeta = \bar{\zeta}^*$ using Kirchhoff modeled data. Correlation with true image = 0.6432.

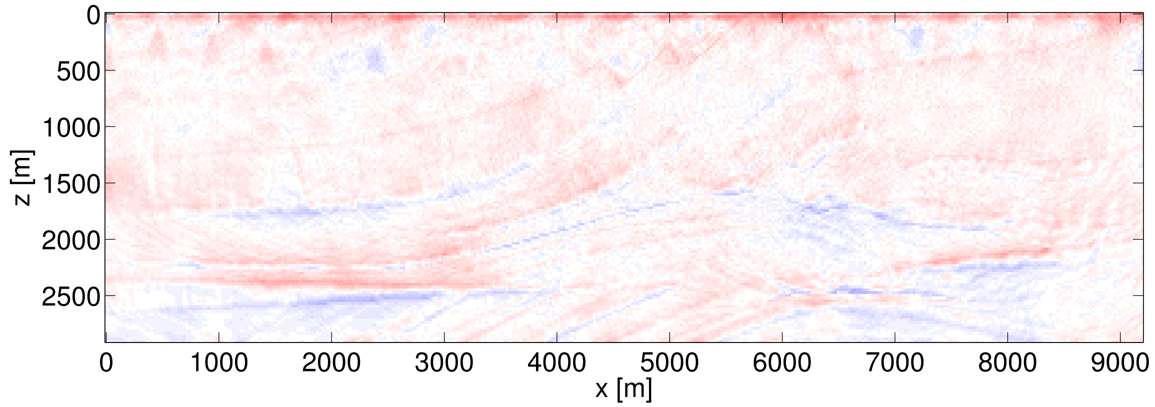


Figure 5-8: Uniformly regularized LSM image (each $\beta_{ij} = 1$) with $\lambda = 10\bar{\lambda}^*$ and $\zeta = \bar{\zeta}^*$ using Kirchhoff modeled data. Correlation with true image = 0.4870.

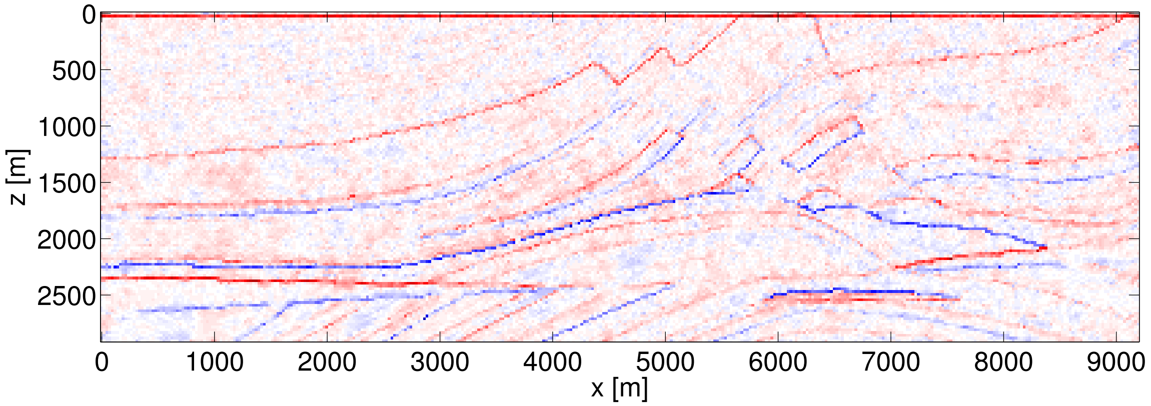


Figure 5-9: Variational Bayes MAP image using Kirchhoff modeled data. Correlation with true image = 0.8218.

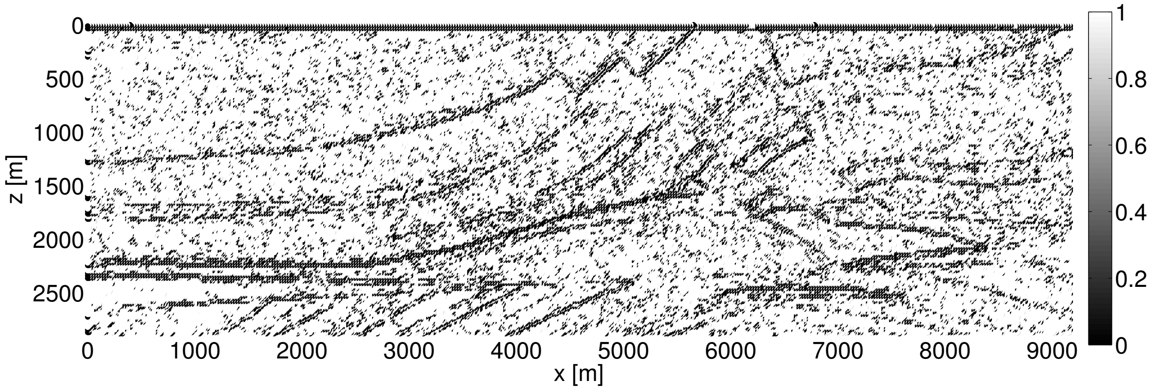


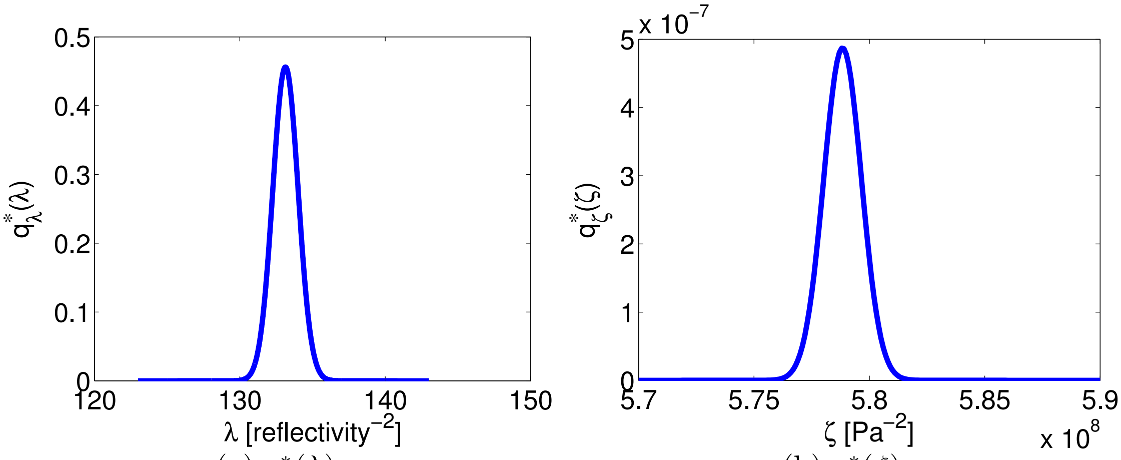
Figure 5-10: Edge strengths $\boldsymbol{\beta}^*$ estimated with VB using Kirchhoff modeled data.

computed from the trace of the prior model covariance matrix defined by $\bar{\lambda}^*$ and β^*). We find that while $\bar{\lambda}^* = 133.14$ (note that the reflectivity model is dimensionless, and hence so is λ), which yields an average prior model variance of $5.90 \cdot 10^{-3}$, the empirical variance of the true model is computed to be $5.66 \cdot 10^{-3}$. Hence, the value of $\bar{\lambda}^*$ computed by VB slightly overestimates the true model variance.

Acoustic wave-equation modeled data

Since the wave-equation modeled data arises from a more complex physical model than the Kirchhoff modeling operator used by our algorithm, we first examine the behavior of our algorithm on the example of a simple three layer model before moving on to the more complex Marmousi example.

For both examples, the synthetic data are acquired from a set of collocated surface sources and receivers, with an even spacing of 240 m, where the source wavelet is a 25 Hz Ricker wavelet. We note that, due to the complex nature of this data, we used smaller source and receiver spacings than what was used for the Kirchhoff modeled data (where for that data the source and receiver spacing was 480 m). Even still, this receiver spacing is far larger (and the dataset far sparser) than what is typically used in Kirchhoff migration applications; hence the choice of regularization becomes extremely significant in this case. We did not add random noise to the data, since the residual between the predicted data (by Kirchhoff modeling) and the wave-equation data already serves as a significant source of noise. For these examples, the VB estimates of the marginal posteriors were again obtained with 3 (outer) t-iterations (to update $\boldsymbol{\beta}$) and 10 (inner) s-iterations (to update q_λ,q_ζ,$q_{\mathbf{m}}$) of the VB algorithm (see Algorithm 5.2), where each outer t-iteration of VB took approximately 12 hours on a quad core Intel$^{\text{TM}}$ Xeon W3550 3.0GHz processor (hence the entire VB algorithm ran in approximately 36 hours). Again, the LSM inversions are computed using the method of conjugate gradients (CG), which is run for 200 iterations per LSM. Due to the increased number of sources and receivers, each iteration of CG costs approximately 2.8 seconds (on the same machine).



(a) $q^*_\lambda(\lambda)$ (b) $q^*_\zeta(\zeta)$

Figure 5-11: The variational Bayesian approximations to the posterior distributions for the parameters scaling the inverse variances of the (a) model λ and (b) noise ζ, using Kirchhoff modeled data.

Three layer example

The results for the three layer example are shown in Figures 5-12–5-22. The true reflectivity model is depicted in Figure 5-12, and again we show the unregularized LSM image (Figure 5-13) along with the uniformly regularized images for different values of λ (Figures 5-14–5-17). The VB-MAP estimate of the image is shown in Figure 5-18, and the corresponding edge strengths β^* are shown in Figure 5-19. As expected, the unregularized image contains a heavy amount of noise due to both the limited acquisition and poor modeling effects. The regularized LSM images (obtained with $\beta_{ij} = 1$) significantly improve upon this result. We find that while a significant amount of noise remains in the image obtained with $\lambda = \bar{\lambda}^*$, the noise is greatly reduced when a value of $\lambda = 10\bar{\lambda}^*$ is used in the regularization (Figure 5-18). This is a somewhat surprising result because, although $\bar{\lambda}^*$ was estimated in conjunction with $\boldsymbol{\beta}^*$, we would expect that, since β_{ij}^* is always between 0 and 1, regularizing with $\bar{\lambda}^*$ and $\bar{\zeta}^*$ and setting all the β_{ij} to 1 will result in over-smoothing the image. Indeed, for the Kirchhoff modeled data example, this choice of parameters did result in over-smoothing in the regularized LSM image (Figure 5-7). Hence, from a purely "results-oriented" perspective, one would expect a higher value of $\bar{\lambda}^*$ than that obtained.

This is also seen in the VB-MAP image (Figure 5-18), where the noise has only been slightly reduced from the unregularized LSM image, once again suggesting that the value of $\bar{\lambda}^*$ is too low. The estimated edge strengths (Figure 5-19), however, are aligned with our expectations. We see that the edge strengths go to 0 above and below the reflectors, while mostly remaining near 1 elsewhere in the image (however, we do also see the effects of the noise on the edge strengths). Hence, using these edge strengths along with a higher value for λ should hopefully give a better image. As noted earlier, the resulting image is the empirical Bayes MAP estimate of the image with λ fixed to a higher value. We experiment with successively higher values of λ (fixing $\boldsymbol{\beta} = \boldsymbol{\beta}^*$ and $\zeta = \bar{\zeta}^*$) in Figures 5-20, 5-21, and 5-22. Indeed, we see that as λ is made larger, the effects of the noise are greatly reduced, but the edge strengths β^* preserve the sharpness at the reflectors. At $\lambda = 10^3\bar{\lambda}^*$ (Figure 5-22), the greatest improvement is seen, and the correlation is highest for this image.

The VB approximations to the posterior marginals for λ and ζ are shown in Figure 5-23. Here we find that VB estimates the inverse noise variance at $\bar{\zeta}^* = 2.15 \cdot 10^{11}$ Pa^{-2}, whereas the inverse variance of the data residuals found with the true model and MAP-VB model are $\zeta = 1.14 \cdot 10^{11}$ Pa^{-2} and $\zeta = 2.16 \cdot 10^{11}$ Pa^{-2}, respectively. Hence, as expected, the estimated inverse noise variance matches well with the inverse variance of the data residual computed with the VB-MAP model (but the "true" inverse noise variance has been significantly overestimated). The VB estimate of λ was found to be $\bar{\lambda}^* = 4.05 \cdot 10^3$, which results in an average prior model variance of $7.66 \cdot 10^{-4}$ (again, found by computing the trace of the prior model covariance matrix given by $\bar{\lambda}^*$ and $\boldsymbol{\beta}^*$). By comparison, the empirical variance of the true model is $2.64 \cdot 10^{-4}$. Hence the true model variance is, as our imaging results seemed to indicate, significantly overestimated (meaning λ is significantly underestimated).

Marmousi model example

The second example we consider is the Marmousi model, where the wave-equation data are generated using the same acquisition geometry as in the three layer example. The complex velocity structure of the Marmousi model makes this example particularly challenging for Kirchhoff-based methods (including LSM), and often a wave-equation based imaging method, such as reverse-time migration (RTM) or RTM-based LSM, is required to correctly image deeper sections of the model [25]. We do not expect that the hierarchical Bayesian version of Kirchhoff-based LSM will be able to remedy this problem, since this is, at heart, a modeling issue and not an inversion issue. Nevertheless, the reflectors in the shallow part of the model can still be described by Kirchhoff modeling, and it is of interest to determine the improvements we might be able to gain with our methodology.

The results for the Marmousi example are shown in Figures 5-24–5-33. The unregularized LSM image is shown in Figure 5-24, and the uniformly regularized LSM images for different values of λ are displayed in Figures 5-25–5-28. Once again, we observe a similar issue with the value estimated for λ^*. The noisy unregularized LSM

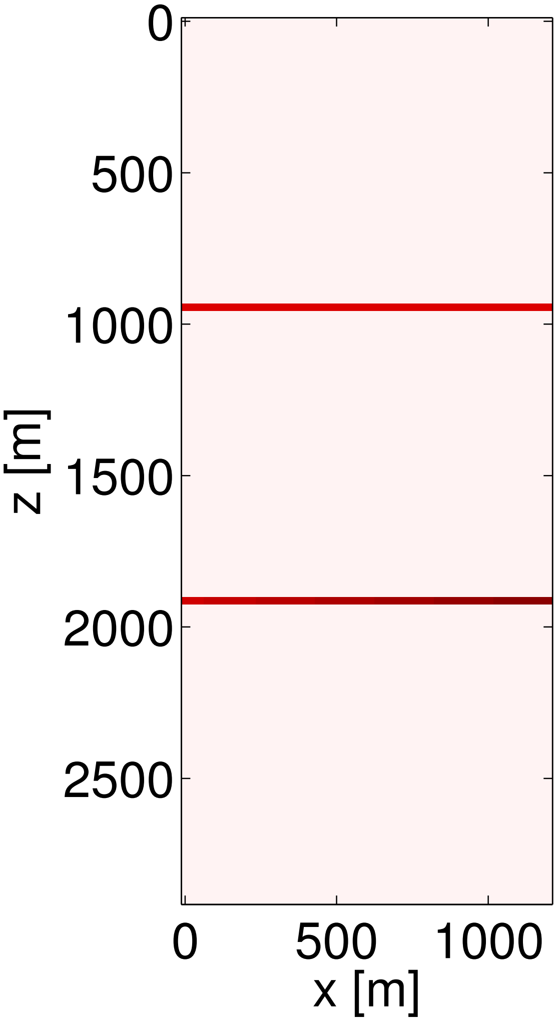


Figure 5-12: The true reflectivity for the three layer model.

