Source estimation and uncertainty quantification for wave-equation based seismic imaging and inversion

by

Zhilong Fang

B.Math., Tsinghua University, 2010

M.Math., Tsinghua University, 2012

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

Doctor of Philosophy

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(Geophysics)

The University of British Columbia

(Vancouver)

February 2018

Abstract

In modern seismic exploration, wave-equation-based inversion and imaging approaches are widely employed for their potential of creating high-resolution subsurface images from seismic data by using the wave equation to describe the underlying physical model of wave propagation. Despite their successful practical applications, some key issues remain unsolved, including local minima, unknown sources, and the largely missing uncertainty analyses for the inversion. This thesis aims to address the following two aspects: to perform the inversion without prior knowledge of sources, and to quantify uncertainties in the inversion.

The unknown source can hinder the success of wave-equation-based approaches. A simple time shift in the source can lead to misplaced reflectors in linearized inversions or large disturbances in nonlinear problems. Unfortunately, accurate sources are typically unknown in real problems. The first major contribution of this thesis is, given the fact that the wave equation linearly depends on the sources, I have proposed on-the-fly source estimation techniques for the following wave-equation-based approaches: (1) time-domain sparsity-promoting least-squares reverse-time migration; and (2) wavefield-reconstruction inversion. Considering the linear dependence of the wave equation on the sources, I project out the sources by solving a linear least-squares problem, which enables us to conduct successful wave-equation-based inversions without prior knowledge of the sources.

Wave-equation-based approaches also produce uncertainties in the resulting velocity model due to the noisy data, which would influence the subsequent exploration and financial decisions. The difficulties related to practical uncertainty quantification lie in: (1) expensive computation related to wave-equation solves, and (2) the nonlinear parameter-to-data map. The second major contribution of this thesis is the proposal of a computationally feasible Bayesian framework to analyze uncertainties in the resulting velocity models. Through relaxing the wave-equation constraints, I obtain a less nonlinear parameter-to-data map and a posterior distribution that can be adequately approximated by a Gaussian distribution. I derive an implicit formulation to construct the covariance matrix of the Gaussian distribution, which allows us to sample the Gaussian distribution in a computationally efficient manner. I demonstrate that the proposed Bayesian framework can provide adequately accurate uncertainty analyses for intermediate to large-scale problems with an acceptable computational cost.

Lay summary

Geophysical prospecting uses sound waves to reveal subsurface structures by detecting differences in the wave-speed of various earth media. I use the mathematical tool called wave equation to simulate the physical process of wave propagation in the Earth, and make the following two contributions in this thesis by addressing some of key issues in using this tool. First, the exact waveform of the sound wave is usually not known, and therefore affects the accuracy of wave simulation. I have proposed an approach to estimate this waveform using observation of sound waves. Second, noise in the observed sound-wave data can lead to uncertainties in the subsurface structures, and to further risks in business decisions regarding oil-well deployment. To address this issue, I propose a method to quantify the uncertainties resulted from the noise in the data.

Preface

All of my thesis work herein presented is carried out under the supervision of Dr. Felix J. Herrmann, in the Seismic Laboratory for Imaging and Modeling at the University of British Columbia.

I prepared Chapter 1 .

A revised version of Chapter 2 has been published [Mengmeng Yang, Philipp A. Witte, Zhilong Fang, and Felix J. Herrmann, 2016, Time-domain sparsity-promoting least-squares migration with source estimation, *in* SEG Technical Program Expanded Abstracts 2016, p. 4225-4229]. The manuscript was written jointly while Mengmeng Yang was the lead investigator and the main manuscript composer. I derived the theory and designed the algorithm for the source estimation. Mengmeng Yang built up the main structure of the time-domain sparsity-promoting least-squares migration. I also contributed to the design and the explaination of the numerical experiment. Mengmeng Yang has granted permission for this chapter to appear in my dissertation.

Chapter 3 contains the text and figures excerpted from a submitted paper [Zhilong Fang, Rongrong Wang, and Felix J. Herrmann, Source estimation for wavefield-reconstruction inversion]. I was the lead investigator and the manuscript composer.

A short version of Chapter 4 has been published [Zhilong Fang and Felix J. Herrmann, Source estimation for Wavefield Reconstruction Inversion, 77th EAGE Conference & Exhibition, 2015, Madrid, Spain], and a revised full version has been acceptted by Journal of Geophysics [Zhilong Fang, Rongrong Wang, and Felix J. Herrmann, Source estimation for wavefield-reconstruction inversion]. I was the lead investigator and the manuscript composer.

A revised version of Chapter 5 has been reported [Bas Peters, Zhilong Fang, Brendan R. Smithyman, and Felix J. Herrmann, 2015, Regularizing waveform inversion by projections onto convex sets - application to the 2D Chevron 2014 synthetic blind-test dataset (Research Report number: TR-EOAS-2015-7)]. The manuscript was written jointly with Bas Peters, Brendan R. Smithyman, and Felix J. Herrmann. I was the lead investigator and contributed to the source estimation part in the algorithm, carried out all the experiments, and produced all the figures. I wrote the sections of source estimation and numerical experiments. Bas Peters and Brendan R. Smithyman designed and implemented the algorithm of optimization with convex constraints. Bas Peters and Felix J. Herrmann wrote the section of the optimization with convex constraints. Given the equivalent contributions made by Bas Peters and I, Bas Peters has granted permission for this chapter to appear in my dissertation.

A short version of Chapter 6 has been published [Zhilong Fang, Chia Ying Lee, Curt Da Silva, Tristan van Leeuwen, and Felix J. Herrmann, Uncertainty quantification for Wavefield Reconstruction Inversion using a PDE free semidefinite Hessian and randomize-then-optimize method, *in* SEG Technical Program Expanded Abstracts 2016, p. 1390-1394], and a revised full version has been submitted [Zhilong Fang, Curt Da Silva, Rachel Kuske, and Felix J. Herrmann, Uncertainty quantification for inverse problems with weak PDE-constraints]. I was the lead investigator and the manuscript composer.

Table of Contents

Abstract ii

Lay summary iv

Preface v

Table of Contents vii

List of Tables x

List of Figures xi

Glossary xvii

Acknowledgements xviii

1 Introduction 1
- 1.1 Least-squares reverse-time migration 7
- 1.2 Full-waveform inversion and wavefield-reconstruction inversion 11
- 1.3 Uncertainty quantification 14
- 1.4 Objectives 18
- 1.5 Thesis outline 19

2 Source estimation for time-domain sparsity-promoting least-squares reverse-time migration 21
- 2.1 Introduction 21
- 2.2 Theory 23
  - 2.2.1 Imaging with linearized Bregman 23
  - 2.2.2 On-the-fly source estimation 25
- 2.3 Numerical experiments 27
  - 2.3.1 RTM with true and wrong source functions 27

2.3.2 LS-RTM with linearized Bregman 29
2.3.3 Linearized Bregman with source estimation 31
2.4 Conclusions 36

3 Full-waveform inversion and wavefield-reconstruction inversion 37
3.1 Introduction 37
3.2 Full-waveform inversion 38
3.3 Wavefield-reconstruction inversion 39
3.4 Variable projection method 40
3.5 FWI v.s. WRI 42
3.6 Conclusions 43

4 Source estimation for wavefield-reconstruction inversion . 45
4.1 Introduction 45
4.2 WRI with source estimation 48
4.3 Optimization scheme 52
4.4 Fast solver for WRI with source estimation 53
4.5 Numerical experiments 57
4.5.1 BG model with frequency-domain data 57
4.5.2 BG model with time-domain data 64
4.5.3 BG model with noisy data 64
4.5.4 Comparison with FWI 67
4.6 Discussions 73
4.7 Conclusions 78

5 Wavefield reconstruction inversion with source estimation and convex constraints — application to the 2D Chevron 2014 synthetic blind-test dataset 80
5.1 Introduction 80
5.2 WRI-SE with regularization by projection onto the intersection of convex sets 81
5.3 Practical considerations 86
5.4 Application to the 2D Chevron 2014 blind-test dataset 87
5.5 Discussions 88
5.6 Conclusions 92

6 Uncertainty quantification for inverse problems with weak PDE constraints and its application to seismic wave-equation-based inversions 93

6.1 Introduction 93
6.2 Bayesian framework for inverse problems with a weak PDE-constraint 96
6.2.1 The posterior PDF 97
6.2.2 Selection of λ 102
6.2.3 Sampling method 106
6.3 Uncertainty quantification for seismic wave-equation constrained inverse problems 108
6.3.1 Steps 1 and 2: designing the posterior and prior PDF 109
6.3.2 Steps 3 and 4: Gaussian approximation 110
6.3.3 Steps 5 and 6: sampling the Gaussian PDFs 114
6.3.4 A benchmark method: the randomized maximum likelihood method 115
6.4 Numerical experiments 116
6.4.1 Influence of the penalty parameter λ 116
6.4.2 A 2D layered model 118
6.4.3 BG Model 122
6.5 Discussion 130
6.6 Conclusions 133

7 Conclusion 135
7.1 Source estimation for time-domain sparsity promoting least-squares reverse-time migration 135
7.2 Source estimation for wavefield-reconstruction inversion . . . 136
7.3 Uncertainty quantification for weak wave-equation constrained inverse problems 137
7.4 Current limitations 138
7.5 Future research directions 138

Bibliography 140

List of Tables

Table 4.1 Comparison of the computational costs for different algorithms. 56
Table 4.2 Comparisons between inversion results obtained by WRI-SE-MS and WRI-SE-WS in terms of the final model errors, source errors, and data residuals. 61
Table 4.3 Comparisons between inversion results obtained by WRI-SE-MS and WRI-SE-WS in terms of the final model errors, source errors, and data residuals. 65
Table 4.4 Comparisons between inversion results obtained by WRI-SE-MS and WRI-SE-WS in terms of the final model errors, source errors, and data residuals. 68
Table 4.5 Comparisons between inversion results obtained by FWI and WRI in terms of the final model errors, source errors, and data residuals. 73

Table 6.1 The selection of λ for each frequency 126

List of Figures

Figure 1.1 (a) Schematic of land seismic survey. Image courtesy ION (www.iongeo.com) (b) Schematic of different marine seismic surveys. (Source: Caldwell and Walker, 2011) 3

Figure 1.2 (a) One seismic trace, (b) one shot record, and (c) one data volume. 4

Figure 1.3 The objective of seismic exploration is to reconstruct the subsurface structures from the observed data. 5

Figure 1.4 The subsurface structure can be divided into two parts: (1) the long-wavelength structure (background velocity model) and (2) the short-wavelength structure (velocity perturbation). 6

Figure 1.5 The objective of migration is to map reflection or diffraction seismic data to their true subsurface positions using a smooth background model that is kinematically correct. 8

Figure 1.6 A 1D example to illustrate the impact of the source function to the observed data. The background velocity is 4 km/s. There is an reflector at the depth of 1 km. The wavefield generated by the source function propagates from the surface, is reflected at the depth of 1 km, returns back to the surface, and is recorded as the observed data. 9

Figure 1.7 Reconstruct the 1D reflectivity model by using the correct (red) and wrong (blue) source functions. Clearly, the time shift in the source function yields the location shift of the reflector. 10

Figure 1.8 Schematic FWI workflow. 13

Figure 1.9 (a) We conduct a simple example on a homogeneous model with a constant velocity of 2 km/s. We simulate one shot (★) and receive the data at one receiver (Δ) using a correct source (red) and a wrong source (blue). (b) There is a time shift of 0.1 s between the correct and wrong sources. (c) There is also a time shift of 0.1 s between the data generated by the correct and wrong sources. 15

Figure 1.10 Schematic workflow of the Bayesian statistical inversion. . 17

Figure 2.1 (a) The smooth background model and (b) the true model perturbation modified from Sigsbee2A. 28

Figure 2.2 Comparison of the true (red) and wrong source (blue) functions. 29

Figure 2.3 RTM images with the correct (a) and wrong (b) source functions. 30

Figure 2.4 LS-RTM images with the true (a) and wrong (b) source functions. 32

Figure 2.5 Data residuals and model errors along iterations 33

Figure 2.6 LS-RTM image with the on-the-fly source estimation. . . 34

Figure 2.7 Comparison of the true (red), initial (blue), and recovered (green) source functions. 34

Figure 2.8 Data residuals and model errors along iterations 35

Figure 3.1 Comparison of objective functions of FWI and WRI as a function of v_0. For WRI, we select $\lambda = 10^2$, 10^3, 10^4, and 10^5. Clearly, as λ increases the objective function of WRI converges to that of the reduced formulation. When λ is small, there are no local minima in the objective function of WRI. 43

Figure 4.1 (a) The true model. (b) Comparison of the correct (blue) and wrong (red) source functions. 47

Figure 4.2 (a) Inversion result with the true source function. (b) Inversion result with the wrong source function. 48

Figure 4.3 (a) True model, (b) initial model, and (c) the difference between them. 59

Figure 4.4 (a) Inversion result with the true source function. (b) Inversion result with WRI-SE-MS. (c) Inversion result with WRI-SE-WS. 60

Figure 4.5 (a) Difference between the inversion results using the true source function and the estimated source function by WRI-SE-MS. (b) Difference between the inversion results using the true source function and the estimated source function by WRI-SE-WS. 61
Figure 4.6 Comparison of the true source function (+) and the estimated source function with WRI-SE-MS (×) and WRI-SE-WS (○). (a) Phase comparison. (b) Amplitude comparison. 62
Figure 4.7 Comparison of the relative model errors. 63
Figure 4.8 Comparison of the data residuals corresponding to the inversion results of WRI-SE-MS (dashed line) and WRI-SE-WS (dotted line). 63
Figure 4.9 Inversion results with (a) WRI-SE-MS and (b) WRI-SE-WS. 65
Figure 4.10 Comparison of the true source function (+) and the estimated source function with WRI-SE-MS (×) and WRI-SE-WS (○). (a) Phase comparison. (b) Amplitude comparison. 66
Figure 4.11 Comparison of the data residuals corresponding to the inversion results of WRI-SE-MS (dashed line) and WRI-SE-WS (dotted line). 67
Figure 4.12 Inversion results with (a) WRI-SE-MS and (b) WRI-SE-WS. 68
Figure 4.13 Comparison of the true source function (+) and the estimated source function with WRI-SE-MS (×) and WRI-SE-WS (○). (a) Phase comparison. (b) Amplitude comparison. 69
Figure 4.14 Comparison of the data residuals corresponding to the inversion results of WRI-SE-MS (dashed line) and WRI-SE-WS (dotted line). 70
Figure 4.15 (a)Initial model and (b) difference between the initial model and true models. 71
Figure 4.16 Comparisons between the observed (dashed line) and predicted data (dotted line) computed with the initial model for a source located at the center of the model. (a) Real part comparison. (b) Imaginary part comparison. 72
Figure 4.17 Inversion results with (a) WRI and (b) FWI. 74