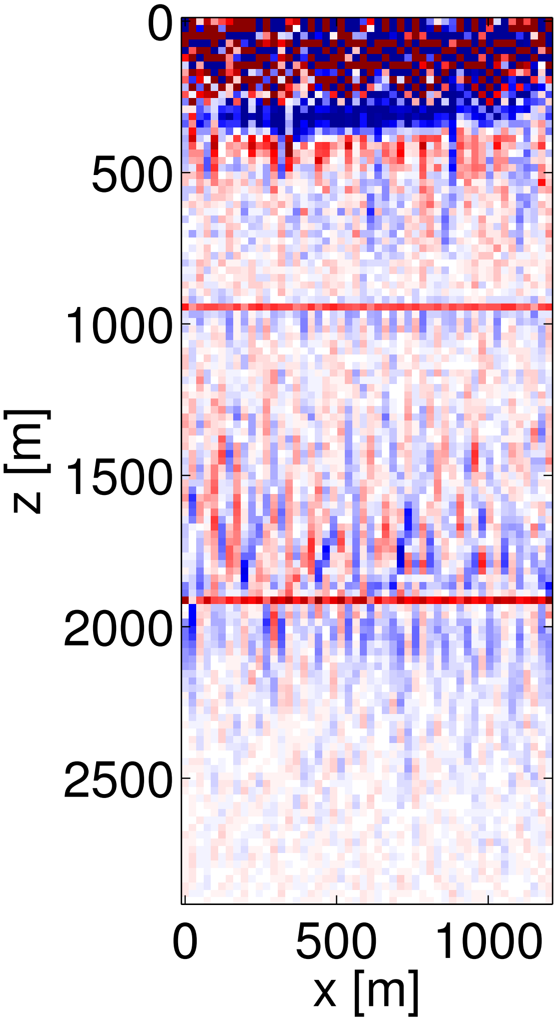


Figure 5-13: Unregularized LSM image (each $\beta_{ij} = 0$) using wave-equation modeled data (three layer model). Correlation with true image = 0.0824.

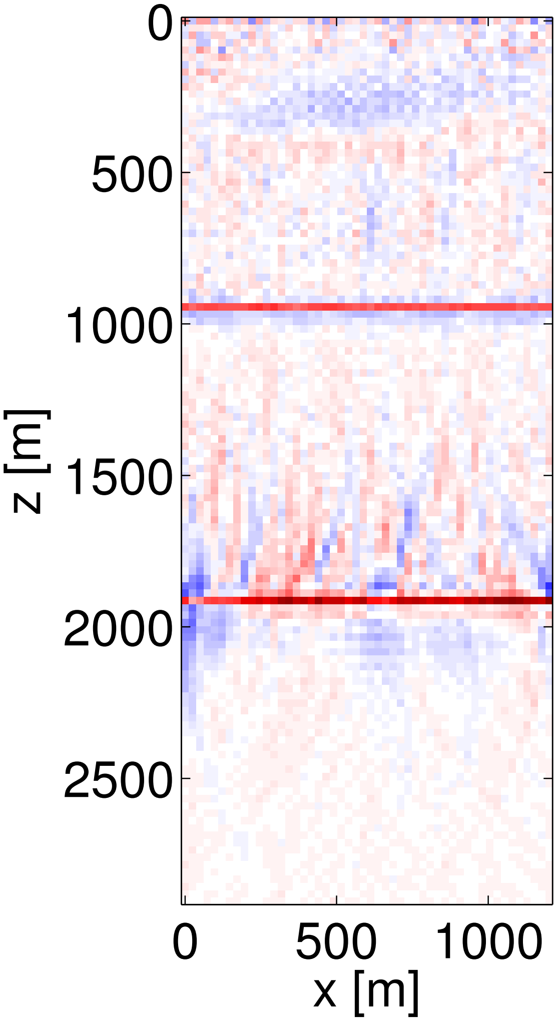


Figure 5-14: Uniformly regularized LSM image (each $\beta_{ij} = 1$) with $\lambda = \bar{\lambda}^*$ and $\zeta = \bar{\zeta}^*$ using wave-equation modeled data (three layer model). Correlation with true image $= 0.7226$.

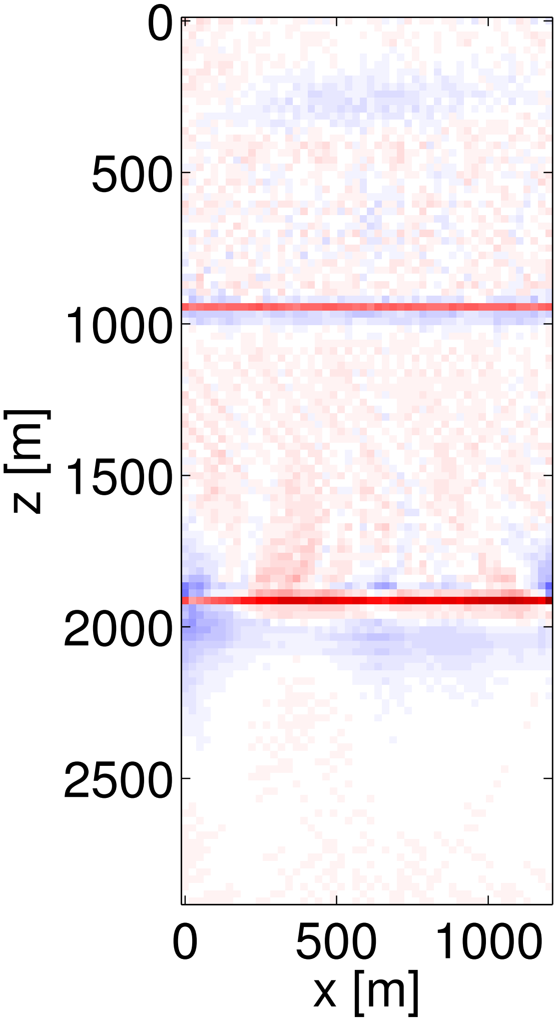


Figure 5-15: Uniformly regularized LSM image (each $\beta_{ij} = 1$) with $\lambda = 10\bar{\lambda}^*$ and $\zeta = \bar{\zeta}^*$ using wave-equation modeled data (three layer model). Correlation with true image $= 0.8309$.

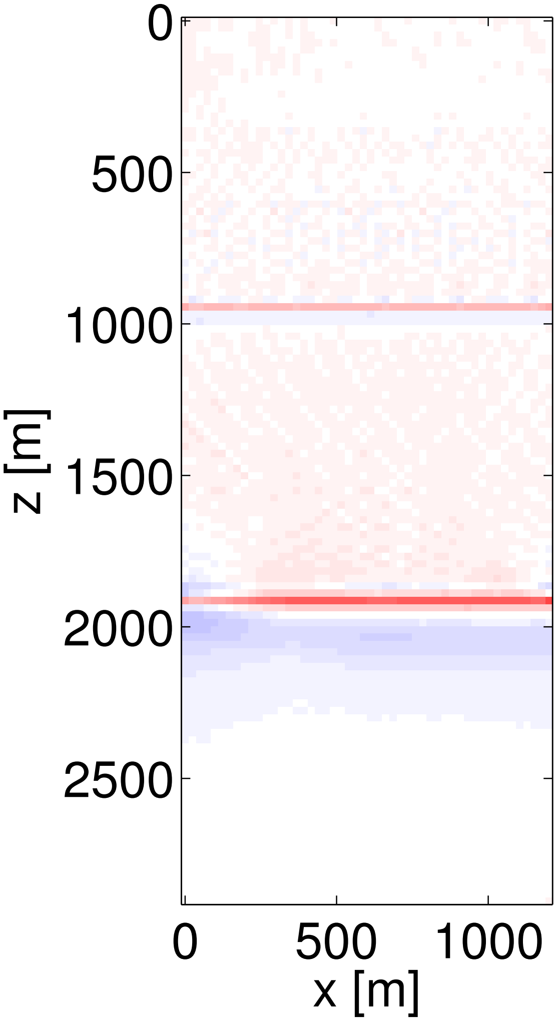


Figure 5-16: Uniformly regularized LSM image (each $\beta_{ij} = 1$) with $\lambda = 10^2\bar{\lambda}^*$ and $\zeta = \bar{\zeta}^*$ using wave-equation modeled data (three layer model). Correlation with true image = 0.7381.

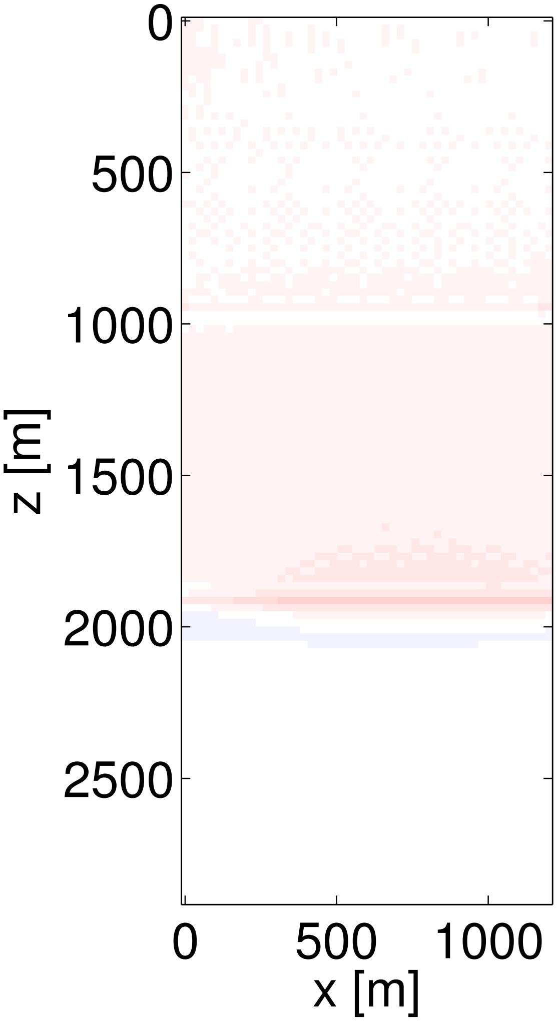


Figure 5-17: Uniformly regularized LSM image (each $\beta_{ij} = 1$) with $\lambda = 10^3\bar{\lambda}^*$ and $\zeta = \bar{\zeta}^*$ using wave-equation modeled data (three layer model). Correlation with true image = 0.4922.

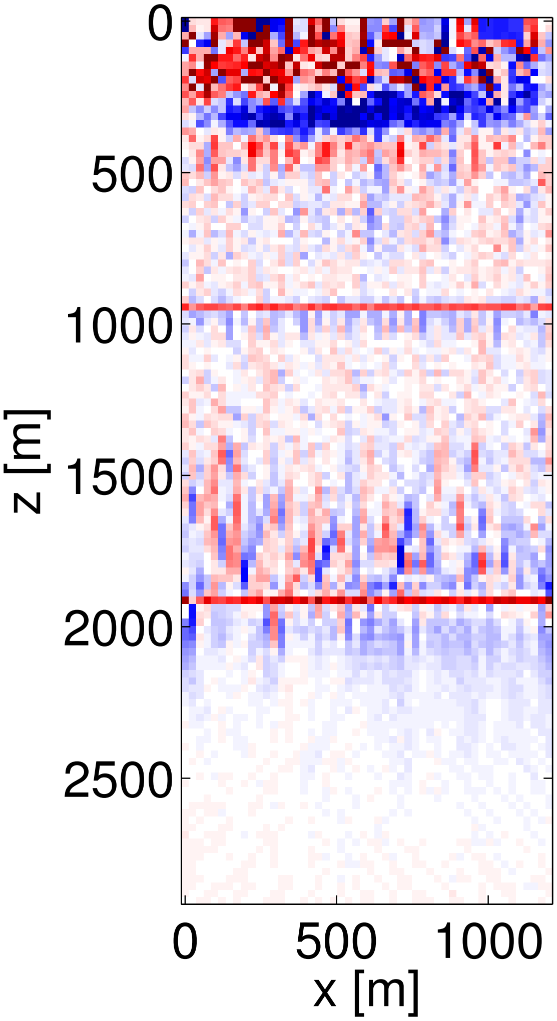


Figure 5-18: Variational Bayes MAP image using wave-equation modeled data (three layer model). Correlation with true image = 0.2291.

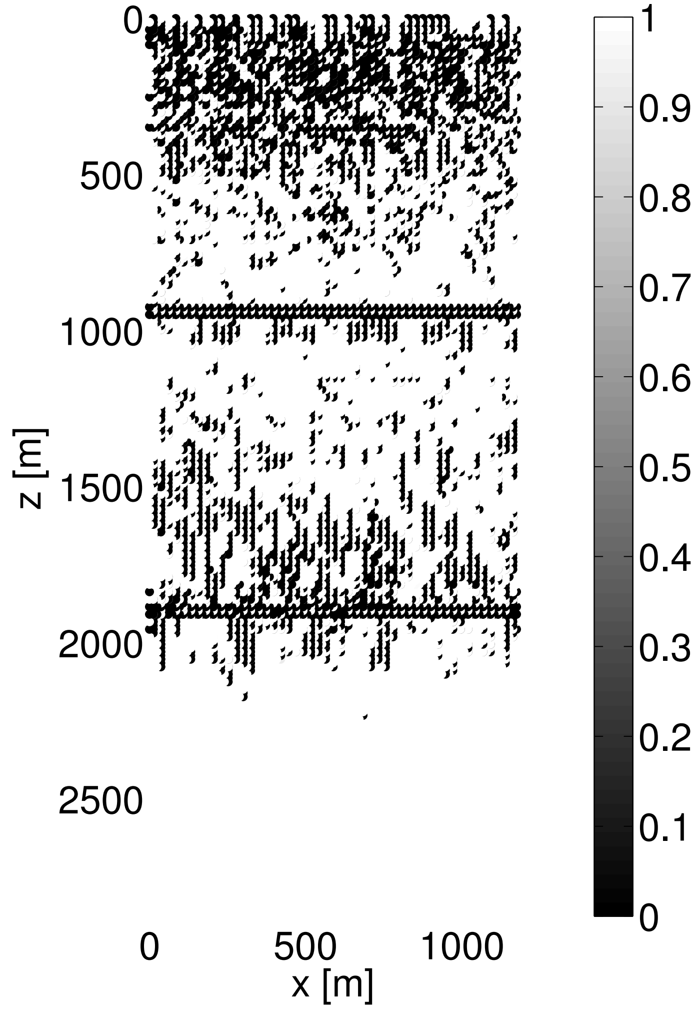


Figure 5-19: Edge strengths $\boldsymbol{\beta}^*$ estimated with VB using wave-equation modeled data (three layer model).

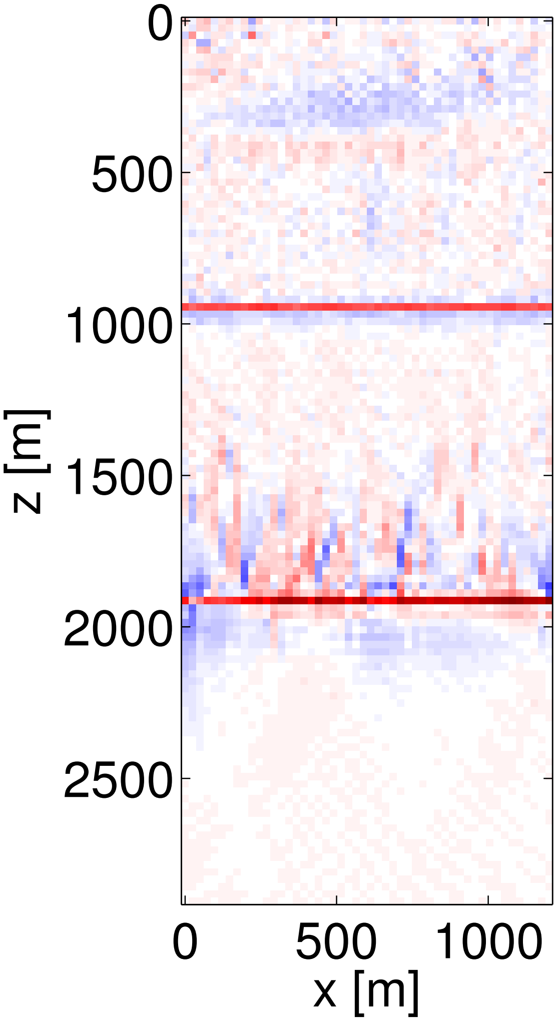


Figure 5-20: Empirical Bayes MAP image obtained with $\boldsymbol{\beta} = \boldsymbol{\beta}^*$, $\lambda = 10\bar{\lambda}^*$, and $\zeta = \bar{\zeta}^*$, using wave-equation modeled data (three layer model). Correlation with true image = 0.7472.

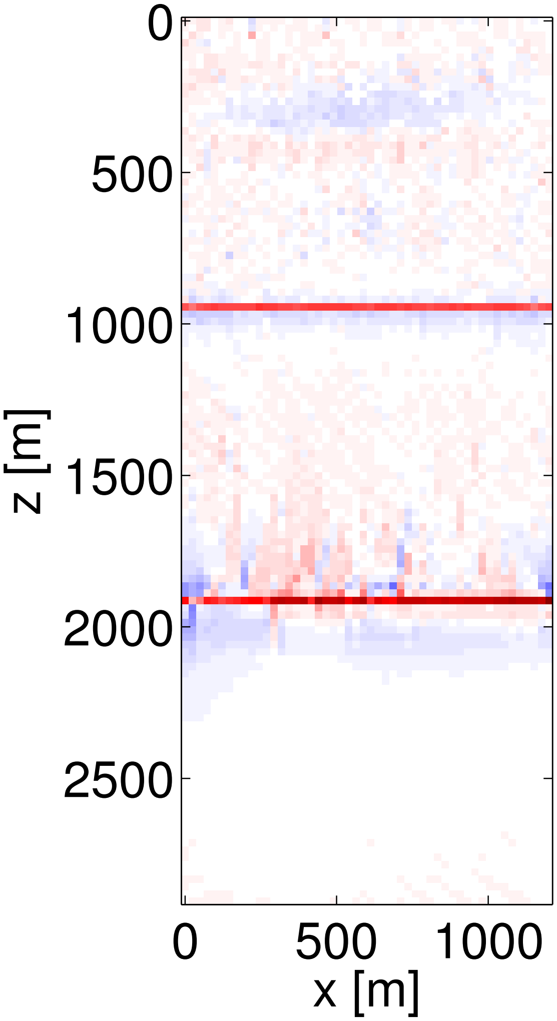


Figure 5-21: Empirical Bayes MAP image obtained with $\boldsymbol{\beta} = \boldsymbol{\beta}^*$, $\lambda = 10^2\bar{\lambda}^*$, and $\zeta = \bar{\zeta}^*$, using wave-equation modeled data (three layer model). Correlation with true image = 0.8724.

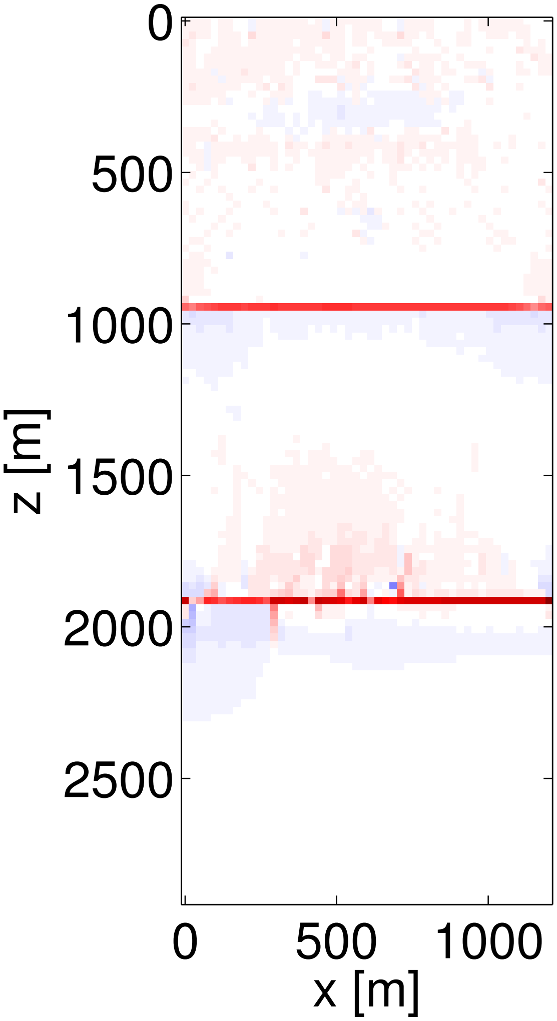
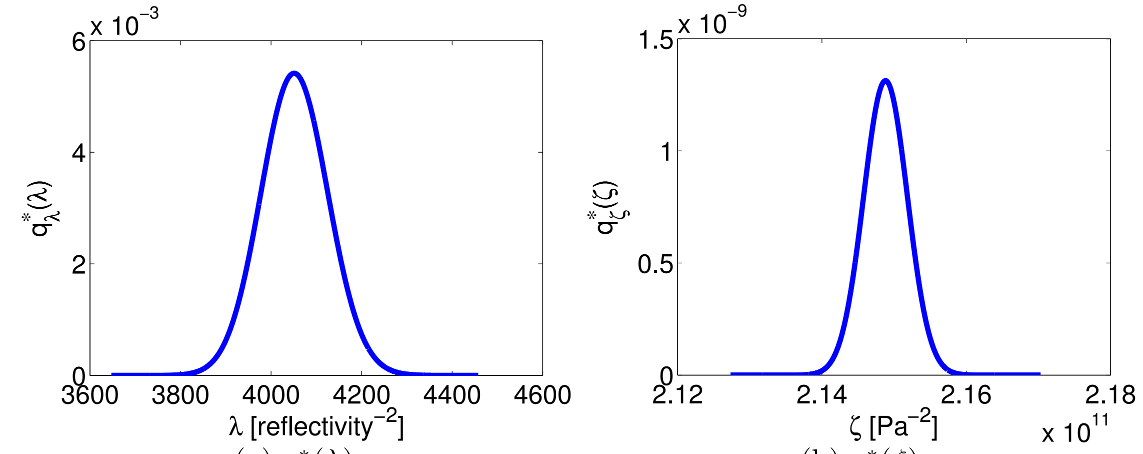


Figure 5-22: Empirical Bayes MAP image obtained with $\boldsymbol{\beta} = \boldsymbol{\beta}^*$, $\lambda = 10^3\bar{\lambda}^*$, and $\zeta = \bar{\zeta}^*$, using wave-equation modeled data (three layer model). Correlation with true image = 0.9349.



(a) $q^*_\lambda(\lambda)$

(b) $q^*_\zeta(\zeta)$

Figure 5-23: The variational Bayesian approximations to the posterior distributions for the parameters scaling the inverse variances of the (a) model λ and (b) noise ζ, using wave-equation modeled data (three layer model).

image is improved upon in the regularized LSM images (with $\beta_{ij} = 1$), but the image obtained with $\lambda = \bar{\lambda}^*$ (Figure 5-25) still contains a significant noise component. Setting λ to $10^2\bar{\lambda}^*$ or higher largely removes the effects of the noise (Figures 5-27–5-28), yet at the cost of over-smoothing the reflectors. Again, our expectation is that setting $\lambda = \bar{\lambda}^*$ should sufficiently remove the noise from the image, and we would thus expect a higher value for $\bar{\lambda}^*$.

This issue can again be seen in the VB-MAP image of Figure 5-29: indeed, it is difficult to observe any qualitative difference between the VB-MAP image and the unregularized LSM image of Figure 5-24. This again suggests that the value for $\bar{\lambda}^*$ is too low. The estimated edge strengths $\boldsymbol{\beta}^*$ are shown in Figure 5-30. We see that, for those reflectors that appeared in the LSM images (i.e. those reflectors that the data were informative about from the perspective of the Kirchhoff modeling operator), the edge strengths correctly go to 0 near the reflectors; elsewhere in the model the edge strengths are high, although the effect of the noise can also be seen in the edge strengths (particularly in the deeper part of the model). As with the three layer example, this suggests that using these edge strengths with a higher value for λ should yield a better image. We compute these images in Figures 5-31–5-33 using increasingly higher values of λ. When λ is set to $10^2\bar{\lambda}^*$ and $10^3\bar{\lambda}^*$ we begin to see some qualitative improvements in the image: the noise is smoothed out, yet some sharpness is preserved in the reflectors captured by $\boldsymbol{\beta}^*$. We note that the correlations of the images with the true model, for all the images obtained in this example, are very low (< 0.1). This is because, no matter what kind of regularization scheme is used, the images obtained only correctly estimate a small portion of the reflectors, since much of the acoustic data is not adequately described by the Kirchhoff operator. As such, even though we notice that the correlations do, for the most part, increase when we see qualitative improvements in the image, they are less informative for this example.

The approximate posterior marginals q^*_λ and q^*_ζ are shown in Figure 5-34. Here, VB estimates the inverse noise variance at $\bar{\zeta}^* = 1.254 \cdot 10^{10}$ Pa^{-2}, whereas the inverse variance of the data residuals found with the true model and MAP-VB model are

$\zeta = 9.858 \cdot 10^8$ Pa^{-2} and $\zeta = 1.269 \cdot 10^{10}$ Pa^{-2}, respectively. Hence, the estimated inverse noise variance greatly overestimates the "true" inverse noise variance (i.e. the inverse variance of the data residual with the true model) and, again, is far more consistent with the inverse variance of the residual the VB-MAP model. The VB estimate of λ was found to be $\bar{\lambda}^* = 150.97$, which results in an average prior model variance of $1.21 \cdot 10^{-2}$. By comparison, the empirical variance of the true model is $5.66 \cdot 10^{-3}$. Hence, again, the true model variance is significantly overestimated, meaning λ is underestimated by VB.

The question remains of why the VB estimate $\bar{\lambda}^*$ is, from both a results-oriented perspective and from comparison with the true model variance, significantly lower than expected. Alternatively, we can ask why the estimated inverse noise variance parameter $\bar{\zeta}^*$ is too high, since it is the ratio $\bar{\lambda}^*/\bar{\zeta}^*$ that ultimately determines the degree of regularization in LSM. We note that this only occurred with the wave-equation modeled data: the estimates of $\bar{\lambda}^*$ and $\bar{\zeta}^*$ for the Kirchhoff modeled data example yielded the expected results. This is likely because, for the wave-equation data, the residual is not well-modeled as zero-mean white noise. Indeed, the Kirchhoff operator models the single-scattering from the reflectors, while the acoustic wave-equation data contain information from the full wavefield, including multiple-scattering, refracted waves, and other coherent effects, which will remain in the residual after subtracting the Kirchhoff modeled data from the acoustic data. This correlation in the residual may lead to very strange models, such as the unregularized LSM images above, that fit the data considerably better than the true model. This data residual using such models may have a significantly lower variance than that of the data residual found with the true model, leading to a larger value being estimated for the inverse noise parameter, $\bar{\zeta}^*$. Indeed, we find that the unregularized LSM image gives a data residual with a variance roughly 1-2 orders of magnitude lower than the variance of the residual found using the true model. This would result in $\bar{\zeta}^*$ being estimated at a significantly higher value than expected, which is likely why we were able to compensate for this by increasing λ. This is no fault of the hierarchical Bayesian framework, but rather, as mentioned previously, a problem with the modeling. The real solution is to use a more realistic forward modeling operator (such as the Born wave-equation operator [31], for example) to better describe the data.

5.3 Conclusions

Finding the optimal regularization scheme in an inverse problem is an open research question which we have attempted to address. In the first part of this chapter we analytically investigated the meaning of the regularization parameters by considering the covariance function that arises in the limiting case of a continuous model. This analysis allows one to heuristically set the regularization parameters based on a desired correlation length and model variance (on the discrete grid). In the second part of this chapter, we attempt to estimate an optimal regularization scheme through the mathematical framework of Bayesian inference. In particular, we applied hierarchical Bayesian analysis to jointly estimate the model and regularization parameters. We utilized variational Bayesian methods to approximate a solution to the hierarchical Bayesian problem, and showed the application of this methodology in the context of least-squares migration. Inferring the regularization parameters allowed for significant improvements in the inferred seismic image, particularly for the Kirchhoff modeled data examples, where we are able to remove the effects of noise while still preserving sharpness at the reflectors in the image. The acoustic wave-equation modeled data examples proved more challenging due to the limited ability of the Kirchhoff modeling operator used in the inversion to fully describe the data; however we were still able to make similar qualitative improvements to the inferred images in the sections of the model that were well-described by the data. As was the case in the previous chapter, the methodology developed herein is applicable to a broad range of linear inverse problems involving spatially-varying model parameter statistics and is not limited to the seismic imaging problem.

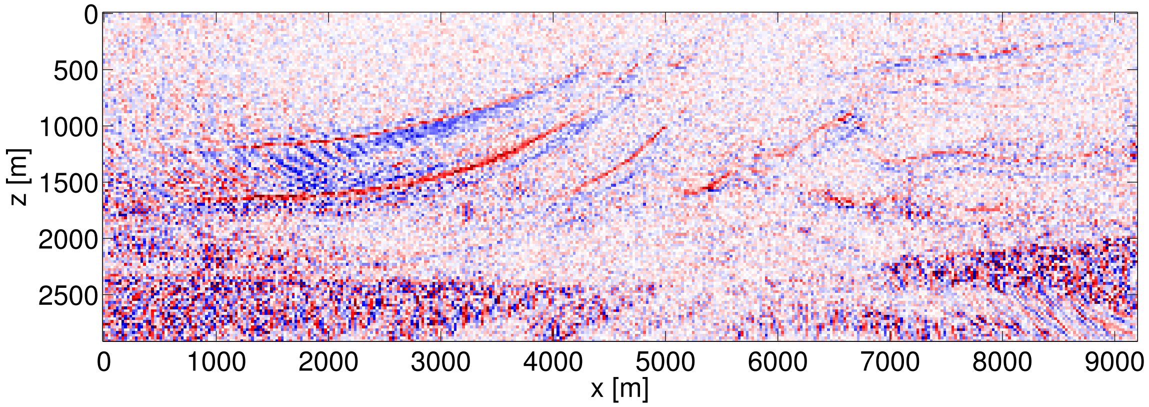


Figure 5-24: Unregularized LSM image (each $\beta_{ij} = 0$) using wave-equation modeled data (Marmousi model). Correlation with true image = 0.0520.

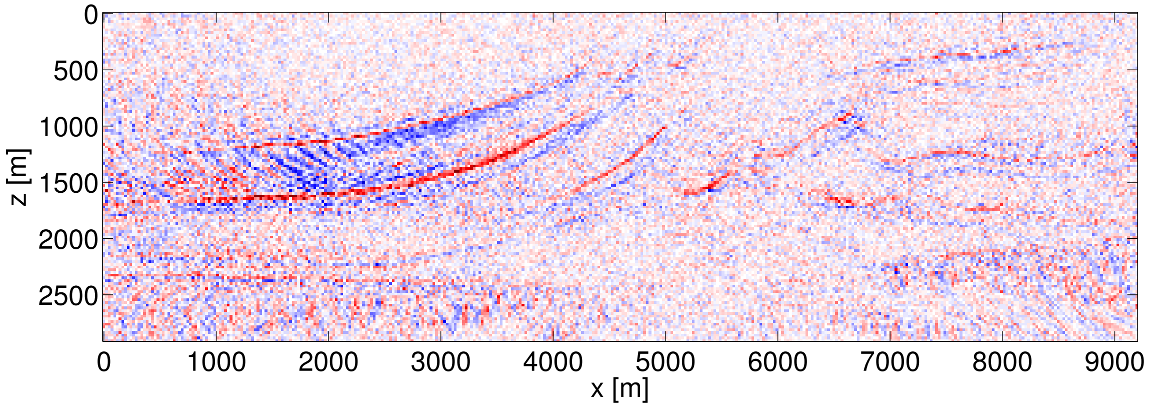


Figure 5-25: Uniformly regularized LSM image (each $\beta_{ij} = 1$) with $\lambda = \bar{\lambda}^*$ and $\zeta = \bar{\zeta}^*$ using wave-equation modeled data (Marmousi model). Correlation with true image = 0.0757.