Figure 4.18 Vertical profiles of the true model (solid line), initial model (dashdot line), and the inversion results with WRI (dashed line) and FWI (dotted line) at (a) $x = 1000\,\mathrm{m}$, (b) $x = 2500\,\mathrm{m}$, and (c) $x = 4000\,\mathrm{m}$. 75
Figure 4.19 Comparison of the true source function (+) and the estimated source function with FWI (×) and WRI ($\circ$). (a) Phase comparison. (b) Amplitude comparison. 76
Figure 4.20 Comparisons between the observed (dashed line) and predicted data (dotted line) computed with the final results of WRI and FWI. (a) Real part and (b) imaginary part of WRI result. (c) Real part and (d) imaginary part of FWI result. 77

Figure 5.1 The trajectory of Dykstra's algorithm for a toy example with constraints $y \leq 2$ and $x^2 + y^2 \leq 3$. Iterates 5, 6 and 7 coincide; the algorithm converged to the point closest to point number 1 and satisfying both constraints. 84
Figure 5.2 Initial model 88
Figure 5.3 Comparison of inversion result with and without smoothness constraint. (a) Inversion result without smoothness constraint. (b) Inversion result with smoothness constraint. 89
Figure 5.4 Comparison of the well-log velocity (red), initial velocity (blue), and inverted velocities with (black) and without (green) the smoothness constraint. 90
Figure 5.5 Phase (a) and amplitude (b) comparisons between the final estimated source wavelet ($\circ$) and the true source wavelet ($*$) 91

Figure 6.1 (a) Snapshot of the time-domain wavefield generated by a single source ($*$); (b) recorded time-domain data at four receiver locations (∇). 103
Figure 6.2 The four posterior PDFs corresponding to the penalty formulations with $\lambda = 10$ (a), 50 (b), and 250 (c), and the reduced formulation (d). 104
Figure 6.3 The Gaussian approximations of the four posterior PDFs associated with the model in equation 6.21. 107
Figure 6.4 The comparison of the negative log-likelihood functions. . 117

Figure 6.5 The comparison of the negative log-likelihood functions of the true distribution (ψ_1), the approximate distribution (ψ_2), and the Gaussian approximation distribution (ψ_3) with the values of $\lambda^2 = 10^{-10}\mu_1$, $10^{-6}\mu_1$, $10^{-4}\mu_1$, $10^{-2}\mu_1$, $10^{0}\mu_1$, and $10^{2}\mu_1$. 119

Figure 6.6 The true model (a) and the prior mean model (b). 120

Figure 6.7 The posterior mean models obtained by the GARTO method (a) and the RML method (b). 122

Figure 6.8 The comparison between the observed data and the predicted data with the posterior mean models obtained by the GARTO method (a) and the RML method (b). 123

Figure 6.9 The posterior standard deviations obtained by the GARTO method (a) and the RML method (b). 124

Figure 6.10 The mean (line) and 95% confidence interval (background patch) comparisons of the GARTO method (blue) and the RML method (red) at x = 500 m, 1500 m, and 2500 m. The similarity between these two results implies that the confidence intervals obtained with the GARTO method is a good approximation of the ones obtain with the RML method. 125

Figure 6.11 The comparison of the prior marginal distribution (solid line) and the posterior marginal distributions obtained by the GARTO method (dotted line) and the RML method (dashed line) at the locations of (x, z) = (1500m, 200m), (1500m, 700m), and (1500m, 1200m). 126

Figure 6.12 (a) The true model, (b) the prior mean model, and (c) the prior standard deviation. 127

Figure 6.13 (a) The posterior MAP estimate of the model, (b) the posterior standard deviation, and (c) the relative difference between the posterior and the prior standard deviations, i.e., $\frac{\mathrm{STD}_{\mathrm{post}}(m_k)-\mathrm{STD}_{\mathrm{prior}}(m_k)}{|\mathrm{STD}_{\mathrm{prior}}(m_k)|}$. 129

Figure 6.14 The mean and standard deviation comparisons of the posterior (blue) and the prior (red) distributions at x = 1000 m, 2500 m, and 4000 m. 130

Figure 6.15 The comparison of the prior (solid line) and the posterior (dotted line) marginal distributions at the locations of (x, z) = (2240m, 40m), (2240m, 640m), (2240m, 1240m), and (2240m, 1840m). 131

Figure 6.16 The 95% confidence intervals and the 10 realizations via the RML method at $x = 1000\,\mathrm{m}$, $2500\,\mathrm{m}$, and $4000\,\mathrm{m}$. . . 132

Glossary

ADMM - Alternating Direction Method of Multipliers
BPDN - Basis Pursuit Denoise
FWI - Full-Waveform Inversion
LASSO - Least Absolute Shrinkage and Selection Operator
LSM - Least Squares Migration
MAP - Maximum a Posterior
McMC - Markov chain Monte Carlo
PNT - Projected Newton-type
RML - Randomized Maximum Likelihood
RTM - Reverse-Time Migration
RTO - Randomized-Then-Optimize
SE - Source Estimation
STD - Standard Deviation
WRI - Wavefield-Reconstruction Inversion

Acknowledgements

First and foremost, I would like to thank my research supervisor Dr. Felix J. Herrmann. Thank him for providing me with the valuable opportunity to study and work in the world-class geophysical lab—SLIM—during the last six years. All the scientific knowledge and skills that I have learned in SLIM from Dr. Herrmann will benefit me in all aspects of my life. I am deeply grateful for his supervision and help during the last six years.

My gratitude also goes to Eldad Haber, Rachel Kuske, and Michael Friedlander for sitting on my supervisory committee during the last six years. Their support and help are very important for my study and research.

Thanks to all my colleagues at the SLIM lab. I would like to thank Henryk Modzelewski, Miranda Joyce, Diana Ruiz, and Ian Hanlon for their professional assistance. I would also like to thank all students and researchers in SLIM—especially Rongrong Wang, Xiang Li, Ning Tu, Mengmeng Yang, Lina Miao, and Yiming Zhang—for their valuable scientific insights as research collaborators, and for the past six-years company and friendship.

Thanks to Dr. James Gunning for the very insightful technical discussions. I also would like to thank the BG group for providing the synthetic 3D compass velocity model that I used in Chapters 4 and 6. Many thanks to Chevron for providing the 2D synthetic dataset that I used in Chapter 5. Thanks to the authors of the Sigsbee2A model that I used in Chapter 2. Thanks to the authors of IWave, SPGℓ_1, and Madagascar. I also would like to acknowledge the collaboration of the SENAI CIMATEC Supercomputing Center for Industrial Innovation, Bahia, Brazil, and the support of BG Group and the International Inversion Initiative Project.

Thanks to my wife for her company during my Ph.D. study. I am utterly fortunate to have her unconditional support and love, which encourage me to overcome all the obstacles in both my research and life. I would also like to thank my parents for providing their support to my study in Canada during the last six years. Thank them for tolerating me to study and live at such a distant place from their home in Hangzhou, China. I am grateful to my sons—they make my life filling with endless happiness.

Thanks to all funding agencies that supported my research—NSERC, and all SINBAD consortium sponsors. I can not accomplish the research in this thesis without their generous financial support.

To my dear wife Shan.
You make me better everyday.

To my dear Rixi and Yuji.
You give me a new life.
And endless happiness.

To all Ph.D. candidates.
Enjoy your life.
You will make it.

Chapter 1

Introduction

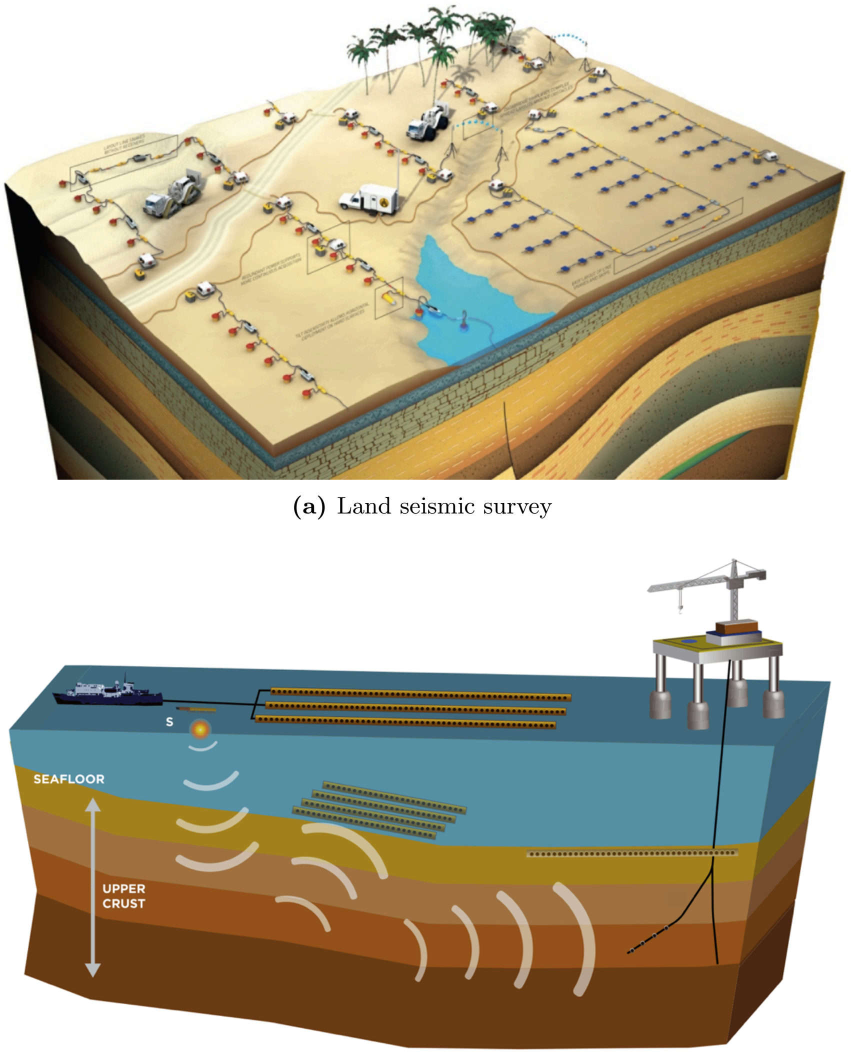
Oil and gas industries rely heavily on images of the subsurface structure to evaluate the location, size, and profitability of a reservoir, as well as to quantify the exploration risk, drilling risk, and volumetric uncertainties. These subsurface images are typically created from different types of geophysical data, ranging from seismic, electrical, electromagnetic, gravitational, and magnetic, which result in different geophysical techniques. Among these methods, seismic exploration is the most important one in terms of capital expenditure because of its higher resolution, higher accuracy, and stronger penetration (Sheriff and Geldart, 1995). Up to now, industries have widely and successfully applied different seismic methods to oil and gas exploration activities around the world (Etgen et al., 2009; Virieux and Operto, 2009).

Seismic exploration technique involves collecting massive volumes of seismic data and processing these data to extract physical properties of subsurface media. During the acquisition of seismic data, an active seismic source, such as a vibroseis truck used on land or an air-gun used offshore, generates acoustic or elastic vibrations that travel from a certain surface location into the Earth, pass through strata with different seismic responses and filtering effects, and return to the surface. The returning waves are detected and recorded as seismic data by receivers including geophones (measure ground motion) or hydrophones (measure wave pressure) (Caldwell and Walker, 2011). Figure 1.1 gives an illustration of standard land and marine seismic data acquisition surveys. The recorded data at each receiver is a time series corresponding to the Earth's response at the receiver location (Figure 1.2a), which is known as the seismic trace. The ensemble of all seismic traces that correspond to the same source experiment forms a shot record (Figure 1.2b), and the ultimate data volume collects hundreds or thousands of shot records (Figure 1.2c). These massive seismic data carry various physical information about subsurface Earth media, and industries aim to extract this information to exploit subsurface oil and gas resources.

After the data acquisition, the next important procedure in seismic exploration is to build up images of subsurface structures from observed seismic data (Figure 1.3). The general workflow of imaging the subsurface structure is a two-step process. The first step is to reconstruct a background velocity model, which describes the long-wavelength characteristics of the subsurface and correctly predicts the kinematics of wave propagation in the true subsurface. The second step is to image the short-wavelength subsurface features—i.e., an image of the reflectivity or the perturbation with respect to the background velocity model—using the velocity model from the first step (Cohen and Bleistein, 1979; Jaiswal et al., 2008). The ensemble of the long and short-wavelength subsurface features (Figure 1.4) provides oil and gas industries with a detailed and informative subsurface image to locate oil and natural gas deposits.

During the last two decades, oil and gas industries have faced the challenge of exploiting reservoirs with complex geology, such as faults, salt bodies, folding, etc (Li, 2017). Therefore, there has been an increasing demand in oil and gas industries for high-resolution images of areas with complicated structures. This demand, along with recent advancements in high-performance computing, has produced a revolution in seismic exploration, as imaging techniques have upgraded from two to three dimensions, from time to depth, from post-stack to pre-stack, and from ray-based methods to wave-equation-based methods. As a result, both academic and industrial efforts have focused on the research of wave-equation-based techniques including full-waveform inversion (Tarantola and Valette, 1982a; Lailly et al., 1983; Pratt, 1999; Virieux and Operto, 2009; Li et al., 2012; van Leeuwen and Herrmann, 2013a; Warner et al., 2013b), reverse-time migration (Baysal et al., 1983; McMechan, 1983; Whitmore, 1983; Etgen et al., 2009), and least-squares reverse-time migration (Aoki and Schuster, 2009; Herrmann and Li, 2012; Tu and Herrmann, 2015b).