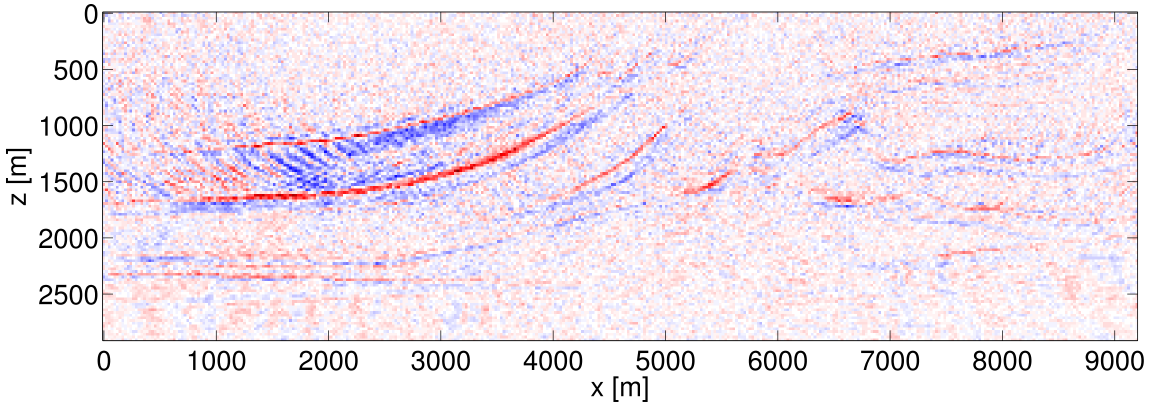


Figure 5-26: Uniformly regularized LSM image (each $\beta_{ij} = 1$) with $\lambda = 10\bar{\lambda}^*$ and $\zeta = \bar{\zeta}^*$ using wave-equation modeled data (Marmousi model). Correlation with true image = 0.0888.

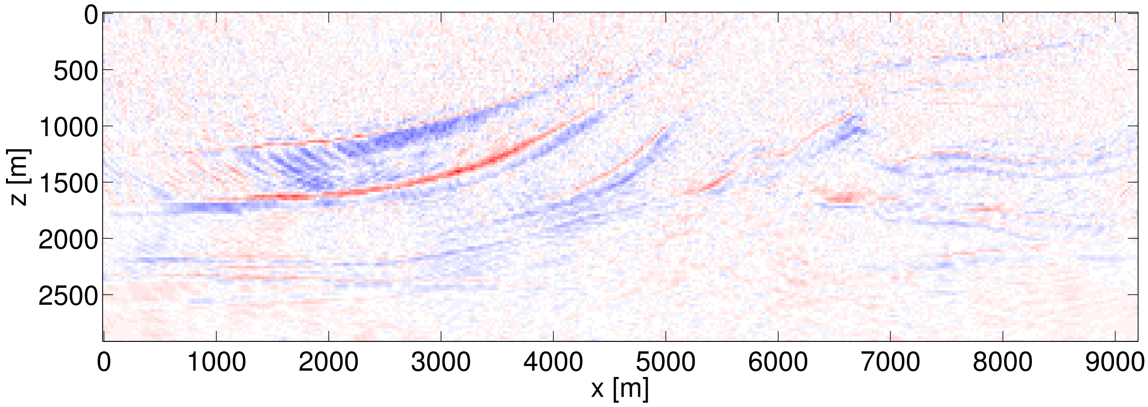


Figure 5-27: Uniformly regularized LSM image (each $\beta_{ij} = 1$) with $\lambda = 10^2\bar{\lambda}^*$ and $\zeta = \bar{\zeta}^*$ using wave-equation modeled data (Marmousi model). Correlation with true image $= 0.0728$.

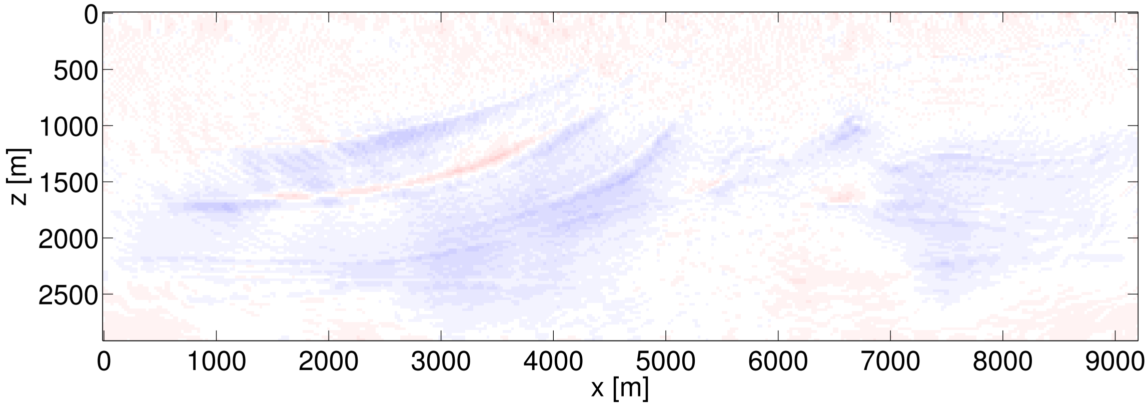


Figure 5-28: Uniformly regularized LSM image (each $\beta_{ij} = 1$) with $\lambda = 10^3\bar{\lambda}^*$ and $\zeta = \bar{\zeta}^*$ using wave-equation modeled data (Marmousi model). Correlation with true image $= 0.0268$.

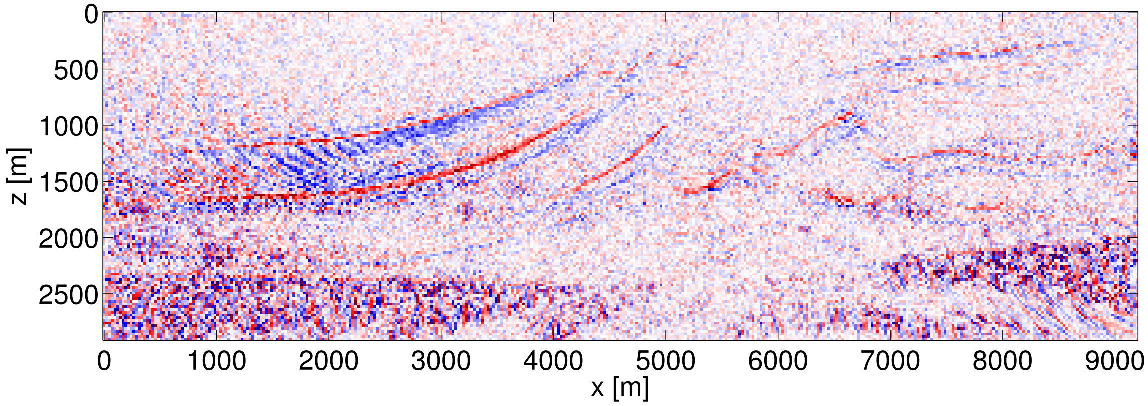


Figure 5-29: Variational Bayes MAP image using wave-equation modeled data (Marmousi model). Correlation with true image $= 0.0557$.

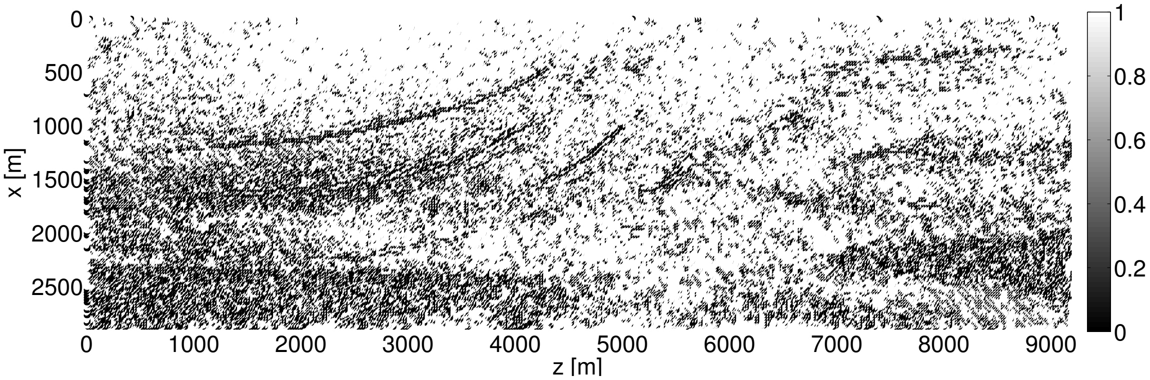


Figure 5-30: Edge strengths $\boldsymbol{\beta}^*$ estimated with VB using wave-equation modeled data (Marmousi model).

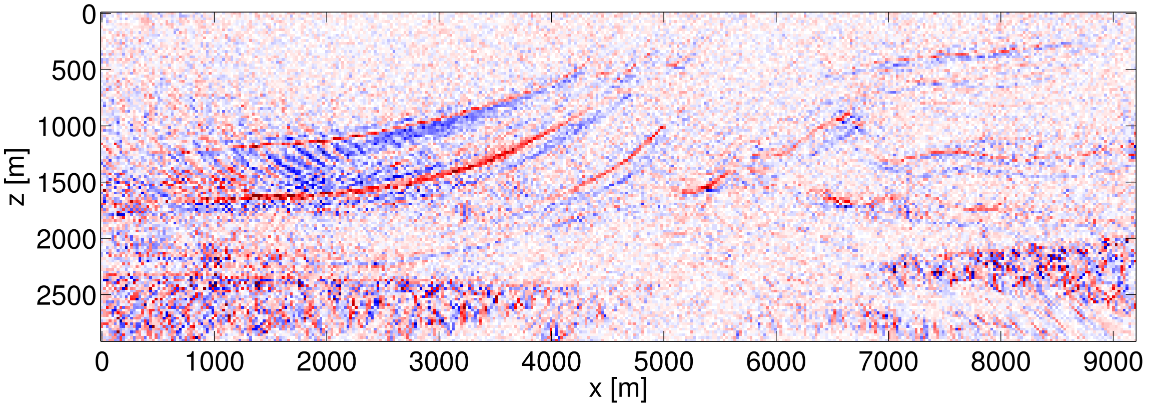


Figure 5-31: Empirical Bayes MAP image obtained with $\boldsymbol{\beta} = \boldsymbol{\beta}^*$, $\lambda = 10\bar{\lambda}^*$, and $\zeta = \bar{\zeta}^*$, using wave-equation modeled data (Marmousi model). Correlation with true image = 0.0672.

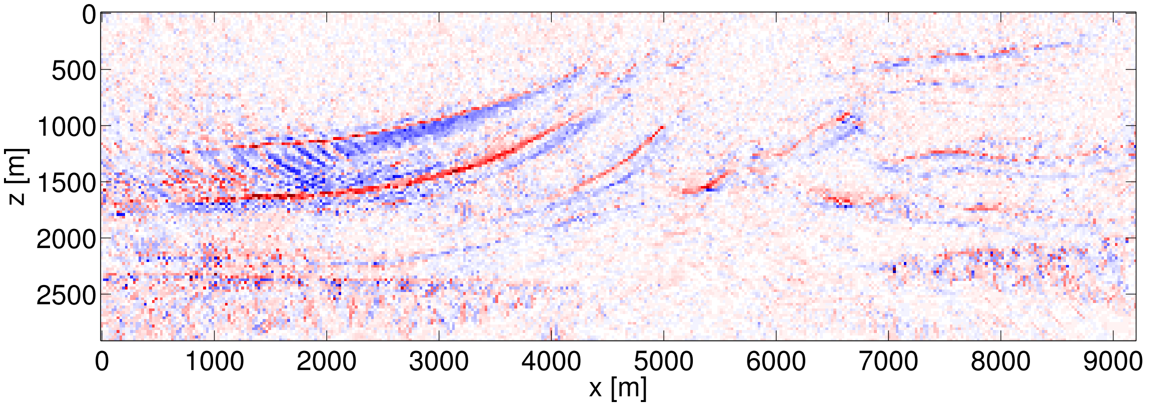


Figure 5-32: Empirical Bayes MAP image obtained with $\boldsymbol{\beta} = \boldsymbol{\beta}^*$, $\lambda = 10^2\bar{\lambda}^*$, and $\zeta = \bar{\zeta}^*$, using wave-equation modeled data (Marmousi model). Correlation with true image = 0.0828.

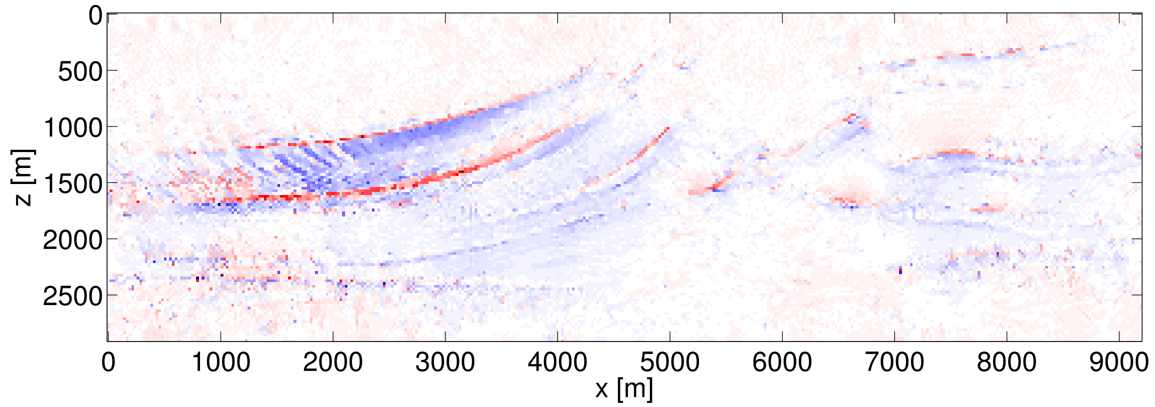
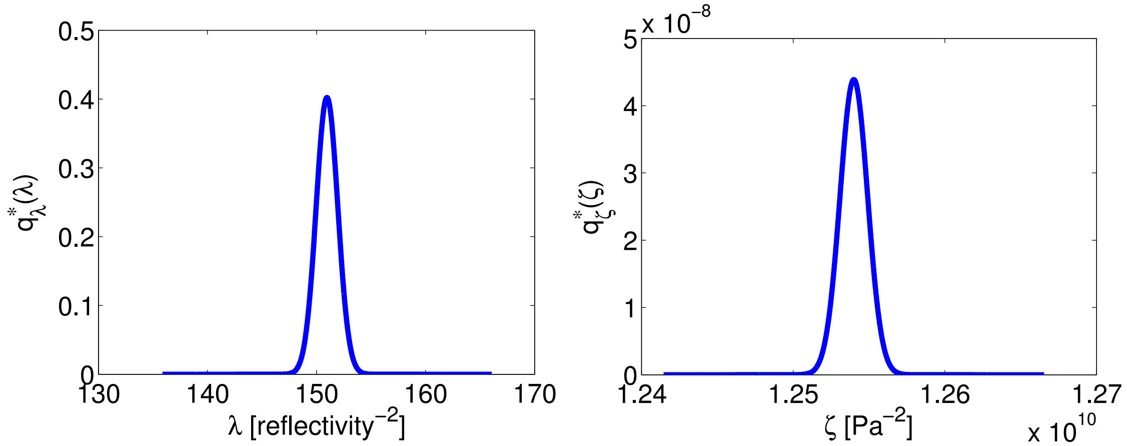


Figure 5-33: Empirical Bayes MAP image obtained with $\boldsymbol{\beta} = \boldsymbol{\beta}^*$, $\lambda = 10^3\bar{\lambda}^*$, and $\zeta = \bar{\zeta}^*$, using wave-equation modeled data (Marmousi model). Correlation with true image = 0.0762.



(a) $q^*_\lambda(\lambda)$ (b) $q^*_\zeta(\zeta)$

Figure 5-34: The variational Bayesian approximations to the posterior distributions for the parameters scaling the inverse variances of the (a) model λ and (b) noise ζ, using wave-equation modeled data (Marmousi model).

Chapter 6

Hierarchical Bayesian Time-Lapse Seismic Processing

6.1 Summary

In this chapter, we describe a methodology for inferring the change in subsurface model parameters from time-lapse seismic data within a hierarchical Bayesian framework, where the time-lapse and baseline surveys may have different acquisition geometries. Conventional methods for processing time-lapse data with differing acquisition geometries involve inverting the baseline and time-lapse datasets separately and subtracting the inverted models; however, such methods do not correctly account for differing model uncertainty between surveys due to differences in illumination and observational noise. Within the hierarchical Bayesian setting, the solution to the time-lapse inverse problem is given by the marginal *maximum a posteriori* (MAP) estimate of the time-lapse change, which seeks the most probable time-lapse change over all probable baseline models described by the data. We present a framework for computing the marginal MAP estimate using the expectation-maximization (E-M) algorithm, which iteratively performs sequential estimation of the time-lapse change and the baseline model. Our algorithm is validated numerically on synthetic data simulated from the Marmousi model (with a time-lapse perturbation), where the hierarchical Bayesian estimates significantly outperform conventional time-lapse in-

version results.

6.2 Introduction

Monitoring changes in the subsurface geophysical properties of a field over time can provide valuable information for, among other things, reservoir modeling and production planning of the field. Repeated time-lapse seismic surveys are commonly used for this purpose, but conventional methods for processing such surveys suffer from a number of limitations. One conventional method for inverting time-lapse seismic data involves subtracting repeated datasets and inverting the differenced data to obtain the changes in the subsurface model parameters [8, 30]; however, the validity of this method requires both that the seismic data depend linearly on the subsurface model and that the repeated datasets have identical acquisition geometries. A recent advance known as double-difference inversion [81] does not require this linearity for correctness but still requires identical acquisition geometries. However, achieving identical acquisitions in sequential surveys can be challenging. Furthermore, time-lapse surveys may differ significantly over time due to the availability of new acquisition technologies. In the case of differing acquisitions, conventional time-lapse inversion involves inverting the datasets separately and subtracting the inverted models to estimate the time-lapse change. However, this method (model subtraction) often performs quite poorly due to differences in illumination and observational noise. It is also quite possible that the time-lapse survey contains information about the baseline model that was not captured by the baseline survey. Ayeni and Biondi [2] describe a framework for target-oriented linear least-squares inversion of multiple time-lapse datasets with differing acquisitions, but their methodology only applies to the case when the data depend linearly on the model and requires explicit computation and storage of a Hessian matrix, which is only feasible for a small target region.

In this chapter, we describe a methodology for estimating the time-lapse change in the model parameters from both datasets simultaneously in a hierarchical Bayesian setting, where we neither require that the data depend linearly on the model nor that the datasets have similar acquisition. In particular, we solve the Bayesian inference problem using a gradient-based implementation of the expectation-maximization algorithm, which iterates between subsequent updates to the background model and time-lapse change, ultimately yielding an estimate of the best time-lapse change over all probable baseline models described by the data.

6.3 Methodology and Bayesian Framework

We consider the case where we have two time-lapse seismic surveys with different source-receiver geometries, a baseline survey yielding data vector $\mathbf{d}_0$ and a monitor survey yielding $\mathbf{d}_1$, and we wish to infer the time-lapse change in some subsurface model parameters $\mathbf{m}$ from $\mathbf{d}_0$ and $\mathbf{d}_1$. Here $\mathbf{m}$ may, for example, be the P-wave propagation velocity or the reflectivity of the medium. We denote the baseline model by $\mathbf{m}_0$, the time-lapse change by $\Delta\mathbf{m}$, and denote by F_0 and F_1 the operators relating the model parameters to the baseline and monitor datasets, respectively. Then we can give the data as

$$\mathbf{d}_0 = F_0(\mathbf{m}_0) + \mathbf{n}_0 \tag{6.1}$$

and

$$\mathbf{d}_1 = F_1(\mathbf{m}_0 + \Delta\mathbf{m}) + \mathbf{n}_1, \tag{6.2}$$

where $\mathbf{n}_0$ and $\mathbf{n}_1$ are noise terms.

The Bayesian inference setting provides a useful and mathematically rigorous framework within which the problem of inferring $\Delta\mathbf{m}$ can be cast. Within the Bayesian context, both the model parameters $\mathbf{m}_0$, $\Delta\mathbf{m}$ and the observed data $\mathbf{d}_0$, $\mathbf{d}_1$ are viewed as random quantities defined by a probabilistic model. In particular, we let $\mathbf{m}_0$ and $\Delta\mathbf{m}$ be independent Gaussian random vectors with prior means $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_\Delta$ and prior covariance matrices C_0 and C_Δ. We relate the prior covariance matrices to prior assumptions about the spatial statistics of $\mathbf{m}_0$ and $\Delta\mathbf{m}$ by specifying them via the same model for the prior precision matrix described in previous chapters:

$$C_i = (\lambda_i(D + \epsilon_i I))^{-1} \qquad i = 0, \Delta, \tag{6.3}$$

where D is a differencing matrix and λ_i and ϵ_i are parameters governing the prior variance and correlation length of spatial variations (cf. Chapter 5). We arrive at a probabilistic model for the data by modeling $\mathbf{n}_0$ and $\mathbf{n}_1$ as zero-mean Gaussian noise (independent of the model) with covariance matrices Σ_0 and Σ_1, respectively, where $\Sigma_i = \sigma_i I$. Bayes' rule then gives the joint posterior distribution for $\mathbf{m}_0$ and $\Delta\mathbf{m}$ given the data:

$$\begin{aligned} p(\mathbf{m}_0, \Delta\mathbf{m}|\mathbf{d}_0, \mathbf{d}_1) &\propto p(\mathbf{m}_0, \Delta\mathbf{m}) p(\mathbf{d}_0, \mathbf{d}_1|\mathbf{m}_0, \Delta\mathbf{m}) \\ &\propto \exp\left\{ -\frac{1}{2} \left[\|\mathbf{m}_0 - \boldsymbol{\mu}_0\|^2_{C_0^{-1}} + \|\Delta\mathbf{m} - \boldsymbol{\mu}_\Delta\|^2_{C_\Delta^{-1}} \right.\right. \\ &\qquad \left.\left. + \|\mathbf{d}_0 - F_0(\mathbf{m}_0)\|^2_{\Sigma_0^{-1}} + \|\mathbf{d}_1 - F_1(\mathbf{m}_0 + \Delta\mathbf{m})\|^2_{\Sigma_1^{-1}} \right] \right\} \end{aligned} \tag{6.4}$$

where we have defined the notation $\|\mathbf{x}\|^2_W \triangleq \mathbf{x}^T W \mathbf{x}$. We note that the joint posterior for $(\mathbf{m}_0, \Delta\mathbf{m})$ is *not* Gaussian when either F_0 or F_1 is non-linear.

In order to infer the time-lapse change $\Delta\mathbf{m}$ with an unknown baseline model $\mathbf{m}_0$, we seek the hierarchical Bayesian *maximum a posteriori* (MAP) estimate for $\Delta\mathbf{m}$. This requires marginalization of the joint posterior distribution (6.4) over the space of all baseline models $\mathcal{M}_0$ to obtain the marginal MAP solution:

$$\Delta\mathbf{m}_{\text{MAP}} = \underset{\Delta\mathbf{m}}{\arg\max} \ \log \int_{\mathcal{M}_0} p(\mathbf{m}_0, \Delta\mathbf{m}|\mathbf{d}_0, \mathbf{d}_1) d\mathbf{m}_0. \tag{6.5}$$

We note that this marginal MAP approach is mathematically quite different from the joint inversion approach: whereas the joint inversion approach would seek the pair $(\mathbf{m}_0, \Delta\mathbf{m})$ that maximizes the joint posterior distribution (i.e. the joint MAP), the marginal MAP approach seeks to find the single best choice for $\Delta\mathbf{m}$ over *all probable* choices for the baseline model $\mathbf{m}_0$. In the special case where F_0 and F_1 are linear, the posterior distribution is Gaussian, and hence the joint MAP solution found by joint inversion of the linear problem will happen to coincide with the marginal MAP solution of (6.5); however, this is not true for general F_0 and F_1. Furthermore, since the posterior distribution is, in general, not Gaussian, the integral in (6.5) might not be analytically tractable. Rather than attempting to numerically explore the high-dimensional model space $\mathcal{M}_0$, we turn to the expectation-maximization algorithm to iteratively solve the marginal MAP problem.

6.3.1 The E-M Algorithm for Time-Lapse Inversion

The E-M algorithm [17, 46] solves maximum likelihood or MAP estimation problems when a subset of the data relevant to the parameter estimation is unobserved (referred to as latent variables). In the time-lapse problem considered here, we view the baseline model $\mathbf{m}_0$ as the latent variables. E-M can be thought of as a coordinate ascent algorithm for solving the marginal MAP optimization problem (6.5), whereby alternating estimations are performed between the latent variables and the parameters of interest $\Delta\mathbf{m}$. Specifically, the E-M algorithm iteratively updates our estimate of the time-lapse model $\widehat{\Delta\mathbf{m}}^{(t)}$ according to:

$$\widehat{\Delta\mathbf{m}}^{(t+1)} = \arg\max_{\Delta\mathbf{m}} \left\{ \log p(\Delta\mathbf{m}) + \mathbb{E}_{p\left(\mathbf{m}_0|\mathbf{d}_0,\mathbf{d}_1,\widehat{\Delta\mathbf{m}}^{(t)}\right)} \left[\log p(\mathbf{m}_0,\mathbf{d}_0,\mathbf{d}_1|\Delta\mathbf{m})\right] \right\}, \quad (6.6)$$

where $\mathbb{E}_p$ denotes an expectation taken with respect to a probability distribution p, which in this case is the posterior distribution of $\mathbf{m}_0$ conditioned on the previous iterate $\widehat{\Delta\mathbf{m}}^{(t)}$. Plugging in for probability distributions, we find the E-M algorithm updates $\widehat{\Delta\mathbf{m}}^{(t)}$ by minimizing a cost function $\phi^{(t)}$ given by

$$\phi^{(t)}(\Delta\mathbf{m}) = \|\Delta\mathbf{m} - \boldsymbol{\mu}_\Delta\|^2_{C_\Delta^{-1}} + \mathbb{E}_{p\left(\mathbf{m}_0|\mathbf{d}_0,\mathbf{d}_1,\widehat{\Delta\mathbf{m}}^{(t)}\right)} \left[\|\mathbf{d}_1 - F_1(\mathbf{m}_0 + \Delta\mathbf{m})\|^2_{\Sigma_1^{-1}}\right]. \quad (6.7)$$

We can perform the above minimization using a first-order gradient-based method, such as gradient descent or non-linear conjugate gradients [27], where the gradient

$\nabla\phi^{(t)}$ is given by

$$\begin{aligned}\nabla\phi^{(t)}(\Delta\mathbf{m}) &= 2C_\Delta^{-1}(\Delta\mathbf{m}-\boldsymbol{\mu}_\Delta)\\ &\qquad + 2\mathbb{E}_{p\left(\mathbf{m}_0|\mathbf{d}_0,\mathbf{d}_1,\widehat{\Delta\mathbf{m}}^{(t)}\right)}\left[A_1^T(\mathbf{m}_0+\Delta\mathbf{m})\Sigma_1^{-1}(\mathbf{d}_1-F_1(\mathbf{m}_0+\Delta\mathbf{m}))\right],\end{aligned} \tag{6.8}$$

where $A_1(\mathbf{m}_0+\Delta\mathbf{m})$ is the Jacobian of F_1 evaluated at $\mathbf{m}_0+\Delta\mathbf{m}$.

In order to compute the expected values in (6.7) and (6.8), we would need to utilize sampling techniques to draw samples from $p(\mathbf{m}_0|\mathbf{d}_0,\mathbf{d}_1,\widehat{\Delta\mathbf{m}}^{(t)})$. At the cost of replacing the marginal MAP solution with the joint MAP solution, a computationally cheaper option would be to approximate $\phi^{(t)}(\Delta\mathbf{m})$ by replacing the conditional expectation of the data misfit with its value at the conditional MAP estimate of $\mathbf{m}_0$, so that the E-M cost function is approximated by

$$\phi^{(t)}(\Delta\mathbf{m}) \approx \|\Delta\mathbf{m}-\boldsymbol{\mu}_\Delta\|^2_{C_\Delta^{-1}} + \left\|\mathbf{d}_1 - F_1\left(\mathbf{m}_{\mathbf{0}\mathrm{MAP}|\widehat{\Delta\mathbf{m}}^{(t)},\mathbf{d}_0,\mathbf{d}_1}+\Delta\mathbf{m}\right)\right\|^2_{\Sigma_1^{-1}} \tag{6.9}$$

and its gradient is

$$\begin{aligned}\nabla\phi^{(t)}(\Delta\mathbf{m}) &\approx 2C_\Delta^{-1}(\Delta\mathbf{m}-\boldsymbol{\mu}_\Delta)\\ &\qquad + 2A_1^T(\mathbf{m}_{\mathbf{0}\mathrm{MAP}|\widehat{\Delta\mathbf{m}}^{(t)},\mathbf{d}_0,\mathbf{d}_1}+\Delta\mathbf{m})\Sigma_1^{-1}(\mathbf{d}_1-F_1(\mathbf{m}_{\mathbf{0}\mathrm{MAP}|\widehat{\Delta\mathbf{m}}^{(t)},\mathbf{d}_0,\mathbf{d}_1}+\Delta\mathbf{m})),\end{aligned} \tag{6.10}$$

where

$$\begin{aligned}\mathbf{m}_{\mathbf{0}\mathrm{MAP}|\widehat{\Delta\mathbf{m}}^{(t)},\mathbf{d}_0,\mathbf{d}_1} = \arg\min_{\mathbf{m}_0}\ & \|\mathbf{m}_0-\boldsymbol{\mu}_0\|^2_{C_0^{-1}} + \|\mathbf{d}_0-F_0(\mathbf{m}_0)\|^2_{\Sigma_0^{-1}}\\ & + \|\mathbf{d}_1-F_1(\mathbf{m}_0+\widehat{\Delta\mathbf{m}}^{(t)})\|^2_{\Sigma_1^{-1}}.\end{aligned} \tag{6.11}$$

To summarize, our implementation of the E-M algorithm for the time-lapse inversion problem is as follows:

Algorithm 6.1 E-M Algorithm for Time-Lapse Inversion

```
Initialize $\widehat{\Delta\mathbf{m}}^{(0)}$. Set $t = 0$. Iterate on $t$:
1. Estimate $\mathbb{E}_{p(\mathbf{m}_0|\mathbf{d}_0,\mathbf{d}_1,\widehat{\Delta\mathbf{m}}^{(t)})}[\|\mathbf{d}_1 - F_1(\mathbf{m}_0 + \Delta\mathbf{m})\|^2_{\Sigma_1^{-1}}]$, the expected value of the
     data misfit in (6.7), and its gradient in (6.8) by either:
     a. Sampling the baseline model from $p\left(\mathbf{m}_0|\mathbf{d}_0, \mathbf{d}_1, \widehat{\Delta\mathbf{m}}^{(t)}\right)$, or
     b. Minimizing (6.11) to obtain the MAP estimate of the baseline model
          and using the approximations of (6.9) and (6.10).
2. Update $\widehat{\Delta\mathbf{m}}^{(t+1)} = \arg\min_{\Delta\mathbf{m}} \phi^{(t)}(\Delta\mathbf{m})$.
```

6.4 Numerical Results

We demonstrate our results numerically on a simple example involving the Marmousi model. As a test case, we consider the seismic imaging problem, where the model parameters $(\mathbf{m}_0, \Delta\mathbf{m})$ are reflectivity (with the smooth part of the velocity model known), and a localized time-lapse change in the model is introduced; the true baseline model and time-lapse changes are shown in Figure 6-1. We take F_0 and F_1 to be the Kirchhoff modeling operators for the source-receiver geometries in $\mathbf{d}_0$ and $\mathbf{d}_1$, respectively. (We note that since this example is for a linear problem, the joint and marginal MAP solutions coincide.) The synthetic datasets are inverse crime data (plus noise) generated from an array of surface receivers responding to a single 20 Hz Ricker source at the surface, where the horizontal position of the source is significantly shifted from $x_{s_0} = 3.4$ km in the baseline survey to $x_{s_1} = 7.2$ km in the monitor survey.