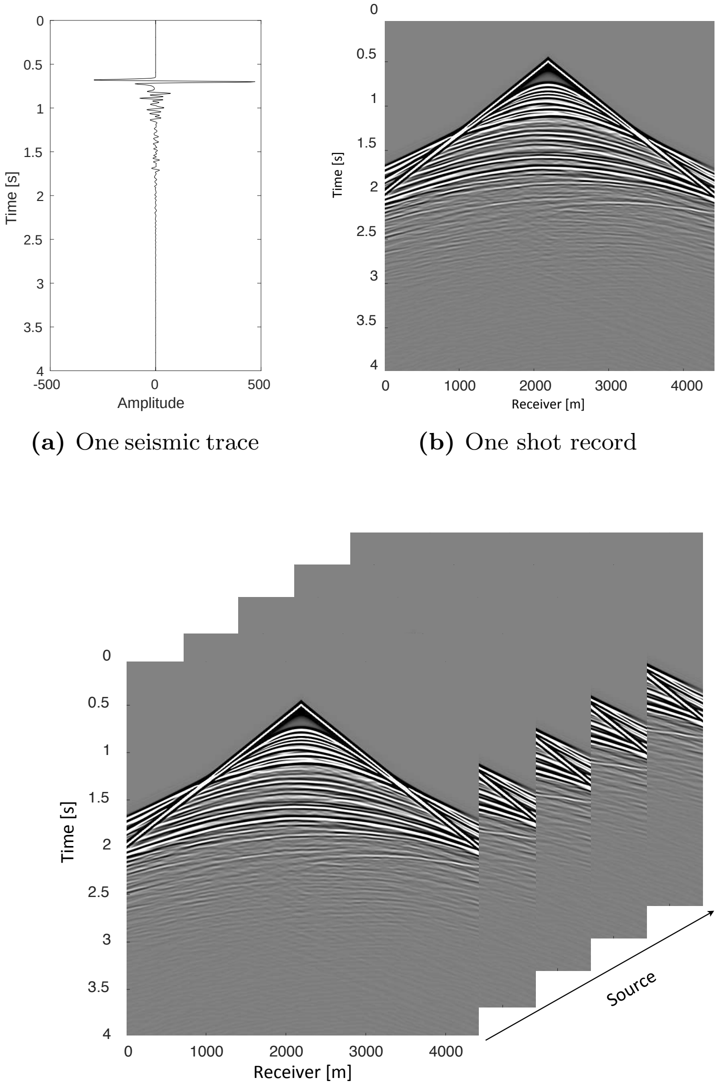
In the oil and gas exploration and production (E&P) business, images of subsurface structures obtained by the aforementioned techniques server as the inputs of the following procedure of structure interpretations, whose results become the inputs to the next step of building up reservoir models. The ultimate stage of this workflow is to analyze financial risks and to make investment decisions (Osypov et al., 2013b). Due to the noise in the observed data and modeling errors, uncertainties exist in images of subsurface structures. Through the workflow, these image uncertainties will propagate to



(a) Land seismic survey

(b) Marine seismic survey

Figure 1.1: (a) Schematic of land seismic survey. Image courtesy ION (www.iongeo.com) (b) Schematic of different marine seismic surveys. (Source: Caldwell and Walker, 2011)



(a) One seismic trace

(b) One shot record

(c) One data volume

Figure 1.2: (a) One seismic trace, (b) one shot record, and (c) one data volume.

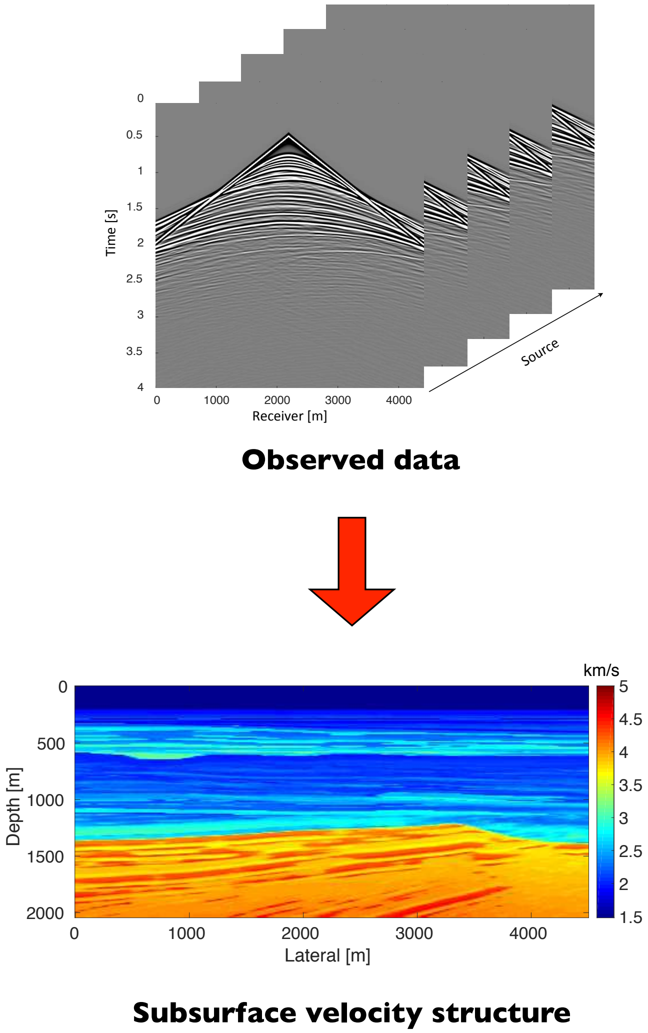
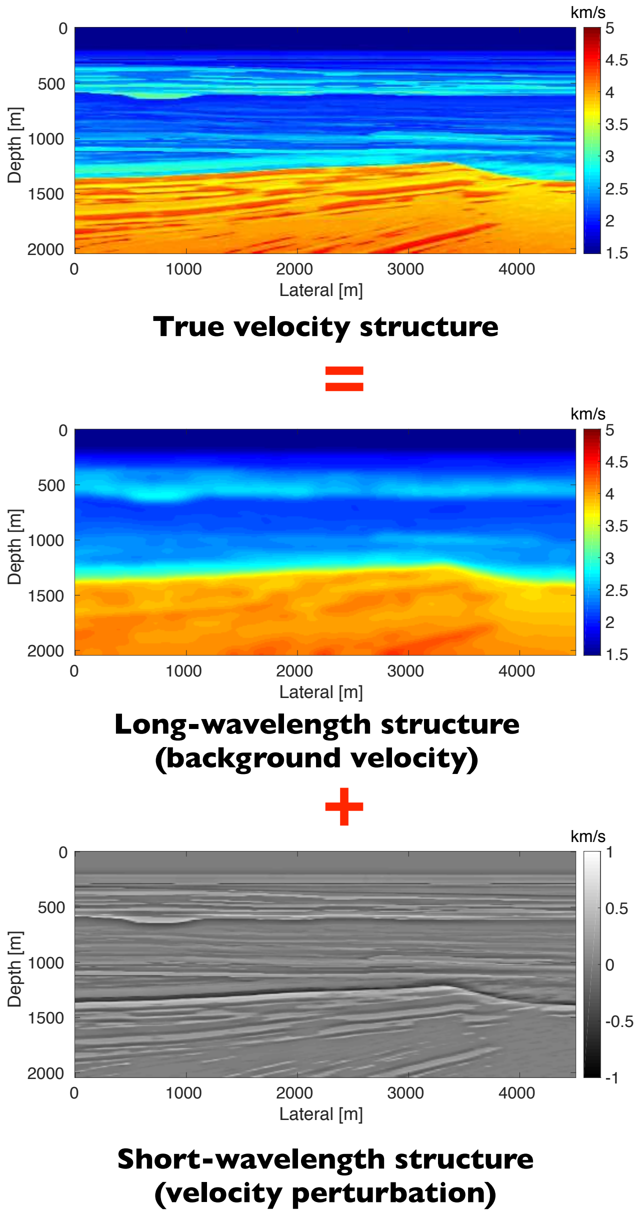


Figure 1.3: The objective of seismic exploration is to reconstruct the subsurface structures from the observed data.



True velocity structure

Long-wavelength structure (background velocity)

Short-wavelength structure (velocity perturbation)

Figure 1.4: The subsurface structure can be divided into two parts: (1) the long-wavelength structure (background velocity model) and (2) the short-wavelength structure (velocity perturbation).

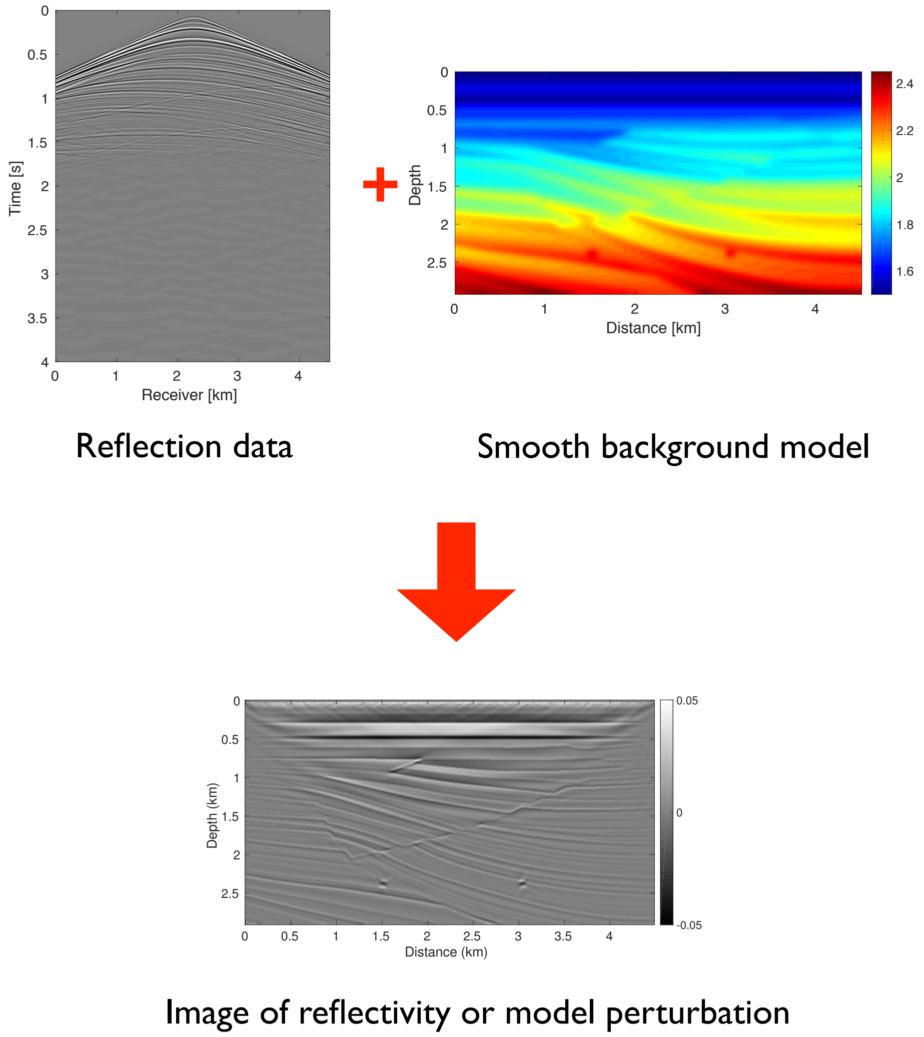
various E&P and financial uncertainties that impact final decision-making. Therefore, understanding and quantifying uncertainties in subsurface images can assist oil and gas industries to have a better understanding of the ultimate financial risks and to make more informed investment decisions (Osypov et al., 2013a).

This thesis first aims to propose robust and computationally efficient wave-equation-based methods that can address practical problems, in which the source functions are typically unknown. Secondly, since it is desirable and important to have a quantitative evaluation of inversion results, this thesis proposes a computationally feasible technique for assessing uncertainties in the velocity model. In order to accomplish these tasks, we will discuss the topics of (1) least-squares reverse-time migration, (2) full-waveform inversion and wavefield-reconstruction inversion, and (3) uncertainty quantification.

1.1 Least-squares reverse-time migration

In the field of seismic exploration, one important and widely-used approach that creates accurate and high-resolution subsurface images is the seismic migration technique. The term migration refers to the process that moves reflection or diffraction seismic energy to their true subsurface positions (Yilmaz, 2001) (Figure 1.5). Conventional migration techniques, including Kirchhoff migration (French, 1975; Schneider, 1978; Biondi, 2001) and reverse-time migration (RTM) (Baysal et al., 1983; Etgen et al., 2009), only utilize the adjoint of the linearized forward modeling operator known as the Born or demigration operator to map the observed data to the image domain. Because the linearized forward operator is not an orthogonal operator in general, these methods do not necessarily produce correct amplitude information on the reflectivity or the perturbations with respect to the background velocity. Moreover, these methods would also produce migration artifacts caused by using the adjoint as an approximated inverse of the linearized operator. Least-squares RTM (LS-RTM) (Aoki and Schuster, 2009; Herrmann and Li, 2012; Tu and Herrmann, 2015b), in contrast, directly inverts the linearized forward operator to reconstruct the true reflectivity or velocity perturbations. As a result, LS-RTM is capable of producing true-amplitude images and mitigating the migration artifacts.

In general, LS-RTM inverts the linearized forward operator by minimizing the ℓ_2-norm of the difference between the observed and predicted reflection data iteratively. Because the predicted and observed data are functions of velocity perturbations and source functions as shown in Figure 1.6, the successful reconstruction of velocity perturbations strongly depends on the

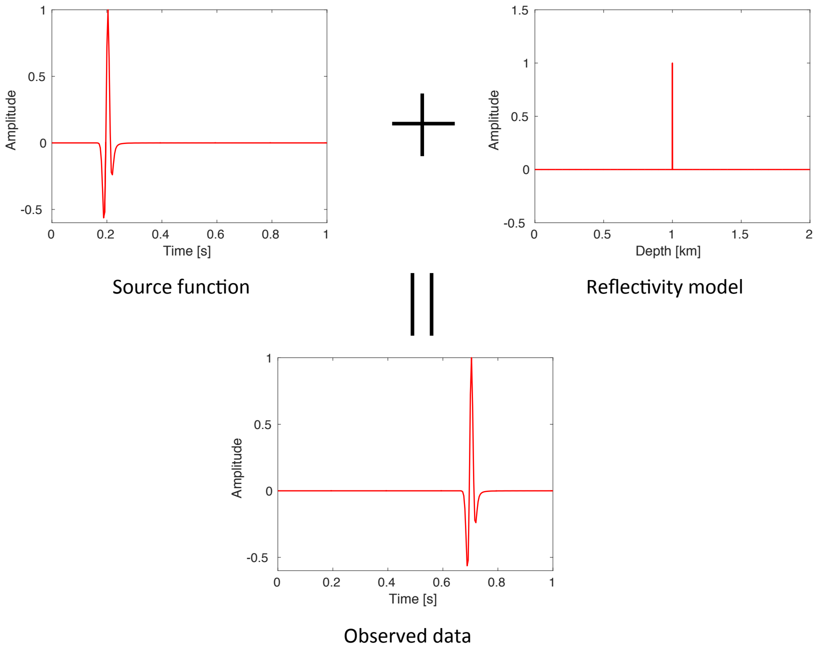


Reflection data

Smooth background model

Image of reflectivity or model perturbation

Figure 1.5: The objective of migration is to map reflection or diffraction seismic data to their true subsurface positions using a smooth background model that is kinematically correct.



Source function

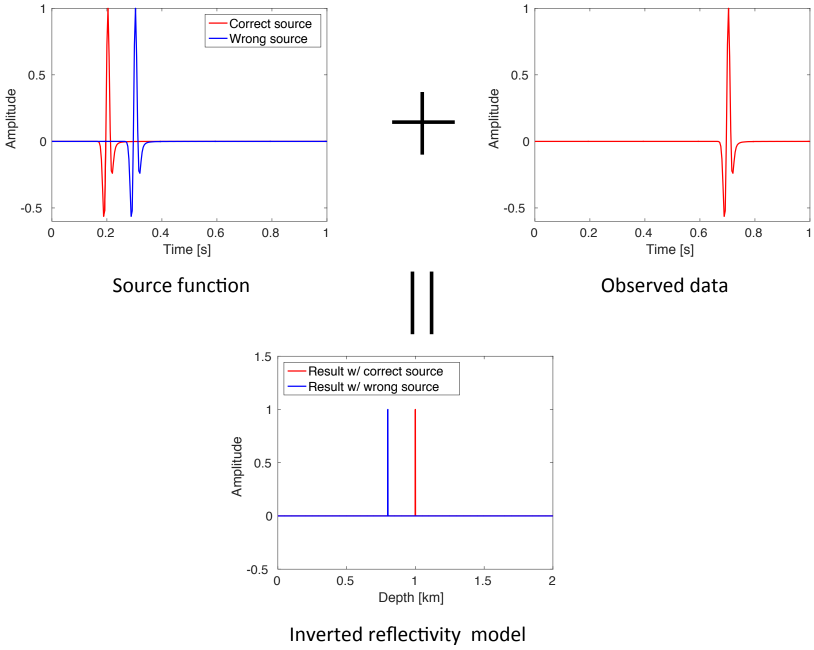
Reflectivity model

Observed data

Figure 1.6: A 1D example to illustrate the impact of the source function to the observed data. The background velocity is 4 km/s. There is an reflector at the depth of 1 km. The wavefield generated by the source function propagates from the surface, is reflected at the depth of 1 km, returns back to the surface, and is recorded as the observed data.

accuracy of source functions. For example, a simple time shift in the source function can lead to a wrongly located reflector as shown in Figure 1.7, and more complicated disturbances in shapes and amplitudes of source functions can result in more unfavorable disturbances in reconstructed images. However, accurate source functions are typically unavailable in practical problems. Therefore, source functions should be estimated along with the inversion of velocity perturbations.

We address the source estimation problem for LS-RTM by considering the fact that the linearized forward modeling operator is linear with respect to the source function. This fact implies that for any fixed velocity perturbation, there is a closed form of the optimal source function that minimizes



Source function

Observed data

Inverted reflectivity model

Figure 1.7: Reconstruct the 1D reflectivity model by using the correct (red) and wrong (blue) source functions. Clearly, the time shift in the source function yields the location shift of the reflector.

the difference between the predicted and observed data. Therefore, we can compute the optimal source function after each update of the velocity perturbation by solving a least-squares linear data fitting problem. However, different from the frequency domain where the source function is a single complex number for each temporal frequency (Tu et al., 2016), the source function is a time series in the time domain. Because both the source function and data are band limited, the solution of the least-squares problem is not unique and unstable. To address this challenge, we regularize the subproblem by simultaneously controlling the oscillations in the source function and the energy of the source function. In this way, we significantly mitigate the nonuniqueness and improve the stability of the estimated source function.

The first contribution of this thesis is that we propose an on-the-fly source estimation technique for the time-domain LS-RTM. During each iteration of LS-RTM, the proposed approach solves an additional regularized least-squares problem to obtain the optimal source function given the current update of the velocity perturbation. We embed the proposed source estimation approach in the recently developed sparsity-promoting LS-RTM (Herrmann et al., 2008; Herrmann and Li, 2012; Tu and Herrmann, 2015b) in the time domain, which produces high-resolution images by combining both stochastic optimization (Bertsekas and Tsitsiklis, 1995; Nemirovski et al., 2009; Shapiro et al., 2009; Haber et al., 2012) and sparsity-promoting optimization (Candès et al., 2006; Donoho, 2006; Candès and Wakin, 2008). The resulting approach is able to create high-resolution images without knowing the accurate source functions.

1.2 Full-waveform inversion and wavefield-reconstruction inversion

The success of an LS-RTM requires an accurate background velocity model that can preserve the travel time kinematics of seismic data. Otherwise, the linearized forward modeling operator cannot generate data at the correct time yielding wrongly positioned reflectors in the final images (Nemeth et al., 1999; Herrmann et al., 2009; Herrmann and Li, 2012). During the past twenty years, many researchers in both industries and academies have reported that the full-waveform inversion technique has the potential capability to produce high-resolution and high-accuracy subsurface velocity models for migrations (Tarantola and Valette, 1982a; Lailly et al., 1983; Pratt, 1999; Virieux and Operto, 2009; Warner et al., 2013b).

Full-waveform inversion is a wave-equation-based technique that aims to obtain the best subsurface velocity model consistent with the observed data and any prior knowledge of the Earth. Through the usage of different types of waves, including diving waves, supercritical reflections, and multiple-scattered waves, FWI possesses the potential capability of reconstructing a high-resolution and high-accuracy subsurface velocity model (Virieux and Operto, 2009). In recent years, many researchers have reported successful applications of FWI to industrial productions and demonstrated its capability in building up velocity models (Virieux and Operto, 2009; Sirgue et al., 2010; Warner et al., 2013b; Vigh et al., 2014).

Most commonly, FWI is formulated as a nonlinear optimization problem that inverts for the unknown velocity model by minimizing the dissimilarities between observed data and predicted data computed by solving wave equations. This formulation is known as the adjoint-state method (Plessix, 2006;

Hinze et al., 2008) and its schematic workflow is shown in Figure 1.8. However, it is well known that the adjoint-state method suffers from local minima in the objective function associated with the well-known cycle-skipping problem (Bunks et al., 1995; Sirgue and Pratt, 2004). More specifically, if the starting model does not produce the predicted data within half a wavelength of the observed data, iterative optimization methods may stagnate at physically meaningless solutions (Warner et al., 2013a; van Leeuwen and Herrmann, 2013b; Huang et al., 2017). As a result, a successful inversion conducted by the adjoint-state method typically requires a good starting model and data containing enough low frequencies and long offsets to prevent cycle skipping from occurring (Bunks et al., 1995; Pratt, 1999; Sirgue and Pratt, 2004; Vigh et al., 2013).

To help mitigate the problem of local minima, van Leeuwen and Herrmann (2013b) and van Leeuwen and Herrmann (2015) proposed a penalty formulation of FWI, which is known as wavefield-reconstruction inversion (WRI). Instead of solving the wave equation during each iteration as the adjoint-state method does, WRI considers the wave-equation as an ℓ_2-norm penalty term in the objective function, which is weighted by a penalty parameter. As a result, compared to the adjoint-state method, WRI enlarges the search space by considering wavefields as additional unknowns. During each iteration, WRI aims to reconstruct wavefields that simultaneously fit both observed data and wave equations, which enables WRI to produce non-cycle-skipped data even for models further from the global optimum. This property provides WRI with the potential to start an inversion with a less accurate initial model and higher-frequency data in comparison with the adjoint-state method (van Leeuwen and Herrmann, 2015; Peters et al., 2014). Indeed, in recent years, the strategy of avoiding cycle-skipped data by enlarging the search space has been also utilized by other extension-based modifications to the conventional FWI technique, including adaptive waveform inversion (AWI) (Warner and Guasch, 2014) and FWI with source-receiver extension (Huang et al., 2017).

Despite the encouraging initial results of WRI, successful applications of WRI also strongly depend on the accuracy of the source function. Any disturbances presented in the source functions will propagate into the predicted data as shown in Figure 1.9. Because the data nonlinearly depends on the velocity model, the additional data misfit introduced by the disturbances in source functions may yield large discrepancies in the recovered velocity model (Pratt, 1999; Zhou and Greenhalgh, 2003; Sun et al., 2014). Moreover, due to the accumulation of errors, these discrepancies can further grow as the depth increases. Hence, accurate source functions play a cru-

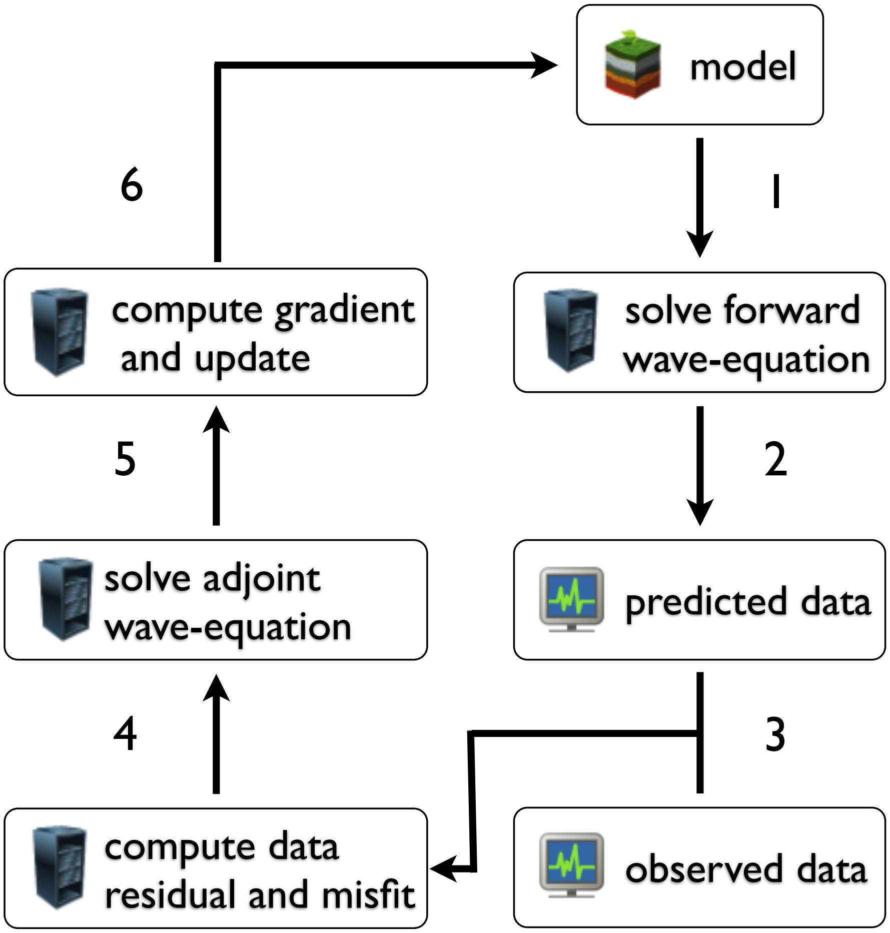

model
1
solve forward wave-equation
2
predicted data
3
observed data
compute data residual and misfit
4
solve adjoint wave-equation
5
compute gradient and update
6

Figure 1.8: Schematic FWI workflow.

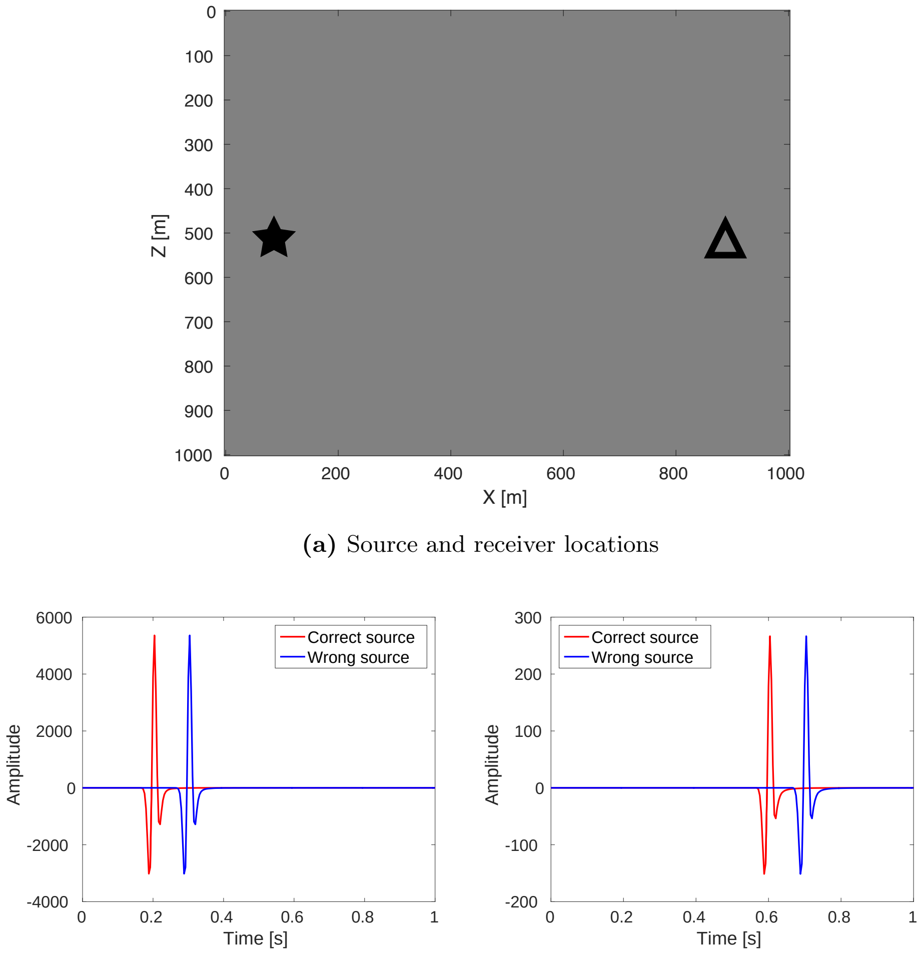
cial role in WRI for the reconstruction of the velocity model. Nevertheless, accurate source functions are generally unknown in practice (Pratt, 1999; Virieux and Operto, 2009; Li et al., 2013). As a result, an embedded procedure of source estimation becomes necessary for WRI to conduct a successful practical application.