Figure 6-2 shows the results of the conventional inversion for the time-lapse change found by separately inverting the two datasets and subtracting the inverted models. Figure 6-3 shows the marginal MAP estimate for the time-lapse change and the expected baseline model obtained after 20 iterations of the E-M algorithm. In order to remove the effects of the additive noise and better capture true change in the model, we thresholded the estimates of the time-lapse changes above the noise level. Since the source locations between the two surveys are so far apart, the surveys illuminate

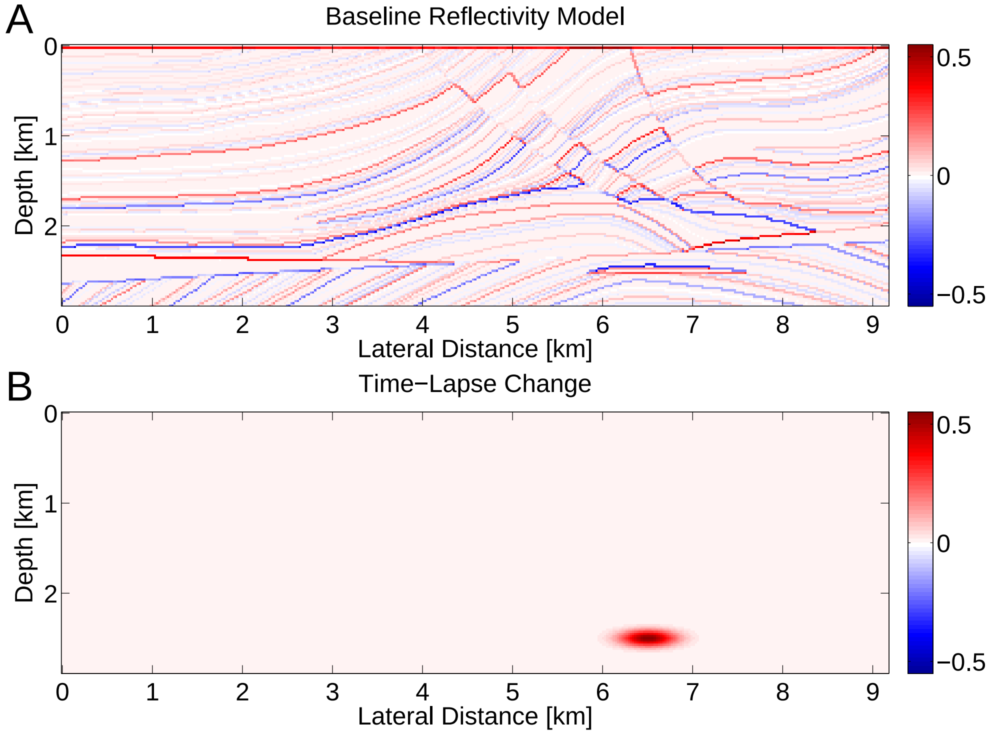


Figure 6-1: (A) True baseline reflectivity model and (B) time-lapse change.

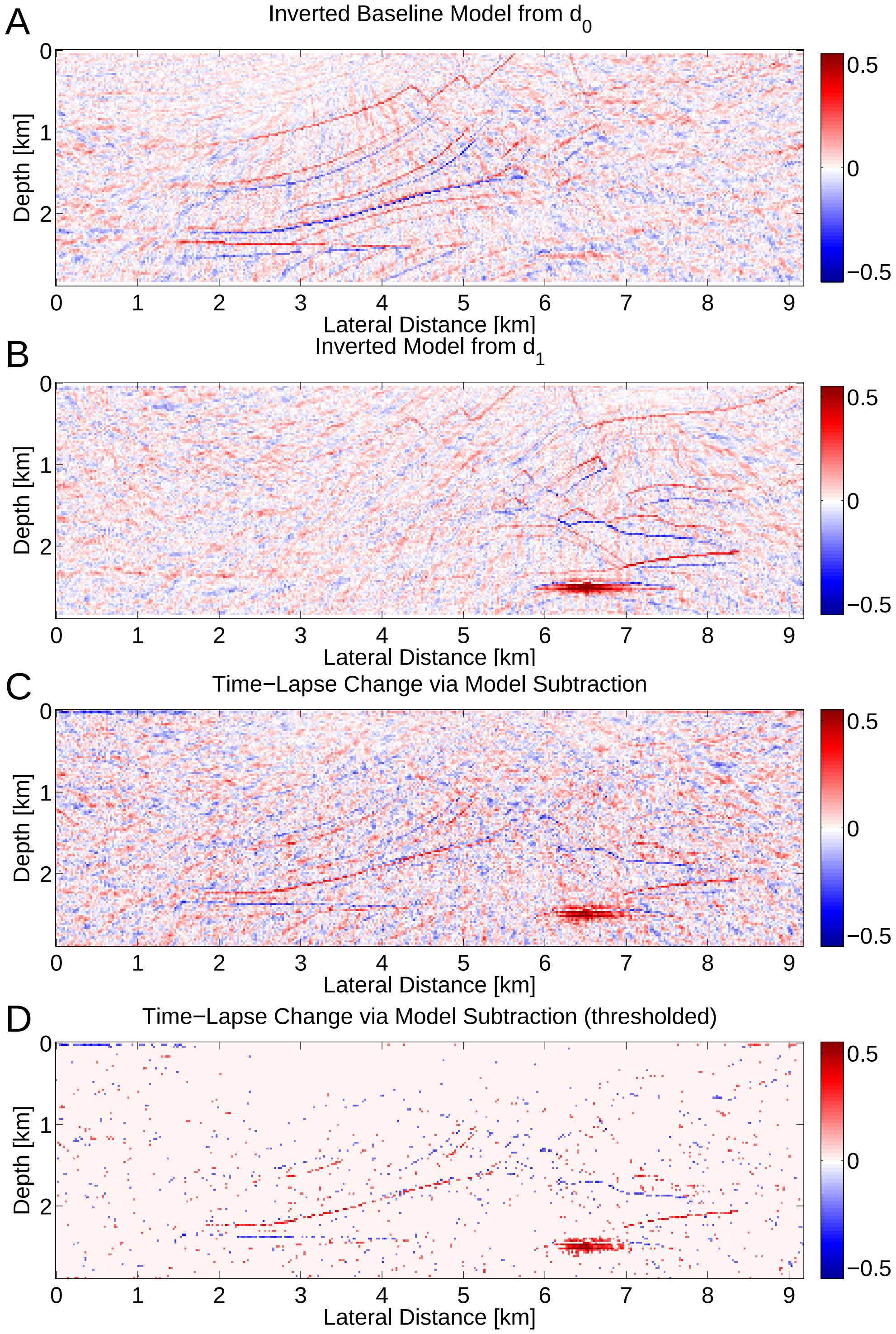


Figure 6-2: Conventional inversion results for the baseline model and time-lapse change. (A,B) $\mathbf{d}_0$ and $\mathbf{d}_1$ are inverted separately to estimate $\mathbf{m}_0$ and $\mathbf{m}_0 + \Delta\mathbf{m}$. (C) The time-lapse change is estimated by subtracting the inversion results and (D) thresholded to remove the effects of noise.

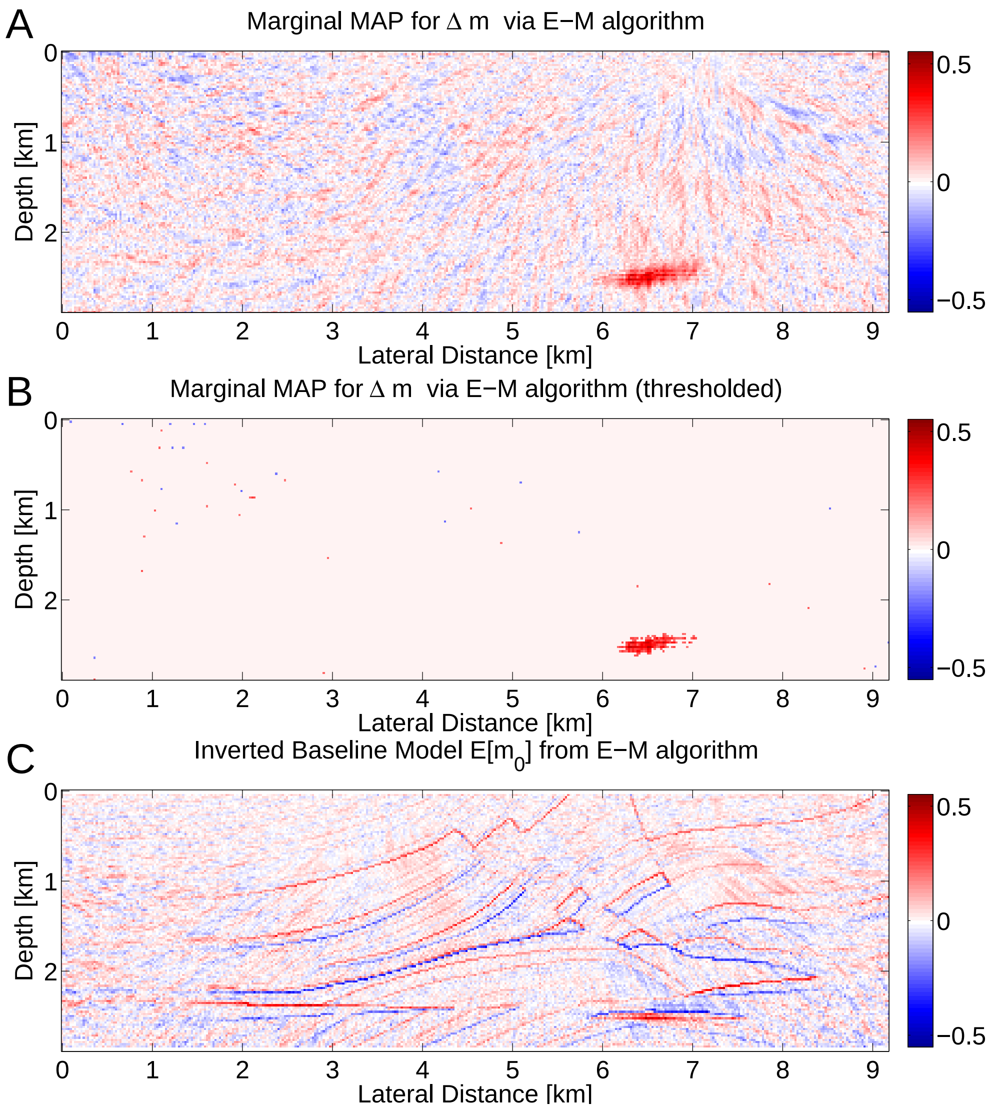


Figure 6-3: Inversion results obtained with our hierarchical Bayesian framework. (A) The marginal MAP solution $\widehat{\Delta\mathbf{m}}^{(t)}$ after 20 E-M iterations which is then (B) thresholded to remove the effects of noise. (C) The E-M estimate of the baseline model $\mathbb{E}[\mathbf{m}_0|\widehat{\Delta\mathbf{m}}^{(t)}, \mathbf{d}_0, \mathbf{d}_1]$.

different (but overlapping) sections of the model domain (as seen in Figure 6-2(A,B)). This results in sections of the baseline model appearing in the time-lapse estimate obtained from model subtraction, as is evident from both Figures 6-2(C) and 6-2(D). In contrast, even prior to thresholding, we see that the structure of the baseline model does not appear in the marginal MAP estimate obtained from the E-M algorithm in Figure 6-3(A) (indeed the noise is truly just noise), and, after thresholding (Figure 6-3(B)), the time-lapse change has been completely isolated from the background. The E-M algorithm also gives the expected baseline model (Figure 6-3(C)), where we see far better illumination since both datasets are used to compute this model.

6.5 Conclusions

In this chapter, we applied the hierarchical Bayesian framework to the problem of estimating the change in subsurface model parameters from time-lapse seismic datasets with differing acquisition geometries. In particular, our gradient-based implementation of the expectation-maximization algorithm solved the marginal MAP problem for the time-lapse change model by performing subsequent updates to the time-lapse change and baseline models. As verified by our numerical results, the marginal MAP estimate for the time-lapse change did not contain structure from the baseline model, and, also important, the estimated baseline model is constrained by both datasets. While we provided a numerical example for the time-lapse seismic imaging of reflectivity, the framework and algorithm detailed in this chapter are general and can be applied to other time-lapse inverse problems, without any requirement on the linearity of the forward operators F_0 and F_1. Although our method is computationally intensive (each iteration of the E-M algorithm requires solving two inverse problems), we find that the algorithm performs well with relatively few iterations; hence, it should be feasible to run this algorithm if the cost of a single inversion is reasonable. A straight-forward extension of this work would be to the case of inferring a series of time-lapse changes in the model from multiple monitor datasets (rather than just one).

Chapter 7

Conclusions

In this thesis, we explored the application of Bayesian inference methods to different geophysical inverse problems involving seismic data. Of particular focus was the question of how to appropriately introduce regularization in an inverse problem where the smoothness properties of the underlying earth model may vary with space. This question can be equivalently posed as how to pick an appropriate prior distribution in the Bayesian inference setting. To address this question, we defined the prior distribution on the model via a Markov random field and parameterized the edges of the MRF with edge strengths that capture the local smoothness properties of the prior. We explored the utility of this representation through its application to different geophysical inverse problems. Below, we summarize the main contributions and conclusions of these studies.

7.1 Summary of Main Contributions

Chapter 3 serves as our first study into the application of Bayesian inference and probabilistic graphical models to a geophysical inverse problem. Here we apply a non-hierarchical Bayesian inference framework (where the edge strengths of the MRF were predetermined) to the problem of characterizing the fractured nature of a reservoir from seismic data. The Bayesian setting allows for combining scattering data and amplitude measures that contain information about anisotropy under a single inversion framework, thereby allowing for the inversion of fracture properties under a larger physical regime than would be attainable using only one of these data types. We further show the capability of the edge strengths to both enforce smoothness in the estimates of the fracture properties and capture *a priori* information about geological features in the model, such as a discontinuity arising from a fault whose location is known.