The second contribution of this thesis is that it equips WRI with an on-the-fly source estimation technique. As the source functions are also unknown, we face minimizing an objective function with three unknowns, i.e., the velocity model, wavefields, and source functions. We observe that for any fixed velocity, the objective function is quadratic with respect to both wavefields and source functions. Therefore, during each iteration, we apply the so-called variable projection method (Golub and Pereyra, 2003) to simultaneously project out the source functions and wavefields. After the projection, we obtain a reduced objective function that only depends on the velocity model, and we invert for the unknown velocity model by minimizing this reduced objective. Finally, we illustrate that the proposed inversion scheme can recover the unknown velocity model without prior information of the source functions, which enhances the feasibility of WRI to practical problems.

1.3 Uncertainty quantification

The seismic data used by FWI and WRI always contain noise that results from nearby human activities (such as traffic or heavy machinery), animals' movements, winds, and ocean waves, etc. During inversion, these uncertainties in the observed data propagate into the inverted velocity model, yielding uncertainties there. Because the inverted velocity model is the input of the subsequent processing and interpretations, tracking uncertainties in the velocity model can lead to significant improvements in the quantifications of exploration, drilling, and financial risks (Osypov et al., 2013b).

One systematic way to quantify these uncertainties is the Bayesian approach (Kaipio and Somersalo, 2006), which has been applied by geophysical researchers for more than 30 years (Tarantola and Valette, 1982b; Duijndam, 1988; Scales and Tenorio, 2001; Sambridge et al., 2006; Osypov et al., 2008; Ely et al., 2017). The Bayesian approach formulates an inverse problem in the framework of statistical inference, describing all unknowns as random variables and incorporating uncertainties in the observed data, the underlying modeling map, as well as one's prior knowledge on the unknown variables. The randomness of the variables indicates the degree of the belief of their values. The solution of such statistical inverse problem is expressed



(a) Source and receiver locations

(b) Correct source v.s. wrong source

(c) Data w/ correct source v.s. data w/ wrong source

Figure 1.9: (a) We conduct a simple example on a homogeneous model with a constant velocity of 2 km/s. We simulate one shot (★) and receive the data at one receiver (Δ) using a correct source (red) and a wrong source (blue). (b) There is a time shift of 0.1 s between the correct and wrong sources. (c) There is also a time shift of 0.1 s between the data generated by the correct and wrong sources.

in terms of the posterior probability distribution or posterior probability density function (PDF) that incorporates all available statistical information from both the observations in a likelihood distribution and the prior knowledge in a prior distribution. This posterior distribution allows us to extract statistics of interests about the unknown variables either by directly sampling the distribution using approaches like Markov chain Monte Carlo methods (Kaipio and Somersalo, 2006; Matheron, 2012) or by locally approximating the distribution with an easy-to-study Gaussian distribution (Bui-Thanh et al., 2013; Osypov et al., 2013b; Zhu et al., 2016). These extracted statistics provide us with a complete description of the degree of confidence about the unknown variables. Specifically, in seismic exploration, these statistics allow us to assess the uncertainties of the inverted velocity model and identify the areas with high/low reliability in the model. Figure 1.10 shows the standard workflow of the Bayesian statistical inversion.

The key procedure in the Bayesian approach is to explore the posterior distribution. For small-scale problems with a fast simulation driver, we can directly sample the posterior distribution by the best-practice McMC type methods (Cordua et al., 2012; Martin et al., 2012; Ely et al., 2017). However, for wave-equation-based problems, the evaluation of the posterior distribution typically involves computationally expensive wave-equation simulations, and the unknown variables live in a high-dimensional space that arises from the discretization of the model. Moreover, the number of samples required by McMC type methods increases rapidly with the dimensionality (Roberts et al., 2001). All of these facts hinder the application of McMC type methods to large-scale wave-equation-based problems. An alternative approach with computationally low cost is to approximate the distribution with an easy-to-study Gaussian distribution (Bui-Thanh et al., 2013; Osypov et al., 2013b; Zhu et al., 2016). Through exploring the Gaussian distribution, this approach provides an approximated analysis of the uncertainties in the inverted velocity model. Without the requirement of iteratively evaluating the expensive posterior distribution, this approaches can significantly reduce the computational cost. Aside from the gain in the computational cost, the accuracy of the resulting uncertainty analysis strongly relies on the fact that the posterior distribution can be adequately approximated by a Gaussian distribution.

The wave-equation simulation in conventional FWI is strongly nonlinear with respect to the velocity model. As a result, the posterior distribution of the statistical FWI may be only valid in a small neighbor of the maximum a posteriori (MAP) estimator of the posterior distribution (Bui-Thanh et al., 2013; Zhu et al., 2016). Different from conventional FWI, we borrow the

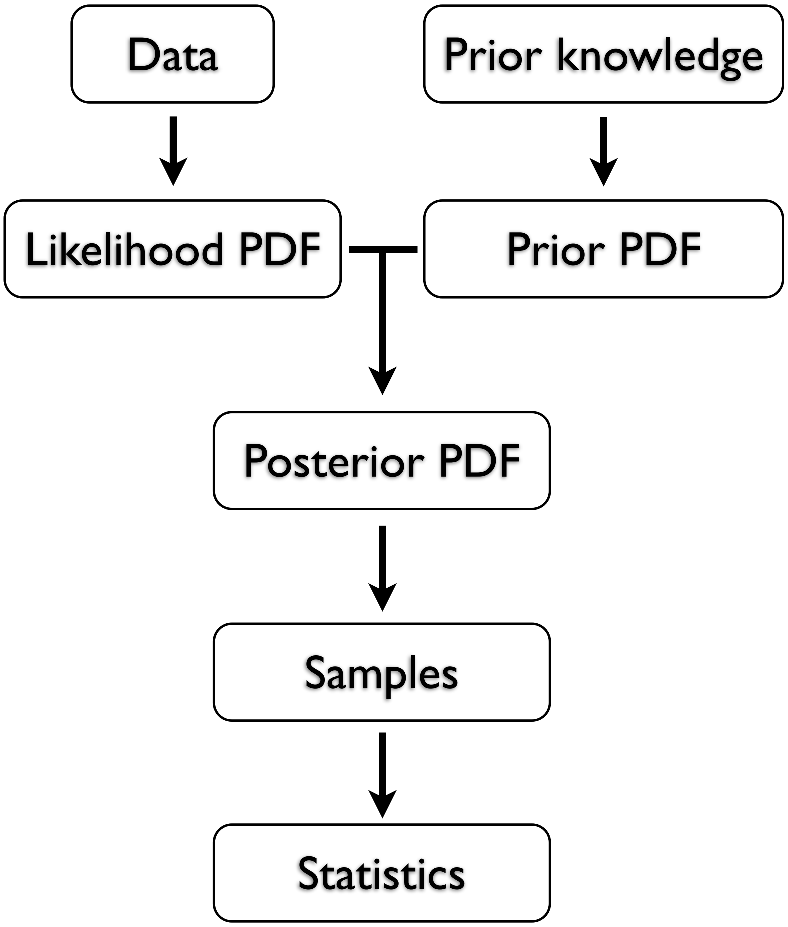


Figure 1.10: Schematic workflow of the Bayesian statistical inversion.

insights from wavefield-reconstruction inversion and propose a new formulation for the posterior distribution to enlarge the region that the Gaussian approximation can hold true. Instead of eliminating the wave-equation constraint by solving the wave equation explicitly, we introduce the wavefields as auxiliary variables and allow for Gaussian deviations in the wave equation. The relaxation of the wave-equation constraint helps weaken the nonlinearity between the velocity model and simulated data. Therefore, the Gaussian approximation of the resulting posterior distribution can be valid in a larger region than that of conventional FWI, which improves the accuracy of the statistics extracted from the Gaussian distribution.

The third main contribution of this thesis, therefore, is that we propose a computationally feasible framework to assess uncertainty in velocity models. Through relaxing the wave-equation constraint, we obtain a posterior distribution that can be adequately approximated by a Gaussian distribution. As a result, we can obtain approximated analyses of uncertainties in velocity models from sampling the Gaussian distribution.

1.4 Objectives

To summarize, we aim to achieve the following objectives:

1. To develop a robust time-domain sparsity promoting LS-RTM with an on-the-fly source estimation. We borrow insights from both stochastic optimization and compressive sensing to reduce the large computational cost of LS-RTM without compromising the image quality. Moreover, we aim to develop an on-the-fly source estimation technique to enhance applications to realistic problems.

2. To develop on-the-fly source estimation technique for WRI. When applied to realistic problems, WRI faces the challenge that the source functions are unknown. We aim to develop an inversion scheme that can remove the demand of accurate source functions.

3. To develop a computationally feasible approach to quantify uncertainties in the unknown velocity model by relaxing the wave-equation constraints. Uncertainty quantification for wave-equation based inverse problem face the challenges that arise from the high-dimensional space and computationally expensive parameter-to-data map with strong nonlinearity. We aim to weaken the nonlinear dependence of the data on the velocity model and to propose a computationally tractable approach to extracting statistical quantities of interests.

1.5 Thesis outline

In chapter 2, we propose an on-the-fly source estimation technique for time-domain LS-RTM. We first introduce the time-domain sparsity promoting LS-RTM as a Basis Pursuit Denoise (BPDN) optimization problem. Then we introduce an easy-to-implement algorithm—the linearized Bregman method—that solves the optimization problem. Following that, we derive the proposed on-the-fly source estimation technique and introduce how to embed it in the linearized Bregman method. Finally, We investigate the effectiveness of the proposed approach using a realistic 2D synthetic example.

In chapter 3, we first introduce the basic theory and implementation of FWI and WRI. Then we describe the variable projection method that is used to solve WRI. We conclude by an intuitive example comparing the objective functions of FWI and WRI, which illustrates the potential advantages of WRI over FWI.

In chapter 4, we propose an on-the-fly source estimation technique for WRI. We first formulate the problem of WRI with source estimation as an optimization problem with respect to the velocity model, wavefields, and source functions. Then we derive the proposed source estimation technique, in which we use the variable projection method to eliminate the dependence of the optimization problem on the wavefields and source functions. Finally, we verify the effectiveness of the proposed source estimation technique by a series of numerical experiments.

In chapter 5, we apply the proposed technique of WRI with the on-the-fly source estimation to the Chevron2014 blind test dataset. We first incorporate the proposed optimization scheme with several regularizations on the velocity model that penalizes unwanted and unphysical events. Then we use the resulting optimization approach to the blind test dataset. We verify the feasibility of the proposed source estimation technique in recovering the source functions and velocity model when addressing real data.

In chapter 6, we propose a computationally efficient approach to quantify uncertainties in the result of wave-equation based approaches. We first explain how to derive a posterior distribution with weak wave-equation constraints. Then, we study how to select the variance for the wave-equation misfit term so that it renders a posterior distribution that can be appropriately approximated by a Gaussian distribution. Following that, we explain how to extract statistical quantities from this posterior distribution with a computationally acceptable cost. We finalize this chapter by presenting several numerical examples to investigate the performance of the proposed uncertainty quantification approach.

In chapter 7, we provide a summary of the thesis. We also discuss the limitation of the approaches provided in the thesis and the possible works in the future.

Chapter 2

Source estimation for time-domain sparsity-promoting least-squares reverse-time migration

2.1 Introduction

In seismic exploration, prestack depth migration techniques, including reverse-time migration (RTM) and least-squares reverse-time migration (LS-RTM), aim to transform the reflection data into a high-resolution image of the interior of the Earth. To accomplish this task, traditional RTM approximates the inverse of the linearized Born modeling operator by its adjoint, which is well known as the migration operator. As a consequence, RTM without extensive preconditioning—generally in the form of image domain scalings and source deconvolution prior to imaging—cannot produce high-resolution and true-amplitude images. By fitting the observed reflection data in the least-squares sense, least-squares migration, and in particular least-squares reverse-time migration (LS-RTM), overcomes these issues except for two main drawbacks withstanding their successful applications. Firstly, the computational cost of conducting demigrations and migrations iteratively is

A version of this chapter has been published in the proceedings of SEG Annual Meeting, 2016, Dallas, USA very large. Secondly, minimizing the ℓ_2-norm of the data residual can lead to model perturbations with artifacts caused by overfitting. These are due to the fact that LS-RTM involves the inversion of a highly overdetermined system of equations where it is easy to overfit noise in the data.

One approach to avoid overfitting is to apply regularizations to the original formulation of LS-RTM and to search for the sparsest possible solution by minimizing the ℓ_1-norm on some sparsifying representation of the image. Motivated by the theory of Compressive Sensing (Donoho, 2006), where sparse signals are recovered from severe undersamplings, and considering the huge cost of LS-RTM, we reduce the size of the overdetermined imaging problem by working with small subsets of sources at each iteration of LS-RTM, bringing down the computational costs.

Herrmann et al. (2008) found that, as a directional frame expansion, curvelets (Ying et al., 2005; Candes et al., 2006) lead to the sparsity of seismic images in imaging problems. This property led to the initial formulation of curvelet-based "Compressive Imaging" (Herrmann and Li, 2012), which formed the bases of later work by Tu et al. (2013) who included surface-related multiples and on-the-fly source estimation via variable projection. While this approach represents major progress, the proposed method, which involves drawing new independent subsets of shots after solving each ℓ_1-norm constrained subproblem, fails to continue to bring down the residual. This results in remaining subsampling related artifacts. In addition, the proposed method relies on a difficult-to-implement ℓ_1-norm solver. To overcome these challenges, Herrmann et al. (2015) motivated by Lorenz et al. (2014) relaxed the ℓ_1-norm objective of Basis Pursuit Denoise (Chen et al., 2001) by an objective given by the sum of the λ-weighted ℓ_1- and ℓ_2-norms. This seemingly innocent change resulted in a greatly simplified implementation that no longer relies on relaxing the ℓ_1-norm constraint and more importantly continues to make progress towards the solution irrespective of the number of randomly selected sources participating in each iteration.

Inspired by this work, we propose to extend our earlier work in the frequency domain to the time domain, which is more appropriate for large-scale 3D problems and to include on-the-fly source estimation via variable projection. Both generalizations are completely new and challenging because the estimation of the time-signature of the source function no longer involves estimating a single complex number for each temporal frequency (Tu et al., 2013).