In Chapter 4, we address the question of how to optimize the edge strengths of the MRF in the context of the seismic imaging problem, where the seismic image, consisting of sharp coherent reflectors, naturally tends to exhibit spatially-varying smoothness properties. We formulate the seismic imaging problem within the hierarchical Bayesian framework, treating the edge strengths as random variables to be inferred from the data. The problem of inferring these edge strengths presents significant computational challenges: the cost of evaluating the posterior distribution seemed initially to create a computational bottleneck that prevented our approach from being scalable to large models. The use of the expectation-maximization (E-M) algorithm along with the approximate methods detailed in Chapter 4 was a breakthrough in this regard, allowing the scalability of our method to more realistic-sized models. We obtain the marginal MAP estimate of the edge strengths via the E-M algorithm, thereby resulting in a prior distribution on the model that correctly captures its spatial smoothness properties. This allows for mitigating the effects of limited acquisition and observational noise in the estimated image while still preserving sharpness at the reflectors.

Chapter 5 extends the work of Chapter 4 by providing a methodology for choosing the remaining parameters (other than the edge strengths) that define the prior distribution for the model. Here, we first derive the relationship between these parameters and the prior model covariance and then extend the hierarchical Bayesian framework of Chapter 4 to include these additional parameters. Our derivation provides insight into these parameters and may provide guidance in selecting them in the future. We note that while the work of Chapters 4-5 is presented within the context of the seismic imaging problem, the methodology developed in these chapters can be applied to other linear inverse problems where the model parameters exhibit spatially (or temporally) varying smoothness properties.

In Chapter 6, we explore the application of the hierarchical Bayesian framework to the problem of time-lapse seismic inversion, where the objective is to infer the change in the subsurface model parameters over time by taking repeated seismic surveys. We consider the case where the surveys are taken with different acquisition geometries causing conventional methods for time-lapse inversion to perform poorly. We develop a novel and computationally tractable approach to the time-lapse inversion problem by applying the E-M algorithm to obtain the marginal MAP estimate of the time-lapse change in the model parameters. In contrast to the results obtained by conventional methods, the marginal MAP estimate for the time-lapse change does not contain structure from the background model, and, furthermore, the background model we estimate is based on information in both the baseline and monitor surveys.

7.2 Directions for Future Work

Here we suggest some avenues for future work stemming from the research presented in this thesis.

Alternative Parameterizations

One direction for future work is to explore alternative ways to parameterize the prior on the model within the hierarchical Bayesian setting. While we have shown through various examples that our choice of parameterization using edge strengths is able to capture the spatially-varying smoothness properties of the model, this choice is by no means unique, and it may be possible to improve on our results via alternative parameterizations of the prior. Additionally, throughout this thesis the number of model parameters has been a fixed input to the inference procedure, however the choice of how to discretize the earth model (i.e. what size of grid cells should be used) is non-trivial. In our case, the grid cell size was chosen relative to the seismic wavelength in the medium (as this determines the scale of the heterogeneities that scatter the seismic wavefield), however it is possible to optimize this via Bayesian model selection (sometimes referred to as transdimensional Bayesian inference). Transdimensional Bayesian methods have been previously applied to geophysical inverse problems [9, 43, 44]. A promising approach developed by Bodin [6] utilizes a transdimensional Bayesian method using Voronoi cells with mobile geometry, shape, and number to parameterize the model. Extending our methodology to incorporate these generalizations is hence another fruitful direction for future research. Finally, a gridded model parameterization may not be the most optimal for the seismic imaging problem. Considering the shape of the reflectors commonly encountered in seismic imaging, one may wish to instead parameterize the model using basis functions defined by the reflectors rather than individual grid cells. Here again, Bayesian model selection could be utilized to infer the number of reflectors in addition to the parameters defining each reflector.

Generalization to Non-linear Problems

The methodology presented in Chapters 4-5 for estimating the edge strengths from the data was formulated in the context of the linear inverse problem of seismic imaging when the background velocity model is known. In cases where the velocity model is not given, it must first be estimated from the seismic data, where the problem of estimating these propagation velocities (along with a density model) from the full seismic waveforms is referred to as full-waveform inversion (FWI) [10, 76]. Typically a smoothed version of the velocity model is obtained and then used to solve the linear seismic imaging problem to resolve the discontinuities in the image. However, it would be of interest to be able to apply our methodology for estimating the edge strengths directly to the FWI problem, in hopes of better resolving discontinuities and sharp contrasts in the velocity model (and thereby obviating the need to afterwards solve the imaging problem). FWI, however, is a highly non-linear inverse problem, and thus our algorithm and methodology for estimating the edge strengths is not immediately applicable. Hence, another interesting and useful direction for future work is extending the methodology and algorithms from Chapters 4-5 to non-linear inverse problems.

Time-lapse Seismic Inversion with Multiple Monitor Surveys

A straight-forward extension of the work of Chapter 6 would address the case where one is interested in monitoring the changes in the subsurface as a function of time. Here, in addition to a baseline seismic survey, the data would consist of multiple monitor surveys (with possibly differing acquisition geometries) taken at regular time intervals over a prolonged period (e.g. a seismic survey might be performed every day over the course of one year). Standard approaches from Bayesian inference for time-series analysis, such as hidden Markov models, can be used to extend the hierarchical Bayesian methodology of Chapter 6 to the multiple survey case.

Dynamic Survey Design and Optimization

In the context of the above described time-lapse seismic inversion problem with multiple surveys, since the seismic surveys are repeatedly taken over time, the question arises of how best to design the next survey (i.e. which acquisition geometry should be used). In particular, if the goal is to monitor and track changes in the subsurface, the survey designer may wish to optimize the subsequent survey geometry to best infer the time-evolving change in the subsurface. This dynamic optimization problem provides a further useful extension of the work of Chapter 6.

7.3 Final Remarks

The research presented in this thesis is among a small, but growing number of studies exploring hierarchical Bayesian approaches in geophysics [7, 9, 43, 44, 45]. The results of our research are encouraging and indicate promising directions for further application of Bayesian inference in geophysical inverse problems. We note that one downside of Bayesian approaches is that they typically involve more computation than standard inversion methods require. For example, at the heart of both the E-M and variational Bayes algorithms applied to LSM (Chapters 4-5) are the alternating updates between the image and edge strengths, where each image update step is essentially a standard LSM inversion. While our research has explored ways to exploit multiple LSM runs to solve a Bayesian inference problem, in industry scale applications even a single least-squares migration is often considered to be too expensive for practical use [16]. Nevertheless, despite the increased cost, the improvements made to standard inversion methods by hierarchical Bayesian approaches are significant, as we have demonstrated in this thesis. Hence this area remains a fruitful area for future research and application.

Appendix A

Derivation of the Thomsen anisotropy parameters from excess fracture compliance

Here we derive the Thomsen anisotropy parameters for modeling the P-P reflection coefficient in Chapter 3. We use the linear slip model of Schoenberg and Sayers [66] to express the Thomsen anisotropy parameters of the fractured medium in terms of the excess fracture compliance of the medium. The anisotropy parameters can be expressed in terms of the stiffness tensor of the medium **C** as [60]:

$$\delta^{(V)} = \frac{(C_{13}+C_{55})^2-(C_{33}-C_{55})^2}{2C_{33}(C_{33}-C_{55})}, \tag{A.1}$$

$$\epsilon^{(V)} = \frac{C_{11}-C_{33}}{2C_{33}}, \tag{A.2}$$

$$\gamma^{(V)} = \frac{C_{66}-C_{44}}{2C_{44}}. \tag{A.3}$$

We can relate the fracture properties of the medium to the stiffness tensor by computing the excess compliance tensor of the fractures $\mathbf{S}_{\text{frac}}$ (which is the contribution of the fractures to the overall medium compliance tensor). Schoenberg and Sayers [66] show that, under the simplifying assumption that the behavior of the fracture system is invariant with respect to rotation about the axis normal to the fractures, the excess compliance tensor of the fractures is given by

$$\mathbf{S}_{\text{frac}} = \begin{bmatrix} Z_N & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & Z_T & 0 \\ 0 & 0 & 0 & 0 & 0 & Z_T \end{bmatrix} \tag{A.4}$$

where Z_N and Z_T are the excess normal and tangential compliances of the fracture system, respectively. In our analysis, we assumed that the excess normal and tangential compliances of the fracture system are equal. Thus at grid node (i, j), we have $Z_N = Z_T = 10^{z_{ij}}$. Keeping with our convention of treating zero excess compliance as 10^{-13} Pa^{-1}, if $z_{ij} = -13$ then we set $Z_N = Z_T = 0$ (corresponding to the case of no fractures at node (i, j)). Schoenberg and Sayers [66] further show that the overall medium compliance tensor $\mathbf{S}_{\text{tot}}$ can be expressed as the sum of the fracture excess compliance tensor $\mathbf{S}_{\text{frac}}$ and the background compliance tensor $\mathbf{S}_{\text{back}}$, so that

$$\mathbf{S}_{\text{tot}} = \mathbf{S}_{\text{back}} + \mathbf{S}_{\text{frac}}. \tag{A.5}$$

The background compliance tensor is the inverse of background stiffness tensor $\mathbf{C}_{\text{back}}$, which for an isotropic, homogeneous medium is given by [72]:

$$\mathbf{S}_{\text{back}}^{-1} = \mathbf{C}_{\text{back}} = \begin{bmatrix} \lambda + 2\mu & \lambda & \lambda & 0 & 0 & 0 \\ \lambda & \lambda + 2\mu & \lambda & 0 & 0 & 0 \\ \lambda & \lambda & \lambda + 2\mu & 0 & 0 & 0 \\ 0 & 0 & 0 & \mu & 0 & 0 \\ 0 & 0 & 0 & 0 & \mu & 0 \\ 0 & 0 & 0 & 0 & 0 & \mu \end{bmatrix}, \tag{A.6}$$

where $\mu = \rho\beta^2$ and $\lambda = \rho\alpha^2 - 2\mu$ are Lamé's parameters. The overall stiffness tensor of the medium $\mathbf{C}$ is then found as the inverse of the overall compliance tensor [66]:

$$\mathbf{C} = \mathbf{S}_{\text{tot}}^{-1} = \begin{bmatrix} M_b(1-d_N) & \lambda(1-d_N) & \lambda(1-d_N) & 0 & 0 & 0 \\ \lambda(1-d_N) & M_b(1-r_b^2 d_N) & \lambda(1-r_b d_N) & 0 & 0 & 0 \\ \lambda(1-d_N) & \lambda(1-r_b d_N) & M_b(1-r_b^2 d_N) & 0 & 0 & 0 \\ 0 & 0 & 0 & \mu & 0 & 0 \\ 0 & 0 & 0 & 0 & \mu(1-d_T) & 0 \\ 0 & 0 & 0 & 0 & 0 & \mu(1-d_T) \end{bmatrix}, \tag{A.7}$$

where

$$M_b = \lambda + 2\mu, \quad r_b = \frac{\lambda}{M_b}, \quad 0 \le d_T = \frac{Z_T\mu}{1+Z_T\mu} < 1, \quad 0 \le d_N = \frac{Z_N M_b}{1+Z_N M_b} < 1.$$

Combining all of the above gives the anisotropy parameters of the fractured medium at node (i,j), which we denote by $\delta_{z_{ij}}^{(V)}, \gamma_{z_{ij}}^{(V)}, \epsilon_{z_{ij}}^{(V)}$, in terms of the excess fracture compliance. Using (A.7) in (A.1), (A.2), (A.3), and (3.11) gives the forward model for the P-P reflection coefficient as a function of the fracture parameters at node (i,j).

Bibliography

[1] A. Ali and M. Jakobsen. Seismic characterization of reservoirs with multiple fracture sets using velocity and attenuation anisotropy data. *Journal of Applied Geophysics*, 75(3):590 – 602, 2011. ISSN 0926-9851. doi: 10.1016/j.jappgeo.2011.09.003. URL `http://www.sciencedirect.com/science/article/pii/S0926985111001959`.

[2] G. Ayeni and B. Biondi. Target-oriented joint least-squares migration/inversion of time-lapse seismic data sets. *Geophysics*, 75(3):R61–R73, 2010.

[3] M. J. Beal. *Variational Algorithms for Approximate Bayesian Inference*. PhD thesis, Gatsby Computational Neuroscience Unit, University College London, 2003.

[4] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006. ISBN 0387310738.

[5] N. Bleistein. *Mathematical methods for wave phenomena*. Computer science and applied mathematics. Academic Press, 1984. ISBN 9780121056506. URL `http://books.google.com/books?id=mmnvAAAAMAAJ`.

[6] T. Bodin. *Transdimensional Approaches to Geophysical Inverse Problems*. PhD thesis, The Australian National University, 2010.

[7] T. Bodin, M. Sambridge, N. Rawlinson, and P. Arroucau. Transdimensional tomography with unknown data noise. *Geophysical Journal International*, 189 (3):1536–1556, 2012. ISSN 1365-246X. doi: 10.1111/j.1365-246X.2012.05414.x. URL `http://dx.doi.org/10.1111/j.1365-246X.2012.05414.x`.

[8] A. Buland and Y. El Ouair. Bayesian time-lapse inversion. *Geophysics*, 71(3): R43–R48, 2006.

[9] A. Buland and H. Omre. Joint AVO inversion, wavelet estimation and noise-level estimation using a spatially coupled hierarchical Bayesian model. *Geophysical Prospecting*, 51(6):531–550, 2003. ISSN 1365-2478. doi: 10.1046/j.1365-2478.2003.00390.x. URL `http://dx.doi.org/10.1046/j.1365-2478.2003.00390.x`.

[10] C. Bunks, F. Saleck, S. Zaleski, and G. Chavent. Multiscale seismic waveform inversion. *Geophysics*, 60(5):1457–1473, 1995. doi: 10.1190/1.1443880. URL `http://dx.doi.org/10.1190/1.1443880`.

[11] J. Claerbout. *Earth Soundings Analysis: Processing Versus Inversion*. Stanford Exploration project. Blackwell Scientific Publ., 1992. ISBN 9780865422100. URL `http://books.google.com/books?id=ws1qQgAACAAJ`.

[12] M. Clapp. *Imaging under salt: illumination compensation by regularized inversion*. PhD thesis, Stanford University, July 2005.

[13] P. Clifford. Markov random fields in statistics. In G. Grimmett and D. Welsh, editors, *Disorder in Physical Systems: A Volume in Honour of John M. Hammersley*, pages 19–32. Oxford University Press, 1990.

[14] R. Coates and M. Schoenberg. Finite-difference modeling of faults and fractures. *Geophysics*, 60(5):1514–1526, 1995. doi: 10.1190/1.1443884. URL `http://library.seg.org/doi/abs/10.1190/1.1443884`.

[15] R. S. Crosson. Crustal structure modeling of earthquake data: 1. Simultaneous least squares estimation of hypocenter and velocity parameters. *Journal of Geophysical Research*, 81(17):3036–3046, 1976. ISSN 2156-2202. doi: 10.1029/JB081i017p03036. URL `http://dx.doi.org/10.1029/JB081i017p03036`.

[16] W. Dai. *Multisource Least-squares Migration and Prism Wave Reverse Time Migration*. PhD thesis, University of Utah, December 2012.

[17] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977. ISSN 00359246. doi: 10.2307/2984875.

[18] N. Dubos-Sallée and P. Rasolofosaon. Evaluation of fracture parameters and fluid content from seismic and well data. In *SEG Technical Program Expanded Abstracts*, pages 1511–1515, 2008. doi: 10.1190/1.3059201. URL `http://library.seg.org/doi/abs/10.1190/1.3059201`.

[19] B. Duquet, K. J. Marfurt, and J. A. Dellinger. Kirchhoff modeling, inversion for reflectivity, and subsurface illumination. *Geophysics*, 65(4):1195–1209, 2000. doi: 10.1190/1.1444812. URL `http://link.aip.org/link/?GPY/65/1195/1`.

[20] J. Eidsvik, P. Avseth, H. Omre, T. Mukerji, and G. Mavko. Stochastic reservoir characterization using prestack seismic data. *Geophysics*, 69(4):978–993, 2004.