This chapter is organized as follows. First, we formulate the sparsity-promoting LS-RTM as a Basis Pursuit Denoising problem formulated in the curvelet domain. Next, we relax the ℓ_1-norm and describe the relative simple iterative algorithm that solves this optimization with the mixed ℓ_1- ℓ_2-norm objective. We conclude this theory section by including estimation of the source-time signature that calls for additional regularization to prevent overfitting of the source function. We evaluate the performance of the proposed method on the synthetic Sigsbee2A model (Paffenholz et al., 2002) where we compare our inversion results to LS-RTM with the true source function and standard RTM.

2.2 Theory

2.2.1 Imaging with linearized Bregman

The original sparsity-promoting LS-RTM has the same form, albeit it is overdetermined rather than underdetermined, as a Basis Pursuit Denoise (BPDN, Chen et al. (2001)) problem, which is expressed as below

$$\begin{aligned} &\underset{\mathbf{x}}{\text{minimize}} \ \|\mathbf{x}\|_1 \\ &\text{subject to} \ \sum_i \|\nabla\mathbf{F}_i(\mathbf{m}_0, \mathbf{q}_i)\mathbf{C}^\top\mathbf{x} - \delta\mathbf{d}_i\|_2 \leq \sigma, \end{aligned} \tag{2.1}$$

where the vector $\mathbf{x}$ contains the sparse curvelet coefficients of the unknown model perturbation. The symbols $\|\cdot\|_1$ and $\|\cdot\|_2$ denote ℓ_1- and ℓ_2-norms, respectively. The matrix $\mathbf{C}^\top$ represents the transpose of the curvelet transform $\mathbf{C}$. The vectors $\mathbf{m}_0$ stands for the background squared slowness model and $\mathbf{q}_i = \mathbf{q}_i(t)$ is the time-dependent source function for the i^{th} shot. The vector $\delta\mathbf{d}_i$ represents the observed reflection data along the receivers for the i^{th} shot. The $\sum_i$ runs over all shots. The parameter σ denotes the noise level of the data. The matrix $\nabla\mathbf{F}_i$ represents the linearized Born modeling operator, which is given by:

$$\nabla\mathbf{F}_i = -\mathbf{P}\mathbf{A}(\mathbf{m_0})^{-1}\left(\nabla\mathbf{A}(\mathbf{m_0})\mathbf{u}_i\right), \tag{2.2}$$

where the matrix $\mathbf{A}(\mathbf{m}_0)$ represents the discretized partial differential operator of the time-domain acoustic wave-equation as follows:

$$\mathbf{A}(\mathbf{m_0}) = \Delta - \mathbf{m}_0\frac{\partial^2}{\partial t^2}. \tag{2.3}$$

The vector $\mathbf{u}_i$ is the background forward wavefield given by $\mathbf{u}_i = \mathbf{A}(\mathbf{m_0})^{-1}\mathbf{q}_i$. The matrix $\mathbf{P}$ projects the wavefields onto the receiver locations. The operator $\nabla\mathbf{A}(\mathbf{m_0})\mathbf{u}_i$ denotes the Jacobian matrix.

Despite the success of BPDN in Compressive Sensing, where matrix-vector multiplies are generally cheap, solving equation 2.1 in the seismic setting, where the evaluations of the $\nabla\mathbf{F}_i$'s involve wave-equation solves, is a major challenge because it is computationally unfeasible. To overcome this computational challenge, Herrmann and Li (2012) proposed, motivated by ideas from stochastic gradients (Nemirovski et al., 2009; Shapiro et al., 2009), to select subsets of shots by replacing the sum over $i = 1 \cdots n_s$, with n_s the number of shots, to a sum over randomly selected index sets $\mathcal{I} \subset [1 \cdots n_s]$ of size $n'_s \ll n_s$. By redrawing these subsets at certain points of the ℓ_1-norm minimization, Li et al. (2012) and Tu et al. (2013) were able to greatly reduce the computational costs but at the expense of giving up convergence when the residual becomes small.

For underdetermined problems, Cai et al. (2009) presented a theoretical analysis that proves convergence of a slightly modified problem for iterations that involve arbitrary subsets of rows (= subsets $\mathcal{I}$). Herrmann et al. (2015) adapted this idea to the seismic problem and early results suggested that the work of Cai et al. (2009) can be extended to the overdetermined (seismic) case. With this assumption, we propose to replace the optimization problem in equation 2.1 by

$$\begin{aligned}&\underset{\mathbf{x}}{\text{minimize}}\ \lambda_1\|\mathbf{x}\|_1 + \frac{1}{2}\|\mathbf{x}\|_2^2\\ &\text{subject to}\ \sum_i \|\nabla\mathbf{F}_i(\mathbf{m}_0, \mathbf{q}_i)\mathbf{C}^\top\mathbf{x} - \delta\mathbf{d}_i\|_2 \leq \sigma.\end{aligned} \tag{2.4}$$

The mixed objectives of the above type are known as elastic nets in machine learning and for $\lambda_1 \to \infty$—which in practice means λ_1 large enough—solutions of these modified problems converge to the solution of equation 2.1. While adding the ℓ_2-term may seem an innocent change, it completely changes the iterations of sparsity-promoting algorithms including the role of thresholding (See Herrmann et al. (2015) for more discussion).

Pseudo code illustrating the iterations to solve equation 2.4 are summarized in Algorithm 1. In this algorithm, $t_k = \|\mathbf{A}_k\mathbf{x}_k - \mathbf{b}_k\|_2^2/\|\mathbf{A}_k^\top(\mathbf{A}_k\mathbf{x}_k - \mathbf{b}_k)\|_2^2$ is the dynamic step size, $S_{\lambda_1}(\mathbf{z})$ is the soft thresholding operator (Lorenz et al., 2014), and $\mathcal{P}_\sigma$ projects the residual onto an ℓ_2-norm ball given by the size of the noise σ. To avoid too many iterations, the threshold λ_1, which has nothing to do with the noise level but with the relative importance of the ℓ_1 and ℓ_2-norm objectives, should be small enough. Usually, we set it proportional to the level of the maximum of $\mathbf{z}_k$ to let entries of $\mathbf{z}_k$

enter into the solution.

Algorithm 1 Linearized Bregman for LS-RTM

```
Initialize $\mathbf{x}_0 = \mathbf{0}$, $\mathbf{z}_0 = \mathbf{0}$, $\mathbf{q}$, $\lambda_1$, batchsize $n'_s \ll n_s$
for  $k = 0, 1, \cdots$
    Randomly choose shot subsets $\mathcal{I} \subset [1 \cdots n_s]$, $|\mathcal{I}| = n'_s$
    $\mathbf{A}_k = \{\nabla \mathbf{F}_i(\mathbf{m}_0, \mathbf{q}_i)\mathbf{C}^\top\}_{i \in \mathcal{I}}$
    $\mathbf{b}_k = \{\delta \mathbf{d}_i\}_{i \in \mathcal{I}}$
    $\mathbf{z}_{k+1} = \mathbf{z}_k - t_k \mathbf{A}_k^\top \mathcal{P}_\sigma(\mathbf{A}_k \mathbf{x}_k - \mathbf{b}_k)$
    $\mathbf{x}_{k+1} = S_{\lambda_1}(\mathbf{z}_{k+1})$
end
note: $S_{\lambda_1}(\mathbf{z}_{k+1}) = \text{sign}(\mathbf{z}_{k+1}) \max\{0, |\mathbf{z}_{k+1}| - \lambda_1\}$
        $\mathcal{P}_\sigma(\mathbf{A}_k \mathbf{x}_k - \mathbf{b}_k) = \max\{0, 1 - \frac{\sigma}{\|\mathbf{A}_k \mathbf{x}_k - \mathbf{b}_k\|}\} \cdot (\mathbf{A}_k \mathbf{x}_k - \mathbf{b}_k)$
```

2.2.2 On-the-fly source estimation

Algorithm 1 requires knowledge of the source-time signatures $\mathbf{q}_i$'s to be successfully employed. Unfortunately, this information is typically not available. Following our earlier work on source estimation in time-harmonic imaging and full-waveform inversion (Tu and Herrmann, 2015a; van Leeuwen et al., 2011), we propose an approach where after each model update, we estimate the source-time function by solving a least-squares problem that matches predicted and observed data.

For the source estimation, we assume that we have an initial guess $\mathbf{q}_0 = \mathbf{q}_0(t)$ for the source, which convolved with a filter $\mathbf{w} = \mathbf{w}(t)$ gives the true source function $\mathbf{q}$. By replacing $\mathbf{q}$ in equation 2.4 by the convolution of $\mathbf{w}$ and $\mathbf{q}_0$, we obtain the following joint optimization problem with respect to both $\mathbf{x}$ and $\mathbf{w}$:

$$\begin{aligned} &\underset{\mathbf{x},\mathbf{w}}{\text{minimize}} \ \lambda_1 \|\mathbf{x}\|_1 + \frac{1}{2}\|\mathbf{x}\|_2^2 \\ &\text{subject to} \ \sum_i \|\nabla \mathbf{F}_i(\mathbf{m}_0, \mathbf{w} * \mathbf{q}_0)\mathbf{C}^\top \mathbf{x} - \delta \mathbf{d}_i\|_2 \leq \sigma, \end{aligned} \tag{2.5}$$

where the symbol $*$ stands for the convolution along time. The filter $\mathbf{w}$ is unknown and needs to be estimated in every Bregman step by solving an additional least-squares subproblem. Generally, estimating $\mathbf{w}$ requires numerous recomputations of the linearized data. This can be avoided by assuming that the linearized data for the current estimate of $\mathbf{w}$ is given by modeling the data with the initial source $\mathbf{q}_0$ and convolving with the filter

$\mathbf{w}$ afterward:

$$\nabla\mathbf{F}_i(\mathbf{m}_0, \mathbf{w} * \mathbf{q}_0) \approx \mathbf{w} * \nabla\mathbf{F}_i(\mathbf{m}_0, \mathbf{q}_0), \tag{2.6}$$

where the right hand side stands for the trace by trace convolution of $\mathbf{w}$ with the shot gather traces.

We expect the estimated source signature to decay smoothly to zero with few oscillations within a short duration of time. Therefore in our subproblem, we add a penalty term $\|\mathrm{diag}(\mathbf{r})(\mathbf{w}*\mathbf{q}_0)\|^2$, where $\mathbf{r}$ is a weighting function that penalizes non-zeros entries at later times. In our example, we choose $\mathbf{r}(t)$ to be a logistic loss function (Rosasco et al., 2004)

$$\mathbf{r}(t) = \log(1 + e^{\alpha(t-t_0)}). \tag{2.7}$$

Here, the scalar α defines the speed of the function $\mathbf{r}(t)$ decreases to 0 as $t \to 0$, and we penalize all the events after $t = t_0$. Furthermore, we add a second penalty term $\lambda_2\|\mathbf{w} * \mathbf{q}_0\|^2$ to control the overall energy of the estimated source function. Our subproblem for source estimation is then given by

$$\underset{\mathbf{w}}{\mathrm{minimize}} \sum_i \|\mathbf{w} * \nabla\mathbf{F}_i(\mathbf{m}_0, \mathbf{q}_0) - \delta\mathbf{d}_i\|_2^2 + \|\mathrm{diag}(\mathbf{r})\mathbf{w} * \mathbf{q}_0\|_2^2 + \lambda_2\|\mathbf{w} * \mathbf{q}_0\|_2^2. \tag{2.8}$$

The algorithm to solve the sparsity-promoting LS-RTM with source estimation using linearized Bregman is summarized in Algorithm 2.

Algorithm 2 LB for LS-RTM with source estimation

```
1. Initialize $\mathbf{x}_0 = \mathbf{0}$, $\mathbf{z}_0 = \mathbf{0}$, $\mathbf{q}_0$, $\lambda_1$, $\lambda_2$, batchsize $n_s' \ll n_s$, weights $\mathbf{r}$
2. for $k = 0, 1, \cdots$
3.     Randomly choose shot subsets $\mathcal{I} \subset [1 \cdots n_s], |\mathcal{I}| = n_s'$
4.     $\mathbf{A}_k = \{\nabla\mathbf{F}_i(\mathbf{m}_0, \mathbf{q}_0)\mathbf{C}^\top\}_{i\in\mathcal{I}}$
5.     $\mathbf{b}_k = \{\delta\mathbf{d}_i\}_{i\in\mathcal{I}}$
6.     $\tilde{\mathbf{d}}_k = \mathbf{A}_k\mathbf{x}_k$
7.     $\mathbf{w}_k = \arg\min_{\mathbf{w}} \|\mathbf{w} * \tilde{\mathbf{d}}_k - \mathbf{b}_k\|_2^2 + \|\mathrm{diag}(\mathbf{r})(\mathbf{w} * \mathbf{q}_0)\|_2^2 + \lambda_2\|\mathbf{w} * \mathbf{q}_0\|_2^2$
8.     $\mathbf{z}_{k+1} = \mathbf{z}_k - t_k\mathbf{A}_k^\top\Big(\mathbf{w}_k \star \mathcal{P}_\sigma(\mathbf{w}_k * \tilde{\mathbf{d}}_k - \mathbf{b}_k)\Big)$
9.     $\mathbf{x}_{k+1} = S_{\lambda_1}(\mathbf{z}_{k+1})$
10. end
```

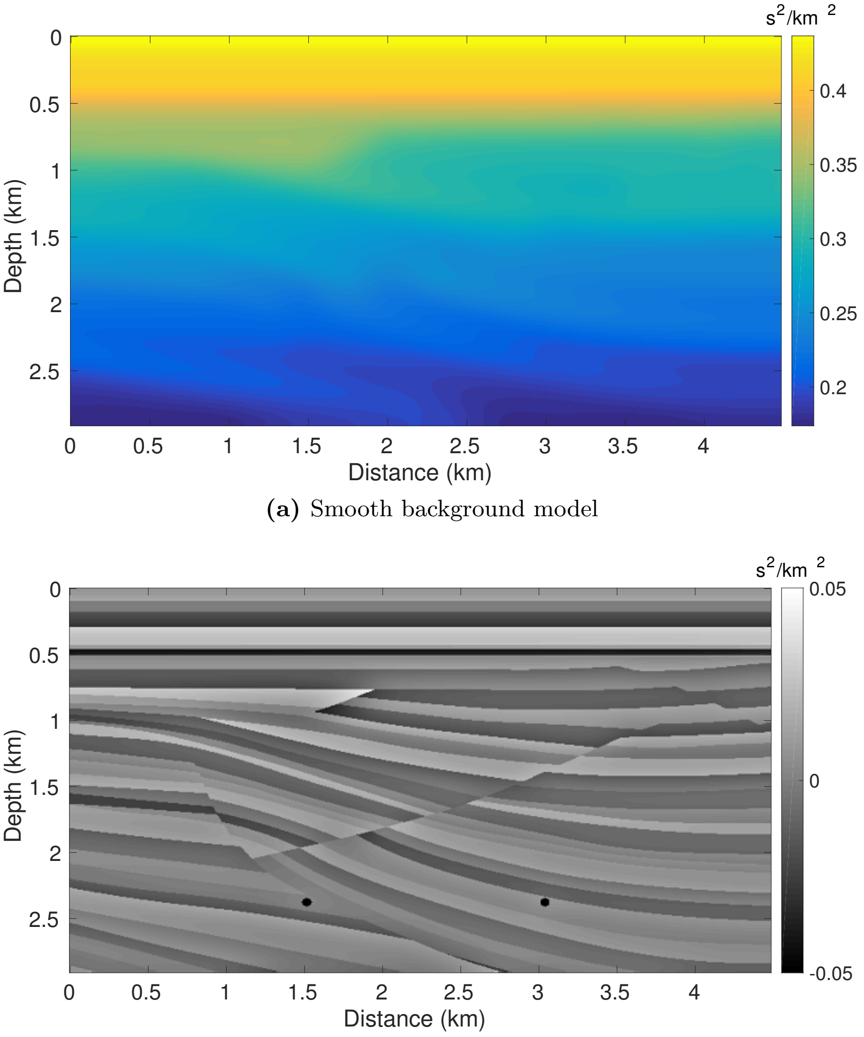
In Algorithm 2, the symbol $\star$ stands for the correlation, which is the adjoint operation of the convolution. Since the initial guess of $\mathbf{x}$ is zero, we omit the filter estimation in the first iteration and update $\mathbf{x}_1$ with the initial guess of the source. In the second iteration, after the filter is estimated for the first time (line 7), $\mathbf{z}_1$ is reset to zero, and $\mathbf{z}_2$ and $\mathbf{x}_2$ are updated according to lines $8-9$. This prevents that the imprint of the initial (wrong) source function pollutes the image updates of later iterations. The subproblem in line 7 can be solved by formulating the optimality condition and solving for $\mathbf{w}_k$ directly.

2.3 Numerical experiments

To test our algorithm, we conduct two types of experiments. First, we test the linearized Bregman method for sparsity-promoting LS-RTM for the case with a known source function $\mathbf{q}$ and a wrong source function $\mathbf{q}_0$. We then compare these results with an example in which the source function is unknown and estimated by the proposed approach. For the experiments, we use the top left part of the Sigsbee2A model, which has a size of 2918m by 4496m and is discretized with a grid size of 7.62 m by 7.62 m. The model perturbation, which is the difference between the true model and a smoothed background model (Figure 2.1a), is shown in Figure 2.1b. We generate linearized and single scattered data using the Born modeling operator. The experiment consists of 295 shot positions with 4 seconds recording time. Each shot is recorded by 295 evenly spaced receivers at a depth of 7.62 m and with a receiver spacing of 15.24 m, yielding a maximum offset of 4 km. We use a source function with a limited frequency spectrum (Figure 2.2) to simulate the data. The source function has a central frequency of 15 Hz with a spectrum ranging from 5 Hz to 35 Hz.

2.3.1 RTM with true and wrong source functions

For later comparisons with the LS-RTM images, we first generate a traditional RTM image (Figure 2.3a) using the true source function. Clearly, the traditional RTM approach creates an image with wrong amplitudes and blurred shapes at interfaces and diffraction points. Furthermore, we also generate an RTM image (Figure 2.3b) using a wrong source function as shown in Figure 2.2. Compared to the true source function, the wrong one contains a spectrum with a flat shape ranging from 12 to 28 Hz and an exponential tapering between 0-12Hz and 28-40 Hz. Furthermore, the phase and shape of the wrong source function also differ from the correct ones. Figure 2.3b illustrates that the usage of the wrong source function enhances the artifacts and destroys the energy's focusing at reflectors and diffraction points.



(a) Smooth background model

(b) Model perturbation

Figure 2.1: (a) The smooth background model and (b) the true model perturbation modified from Sigsbee2A.

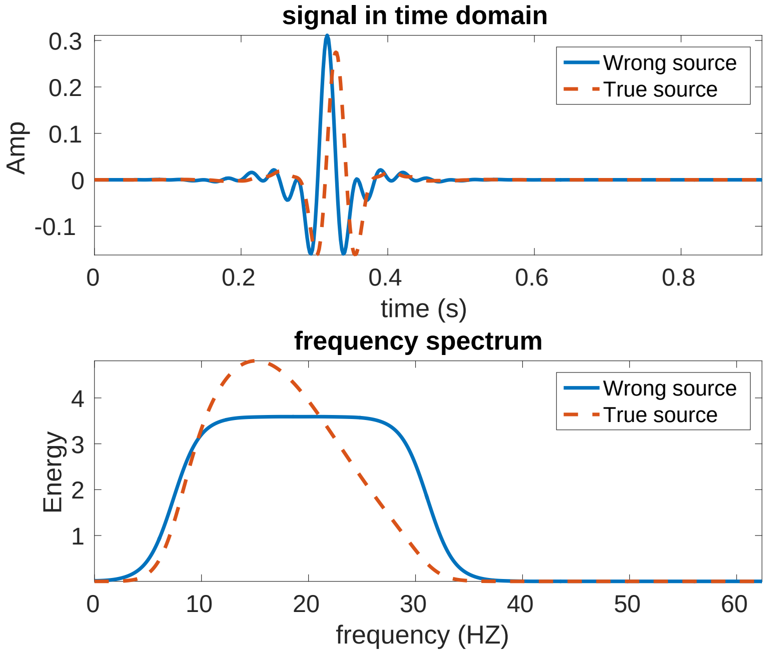
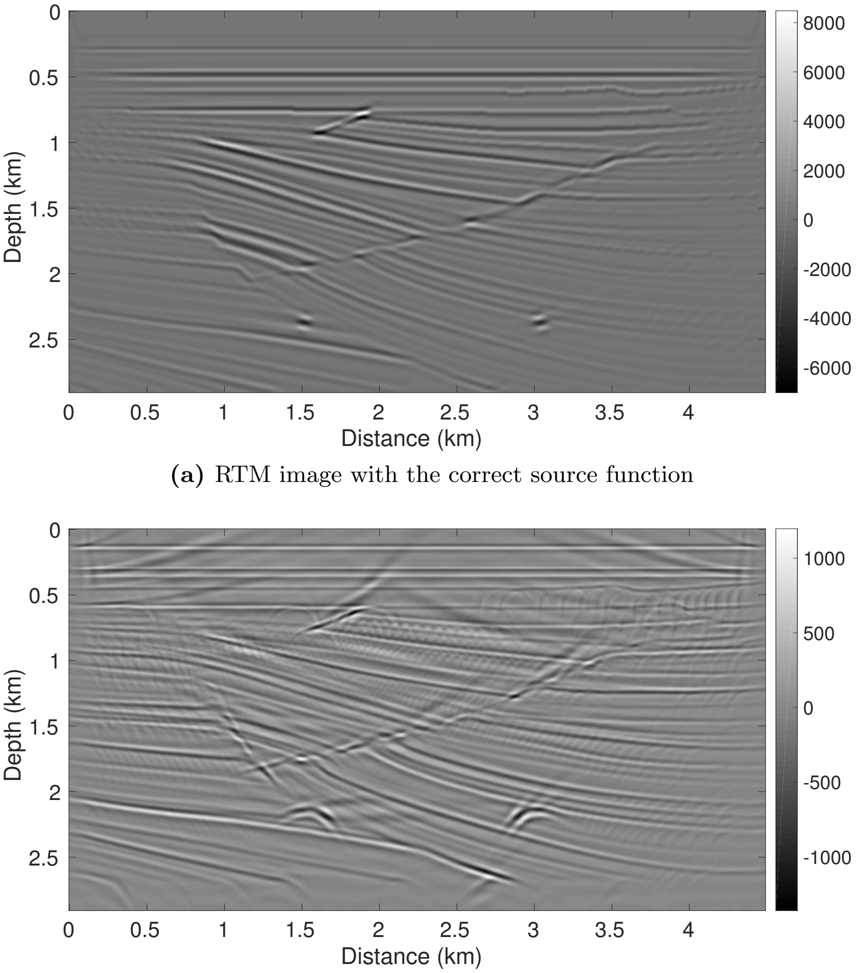


Figure 2.2: Comparison of the true (red) and wrong source (blue) functions.

2.3.2 LS-RTM with linearized Bregman

In this example, we perform LS-RTM via the linearized Bregman method using both the correct source function $\mathbf{q}$ and the wrong source function $\mathbf{q}_0$ shown in Figure 2.2. During the inversion, we perform 40 iterations and draw 8 randomly selected shots in each iteration. Therefore, after 40 iterations, approximately all shot gathers have been used one time (˜one pass through the data). To accelerate the convergence, we apply a depth preconditioner to the model updates, which compensates for the spherical amplitude decay (Herrmann et al., 2008).

The only parameter that needs to be provided by the user for the algorithm is the weighting parameter λ_1 that controls the soft thresholding of the vector $\mathbf{z}$. A large value of λ_1 will threshold too many entries of $\mathbf{z}$ leading to a slower convergence and more iterations, whereas a small λ_1 allows noise and subsampling artifacts to pass the thresholding. In our experiments, we determine λ_1 in the first iteration by setting $\lambda_1 = 0.1 * \mathsf{max}(|\mathbf{z}|)$, which allows roughly 90 percent of the entries in $\mathbf{z}$ to pass the soft thresholding in the



(a) RTM image with the correct source function

(b) RTM image with the wrong source function

Figure 2.3: RTM images with the correct (a) and wrong (b) source functions.

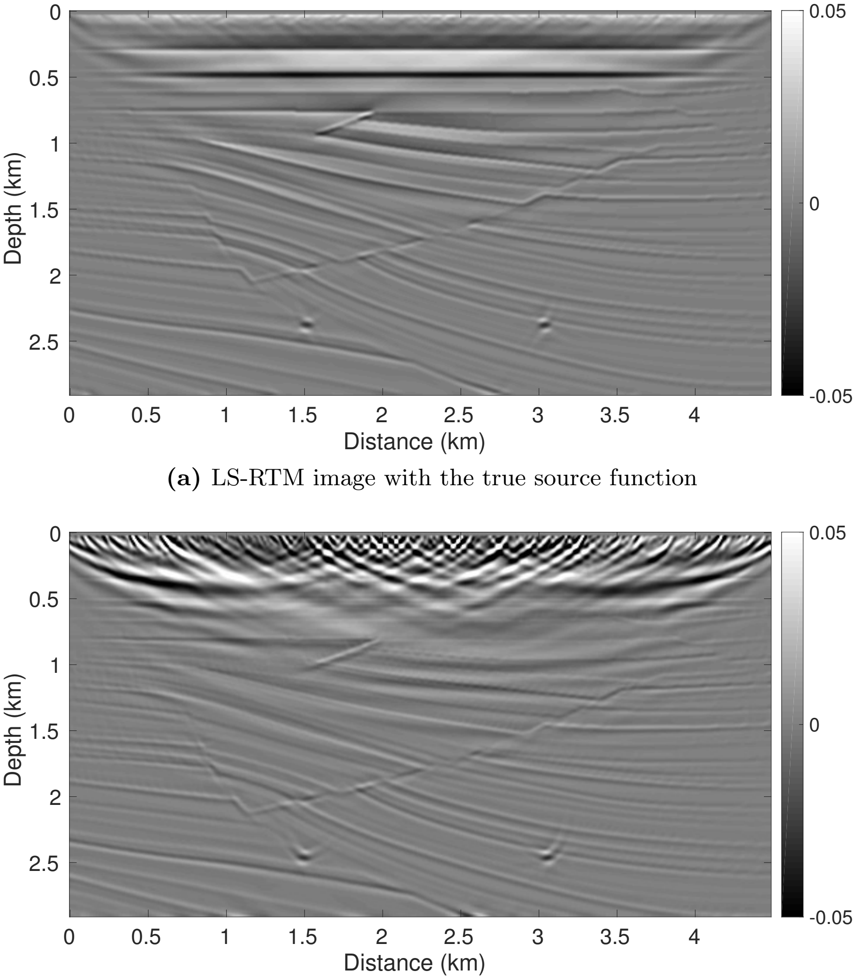
first iteration.

The inversion result of the first experiment with the correct source function is shown in Figure 2.4a. It clearly shows that the faults and reflectors are all accurately imaged. The interfaces are sharp because of the influence of directly inverting the linearized Born modeling operator. On the other hand, when the source function is not correct, the LS-RTM image exhibits strong artifacts, blurred layers, and wrongly located diffractors, as shown in Figure 2.4b. The data residual in the inversion with the correct source function decays as a function of the iteration number (Figure 2.5a), in spite of some jumps occur due to redrawing the shot records in each iteration. On the other hand, when using the wrong source function, the data residual does not decay with the increase of iterations. We can obtain the same observation when comparing the model errors of the two inversion strategies. This example indicates the importance of the correct source function in LS-RTM.

2.3.3 Linearized Bregman with source estimation

To solve the problem with the unknown source function, we utilize the linearized Bregman method with source estimation as described in Algorithm 2. To start the inversion, we select the wrong source function in Figure 2.2 as the initial guess for the source function $\mathbf{q}_0$. To successfully conduct the source estimation, the user also needs to supply the parameter α in the damping function (c.f. equation 2.7). From our experience, once α is big enough, the recovered source is not sensitive to the change of α. In our example, we select $\alpha = 8$ and $\lambda_2 = 1$.