[21] X. Fang, M. Fehler, Z. Zhu, Y. Zheng, and D. Burns. Reservoir fracture characterizations from seismic scattered waves. In *SEG Technical Program Expanded Abstracts 2012*, volume 31, pages 1–6, 2012. doi: 10.1190/segam2012-0813.1. URL `http://library.seg.org/doi/abs/10.1190/segam2012-0813.1`.

[22] X. Fang, M. C. Fehler, Z. Zhu, Y. Zheng, and D. R. Burns. Reservoir fracture characterization from seismic scattered waves. *Geophysical Journal International*, 196(1):481–492, 2014. doi: 10.1093/gji/ggt381. URL http://gji.oxfordjournals.org/content/196/1/481.abstract.

[23] J. Gaiser and R. Van Dok. Green river basin 3-D/3-C case study for fracture characterization: Analysis of PS-wave birefringence. In *SEG Techincal Program Expanded Abstracts*, volume 20, pages 764–767, 2001.

[24] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC Texts in Statistical Science. CRC Press, Boca Raton, FL, USA, third edition, 2013. ISBN 9781439840955.

[25] S. Geoltrain and J. Brac. Can we image complex structures with first-arrival traveltime? *Geophysics*, 58(4):564–575, 1993. doi: 10.1190/1.1443439. URL http://geophysics.geoscienceworld.org/content/58/4/564.abstract.

[26] H. Gjoeystdal and B. Ursin. Inversion of reflection times in three dimensions. *Geophysics*, 46(7):972–973, 1981. doi: 10.1190/1.1441246. URL http://geophysics.geoscienceworld.org/content/46/7/972.abstract.

[27] W. W. Hager and H. Zhang. A survey of nonlinear conjugate gradient methods. *Pacific journal of Optimization*, 2(1):35–58, 2006.

[28] J. Hammersley and P. Clifford. Markov fields on finite graphs and lattices. Unpublished manuscript, 1971.

[29] C. Hanitzsch, J. Schleicher, and P. Hubral. True-amplitude migration of 2D synthetic data. *Geophysical Prospecting*, 42(5):445–462, 1994. ISSN 1365-2478. doi: 10.1111/j.1365-2478.1994.tb00220.x. URL http://dx.doi.org/10.1111/j.1365-2478.1994.tb00220.x.

[30] T. Hong and M. Sen. Joint Bayesian inversion for reservoir characterization and uncertainty quantification. *SEG Technical Program Expanded Abstracts*, 27: 1481–1485, 2008.

[31] J. A. Hudson and J. R. Heritage. The use of the born approximation in seismic scattering problems. *Geophysical Journal of the Royal Astronomical Society*, 66 (1):221–240, 1981. ISSN 1365-246X. doi: 10.1111/j.1365-246X.1981.tb05954.x. URL http://dx.doi.org/10.1111/j.1365-246X.1981.tb05954.x.

[32] M. Hutchinson. A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines. *Communications in Statistics - Simulation and Computation*, 19(2):433–450, 1990. doi: 10.1080/03610919008812866. URL http://dx.doi.org/10.1080/03610919008812866.

[33] J. R. Inman. Resistivity inversion with ridge regression. *Geophysics*, 40 (5):798–817, 1975. doi: 10.1190/1.1440569. URL http://geophysics.geoscienceworld.org/content/40/5/798.abstract.

[34] D. L. B. Jupp and K. Vozoff. Stable iterative methods for the inversion of geophysical data. *Geophysical Journal of the Royal Astronomical Society*, 42(3): 957–976, 1975. ISSN 1365-246X. doi: 10.1111/j.1365-246X.1975.tb06461.x. URL `http://dx.doi.org/10.1111/j.1365-246X.1975.tb06461.x`.

[35] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

[36] H. Kühl and M. D. Sacchi. Least-squares wave-equation migration for AVP/AVA inversion. *Geophysics*, 68(1):262–273, 2003. doi: 10.1190/1.1543212. URL `http://geophysics.geoscienceworld.org/content/68/1/262.abstract`.

[37] G. Lambare, J. Virieux, R. Madariaga, and S. Jin. Iterative asymptotic inversion in the acoustic approximation. *Geophysics*, 57(9):1138–1154, 1992. doi: 10.1190/1.1443328. URL `http://geophysics.geoscienceworld.org/content/57/9/1138.abstract`.

[38] B. Last and K. Kubik. Compact gravity inversion. *Geophysics*, 48(6):713–721, 1983. doi: 10.1190/1.1441501. URL `http://dx.doi.org/10.1190/1.1441501`.

[39] R. LeBras and R. W. Clayton. An iterative inversion of back-scattered acoustic waves. *Geophysics*, 53(4):501–508, 1988. doi: 10.1190/1.1442481. URL `http://geophysics.geoscienceworld.org/content/53/4/501.abstract`.

[40] Y. Li and D. W. Oldenburg. 3-D inversion of gravity data. *Geophysics*, 63(1):109–119, 1998. doi: 10.1190/1.1444302. URL `http://geophysics.geoscienceworld.org/content/63/1/109.abstract`.

[41] E. Liu and A. Martinez. *Seismic Fracture Characterization: Concepts and Practical Applications*. EAGE Publications bv, 2013. ISBN 978-90-73834-40-8.

[42] H. Lynn, S. R. Narhari, S. Al-Ashwak, V. Kidambi, B. Al-Qadeeri, and O. Al-Khaled. PP azimuthal-amplitudes and -acoustic impedance for fractured carbonate reservoir characterization. In *SEG Technical Program Expanded Abstracts*, volume 29, pages 258–262, 2009.

[43] A. Malinverno. A Bayesian criterion for simplicity in inverse problem parametrization. *Geophysical Journal International*, 140(2):267–285, 2000. ISSN 1365-246X. doi: 10.1046/j.1365-246x.2000.00008.x. URL `http://dx.doi.org/10.1046/j.1365-246x.2000.00008.x`.

[44] A. Malinverno. Parsimonious Bayesian Markov chain Monte Carlo inversion in a nonlinear geophysical problem. *Geophysical Journal International*, 151(3): 675–688, 2002. ISSN 1365-246X. doi: 10.1046/j.1365-246X.2002.01847.x. URL `http://dx.doi.org/10.1046/j.1365-246X.2002.01847.x`.

[45] A. Malinverno and V. A. Briggs. Expanded uncertainty quantification in inverse problems: Hierarchical Bayes and empirical Bayes. *Geophysics*, 69(4):

1005–1016, 2004. doi: 10.1190/1.1778243. URL http://link.aip.org/link/?GPY/69/1005/1.

[46] G. J. McLachlan and T. Krishnan. *The EM algorithm and extensions*. Wiley series in probability and statistics. Wiley, Hoboken, NJ, 2nd edition, 2008. ISBN 978-0-471-20170-0. URL http://gso.gbv.de/DB=2.1/CMD?ACT=SRCHA\&SRT=YOP\&IKT=1016\&TRM=ppn+52983362X\&sourceid=fbw_bibsonomy.

[47] M. Metwaly, E. Elawadi, S. Moustafa, and N. Al-Arifi. Combined inversion of electrical resistivity and transient electromagnetic soundings for mapping groundwater contamination plumes in Al Quwy'yia area, Saudi Arabia. *Journal of Environmental and Engineering Geophysics*, 19(1):45–52, 2014. doi: 10.2113/JEEG19.1.45. URL http://library.seg.org/doi/abs/10.2113/JEEG19.1.45.

[48] K. Murphy, Y. Weiss, and M. Jordan. Loopy belief propagation for approximate inference: an empirical study. In *Proceedings of Uncertainty in Artificial Intelligence*, pages 467–475, 1999.

[49] T. Nemeth, C. Wu, and G. T. Schuster. Least-squares migration of incomplete reflection data. *Geophysics*, 64(1):208–221, 1999. doi: 10.1190/1.1444517. URL http://link.aip.org/link/?GPY/64/208/1.

[50] A. Neumaier. Solving ill-conditioned and singular linear systems: A tutorial on regularization. *SIAM Review*, 40:636–666, 1998.

[51] G. Neumann. Determination of lateral inhomogeneities in reflection seismics by inversion of traveltime residuals. *Geophysical Prospecting*, 29(2):161–177, 1981. ISSN 1365-2478. doi: 10.1111/j.1365-2478.1981.tb00399.x. URL http://dx.doi.org/10.1111/j.1365-2478.1981.tb00399.x.

[52] D. Oldenburg. The inversion and interpretation of gravity anomalies. *Geophysics*, 39(4):526–536, 1974. doi: 10.1190/1.1440444. URL http://dx.doi.org/10.1190/1.1440444.

[53] F. Orieux, O. Feron, and J. F. Giovannelli. Sampling high-dimensional Gaussian distributions for general linear inverse problems. *Signal Processing Letters, IEEE*, 19(5):251–254, May 2012. ISSN 1070-9908. doi: 10.1109/LSP.2012.2189104.

[54] M. Oristaglio and M. Worthington. Inversion of surface and borehole electromagnetic data for two-dimensional electrical conductivity models*. *Geophysical Prospecting*, 28(4):633–657, 1980. ISSN 1365-2478. doi: 10.1111/j.1365-2478.1980.tb01248.x. URL http://dx.doi.org/10.1111/j.1365-2478.1980.tb01248.x.

[55] J. Pearl. Reverend Bayes on inference engines: A distributed hierarchical approach. In *Proceedings of the Second National Conference on Artificial Intelligence*, pages 133–136, Menlo Park, California, 1982. AAAI, AAAI Press.

[56] J. Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann Publishers Inc., San Francisco, California, 1988.

[57] R.-E. Plessix and W. A. Mulder. Resistivity imaging with controlled-source electromagnetic data: depth and data weighting. *Inverse Problems*, 24(3):034012, 2008. URL `http://stacks.iop.org/0266-5611/24/i=3/a=034012`.

[58] I. Pšenčík and J. Martins. Properties of weak contrast pp reflection/transmission coefficients for weakly anisotropic elastic media. *Studia Geophysica et Geodaetica*, 45(2):176–199, 2001. ISSN 0039-3169. doi: 10.1023/A:1021868328668. URL `http://dx.doi.org/10.1023/A%3A1021868328668`.

[59] W. L. Rodi and S. C. Myers. Computation of traveltime covariances based on stochastic models of velocity heterogeneity. *Geophysical Journal International*, 194(3):1582–1595, 2013.

[60] A. Rüger. Variation of P-wave reflectivity with offset and azimuth in anisotropic media. *Geophysics*, 63(3):935–947, 1998.

[61] A. Rüger. *Reflection coefficients and azimuthal AVO analysis in anisotropic media*. Society of Exploration Geophysicists, 2002. ISBN 78-1-56080-107-8.

[62] C. Sayers. Seismic characterization of reservoirs containing multiple fracture sets. *Geophysical Prospecting*, 57(2):187–192, 2009. ISSN 0016-8025. doi: 10.1111/j.1365-2478.2008.00766.x. URL `http://dx.doi.org/10.1111/j.1365-2478.2008.00766.x`.

[63] C. Sayers and J. Rickett. Azimuthal variation in AVO response for fractured gas sands. *Geophysical Prospecting*, 45:165–182, 1997.

[64] C. M. Sayers, A. D. Taleghani, and J. Adachi. The effect of mineralization on the ratio of normal to tangential compliance of fractures. *Geophysical Prospecting*, 57(3):439–446, 2009. ISSN 1365-2478. doi: 10.1111/j.1365-2478.2008.00746.x. URL `http://dx.doi.org/10.1111/j.1365-2478.2008.00746.x`.

[65] M. Schoenberg and J. Douma. Elastic wave propagation in media with parallel fractures and aligned cracks. *Geophysical Prospecting*, 36:571–590, 1988.

[66] M. Schoenberg and C. Sayers. Seismic anisotropy of fractured rock. *Geophysics*, 60(1):204–211, 1995.

[67] S. Sil and S. Srinivasan. Stochastic simulation of fracture strikes using seismic anisotropy induced velocity anomalies. *Exploration Geophysics*, 40:257–264, 2009.

[68] D. Simpson, F. Lindgren, and H. Rue. In order to make spatial statistics computationally feasible, we need to forget about the covariance function. *Environmetrics*, 23(1):65–74, February 2012. doi: 10.1002/env.1137. URL `http://opus.bath.ac.uk/32282/`.

[69] R. Snieder and J. Trampert. Inverse problems in geophysics. In A. Wirgin, editor, *Wavefield inversion*, pages 119–190. Springer Verlag, New York, 1999.

[70] A. Tarantola. *Inverse Problem Theory and Methods for Model Parameter Estimation.* Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2004. ISBN 0898715725.

[71] L. Thomsen. Weak elastic anisotropy. *Geophysics*, 51(10):1954–1966, 1986.

[72] I. Tsvankin. *Seismic signatures and analysis of reflection data in anisotropic media*, volume 29 of *Handbook of Geophysical Exploration. Seismic Exploration.* Elsevier, Amsterdam, 2001.

[73] B. Ursin and G. U. Haugen. Weak-contrast approximation of the elastic scattering matrix in anisotropic media. *pure and applied geophysics*, 148(3-4): 685–714, 1996. ISSN 0033-4553. doi: 10.1007/BF00874584. URL `http://dx.doi.org/10.1007/BF00874584`.

[74] J. P. Verdon and A. Wüstefeld. Measurement of the normal/tangential fracture compliance ratio (Z_N/Z_T) during hydraulic fracture stimulation using s-wave splitting data. *Geophysical Prospecting*, 61:461–475, 2013. ISSN 1365-2478. doi: 10.1111/j.1365-2478.2012.01132.x. URL `http://dx.doi.org/10.1111/j.1365-2478.2012.01132.x`.

[75] J. L. Vigneresse. Linear inverse problem in gravity profile interpretations. *Journal of Geophysics*, 43:193–213, 1977.

[76] J. Virieux and S. Operto. An overview of full-waveform inversion in exploration geophysics. *Geophysics*, 74(6):WCC1–WCC26, 2009. doi: 10.1190/1.3238367. URL `http://dx.doi.org/10.1190/1.3238367`.

[77] C. R. Vogel. *Computational Methods for Inverse Problems.* Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2002. ISBN 0898715075.

[78] M. Willis, D. Burns, R. Rao, B. Minsley, M. Toksoz, and L. Vetri. Spatial orientation and distribution of reservoir fractures from scattered seismic energy. *Geophysics*, 71(5):O43–O51, 2006.

[79] J. Zhang and M. N. Toksöz. Nonlinear refraction traveltime tomography. *Geophysics*, 63(5):1726–1737, 1998. doi: 10.1190/1.1444468. URL `http://dx.doi.org/10.1190/1.1444468`.

[80] J. Zhang, U. S. ten Brink, and M. N. Toksöz. Nonlinear refraction and reflection travel time tomography. *Journal of Geophysical Research: Solid Earth*, 103(B12): 29743–29757, 1998. ISSN 2156-2202. doi: 10.1029/98JB01981. URL `http://dx.doi.org/10.1029/98JB01981`.

[81] Z. Zhang and L. Huang. Double-difference elastic-waveform inversion with prior information for time-lapse monitoring. *Geophysics*, 78(6):R259–R273, 2013.

[82] Y. Zheng, X. Fang, M. C. Fehler, and D. R. Burns. Seismic characterization of fractured reservoirs by focusing Gaussian beams. *Geophysics*, 78(4): A23–A28, 2013. doi: 10.1190/geo2012-0512.1. URL http://geophysics.geoscienceworld.org/content/78/4/A23.abstract.

[83] X. Zhu, P. Valasek, B. Roy, S. Shaw, J. Howell, S. Whitney, N. D. Whitmore, and P. Anno. Recent applications of turning-ray tomography. *Geophysics*, 73 (5):VE243–VE254, 2008. doi: 10.1190/1.2957894. URL http://geophysics.geoscienceworld.org/content/73/5/VE243.abstract.