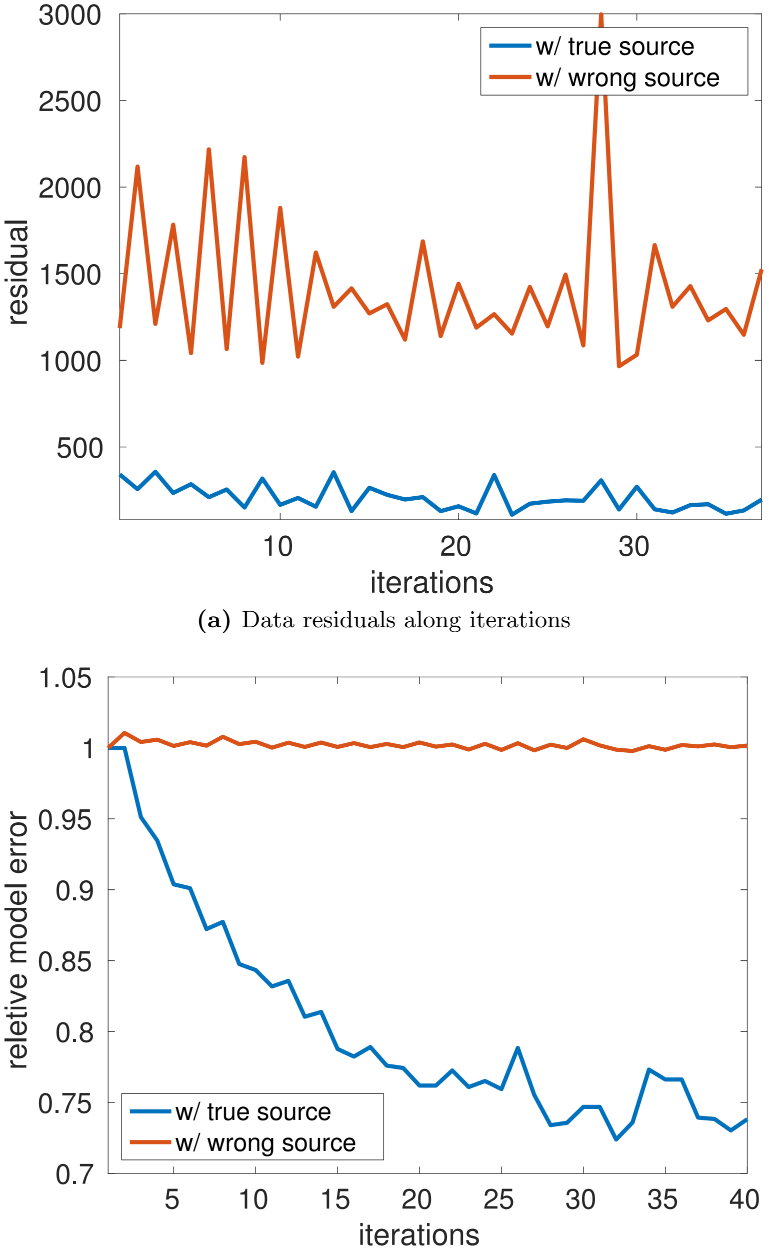
With the same strategy for choosing λ_1 as in the previous example, the inversion result shown in Figure 2.6 is visually as good as the result using the correct source function (Figure 2.4a). Meanwhile, as shown in Figure 2.7, the recovered source function (green dotted line) closely matches the correct source function (red dash-dot line), while some small differences remain in the frequency spectrum. The data residual as a function of the iteration (Figure 2.8a) is similar to the one with the correct source, aside from a slightly larger residual observed at the beginning of the inversion due to the wrong initial source function. We have the same observation for the comparison of the model error history in Figure 2.8b. After 40 iterations, both the final data residual and model error are in the same range as those from the experiment with the correct source function.



(a) LS-RTM image with the true source function

(b) LS-RTM image with the wrong source estimation

Figure 2.4: LS-RTM images with the true (a) and wrong (b) source functions.



(a) Data residuals along iterations

(b) Model errors along iterations

Figure 2.5: Data residuals and model errors along iterations

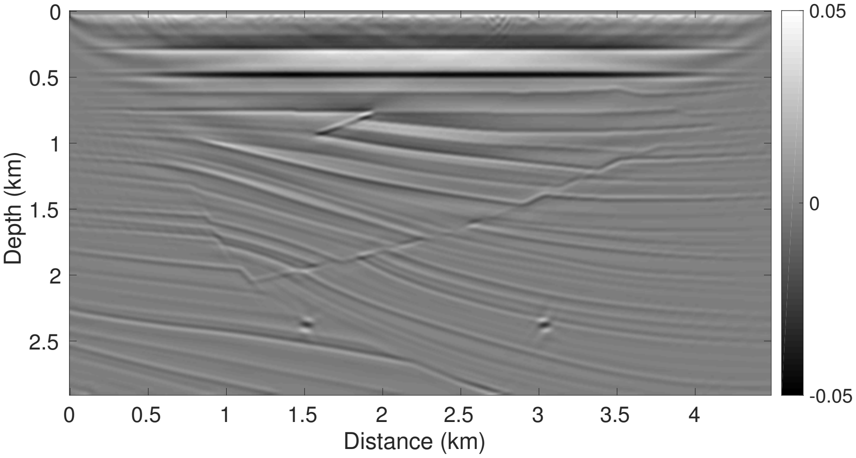


Figure 2.6: LS-RTM image with the on-the-fly source estimation.

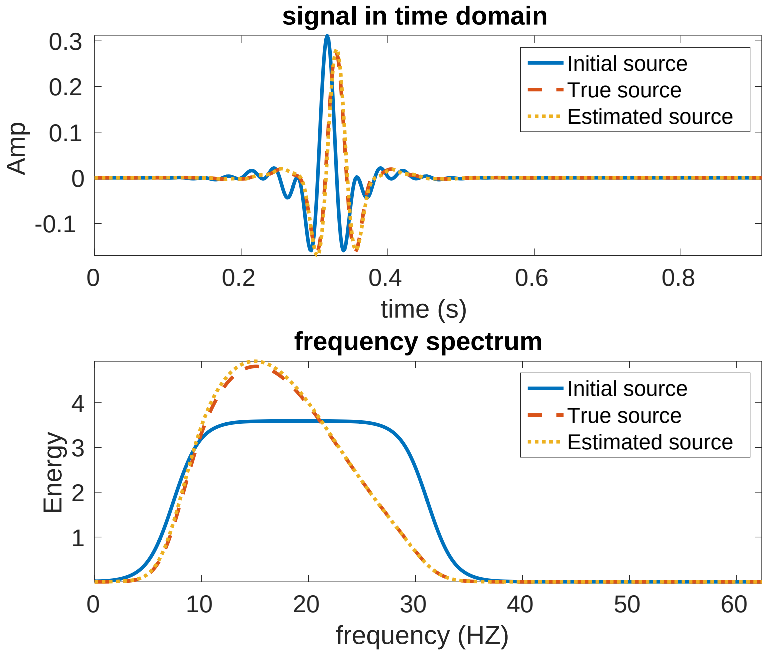
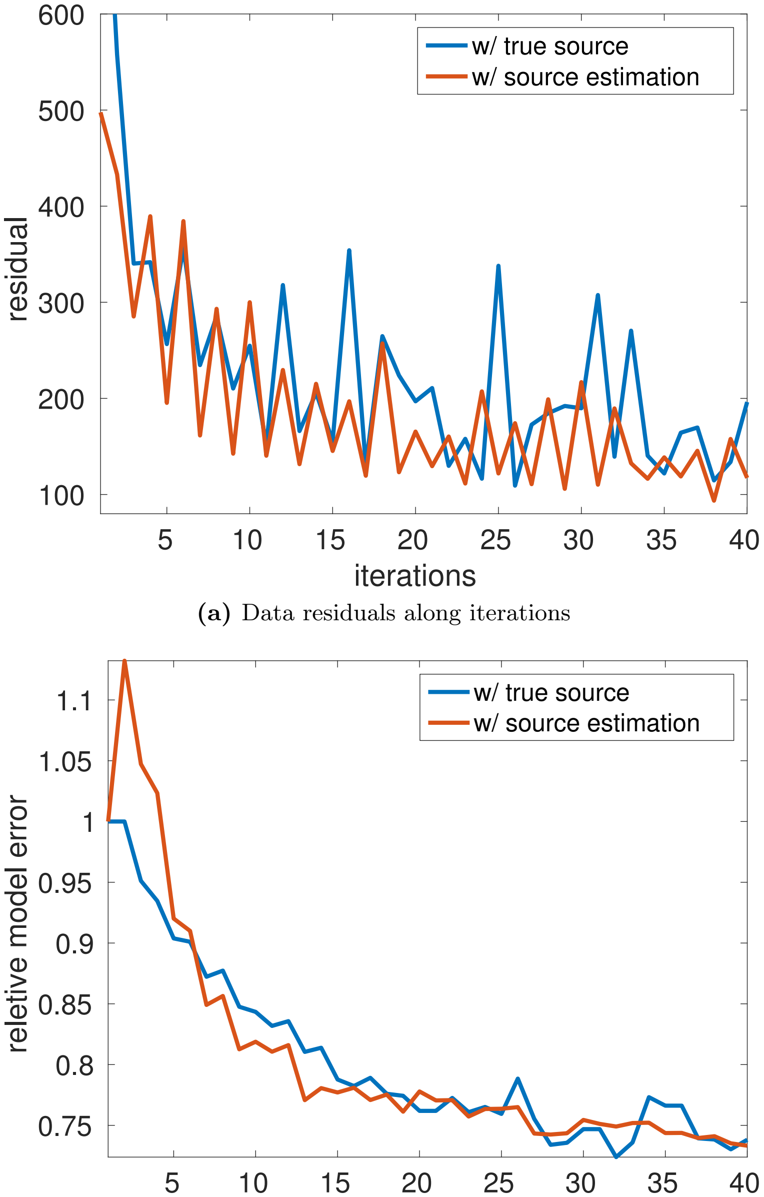


Figure 2.7: Comparison of the true (red), initial (blue), and recovered (green) source functions.



(a) Data residuals along iterations

(b) Model errors along iterations

Figure 2.8: Data residuals and model errors along iterations

2.4 Conclusions

In this chapter, we performed sparsity-promoting LS-RTM in the time domain using the linearized Bregman method. This algorithm is easy to implement and allows us to work with random subsets of data, which significantly reduces the cost of LS-RTM. Furthermore, this algorithm also allows us to incorporate source estimation into the inversion by solving a small additional least-squares subproblem during each Bregman iteration. Ultimately, we proposed a modified version of the LB algorithm including source estimation for sparsity-promoting LS-RTM in the time domain. Numerical experiments illustrated that the proposed algorithm can produce sharp subsurface images with true amplitudes without the requirement of having the correct source functions, while it only needs a computational cost that is in the range of one conventional RTM approach.

Chapter 3

Full-waveform inversion and wavefield-reconstruction inversion

3.1 Introduction

The successful application of LS-RTM introduced in the previous chapter heavily rely on the background velocity model that can preserve the kinematics of the wave propagation. During the past two-decades, full-waveform inversion has been reported by many researchers as a potentially promising approach to build up high-accuracy background velocity models for migration techniques including LS-RTM (Tarantola and Valette, 1982a; Lailly et al., 1983; Pratt, 1999; Virieux and Operto, 2009; Sirgue et al., 2010; Warner et al., 2013b; Vigh et al., 2014).

The objective of full-waveform inversion (FWI) is to compute the best possible Earth model that is consistent with observed data (Tarantola and Valette, 1982a; Lailly et al., 1983; Pratt, 1999; Virieux and Operto, 2009). Mathematically, it can be formulated as an optimization problem that is constrained by wave equations. The conventional FWI approach, also known as the adjoint-state method, eliminates the wave-equation constraints by directly solving the wave equations. However, due to the lack of long-offset data with low-frequency components, the adjoint-state method suffers from the local minima related to the so-called cycle skipping problem (Bunks et al., 1995; Sirgue and Pratt, 2004). To help mitigate the problem of local minima, van Leeuwen and Herrmann (2013b) and van Leeuwen and Herrmann (2015) proposed a penalty formulation, known as wavefield-reconstruction inversion (WRI), where they relaxed the wave-equation constraints by introducing the wavefields as auxiliary variables. Instead of exactly solving the wave equations as the adjoint-state method does, WRI replaces the wave-equation constraint by an ℓ_2-norm penalty term in the objective function. van Leeuwen and Herrmann (2013b) and van Leeuwen and Herrmann (2015) illustrated that through expanding the search space, WRI is less "nonlinear" and may contain less local minima compared to the conventional adjoint-state method.

In this chapter, we first introduce the theoretical formulations of both FWI and WRI in the frequency domain. Then we describe the variable projection method, which is the underlying tool that solves the WRI problem. Finally, we present a simple comparison between the objective functions of FWI and WRI to illustrate the potential benefit of WRI.

3.2 Full-waveform inversion

During FWI, we solve the following least-squares problem for the discretized n_{grid}-dimensional unknown medium parameters, squared slowness $\mathbf{m} \in \mathbb{R}^{n_{\text{grid}}}$, which appear as coefficients in the acoustic time-harmonic wave equation, i.e.,

$$\begin{aligned} \underset{\mathbf{u},\,\mathbf{m}}{\text{minimize}}\, \frac{1}{2} \sum_{i=1}^{n_{\text{src}}} \sum_{j=1}^{n_{\text{freq}}} \|\mathbf{P}_i \mathbf{u}_{i,j} - \mathbf{d}_{i,j}\|_2^2 \\ \text{subject to } \mathbf{A}_j(\mathbf{m})\mathbf{u}_{i,j} = \mathbf{q}_{i,j}. \end{aligned} \tag{3.1}$$

Here, the i, j's represent the source and frequency indices, and the vector $\mathbf{d} \in \mathbb{C}^{n_{\text{freq}} \times n_{\text{src}} \times n_{\text{rcv}}}$ represents monochromatic Fourier transformed data collected at n_{rcv} receiver locations from n_{src} seismic sources and sampled at n_{freq} frequencies. The matrix $\mathbf{A}_j(\mathbf{m}) = \omega_j^2 \text{diag}(\mathbf{m}) + \Delta$ is the discretized Helmholtz matrix at angular frequency ω_j. The symbol Δ represents the discretized Laplacian operator and the matrix $\mathbf{P}_i$ corresponds to a restriction operator for the receiver locations. Finally, the vectors $\mathbf{q}_{i,j}$ and $\mathbf{u}_{i,j}$ denote the monochromatic source term at the i^{th} source location and j^{th} frequency and the corresponding wavefield, respectively. For simplicity, we omit the dependence of $\mathbf{A}_j(\mathbf{m})$ on the discretized squared slowness vector $\mathbf{m}$ from now onwards.

Since the $\mathbf{A}_j$'s are square matrices, we can eliminate the variable $\mathbf{u}$ by solving the discretized partial-differential equation (PDE) explicitly, leading to the so-called adjoint-state method, which has the following reduced form (Virieux and Operto, 2009):

$$\underset{\mathbf{m}}{\text{minimize}}\, f_{\text{red}}(\mathbf{m}) = \frac{1}{2}\sum_{i=1}^{n_{\text{src}}}\sum_{j=1}^{n_{\text{freq}}} \|\mathbf{P}_i\mathbf{A}_j^{-1}\mathbf{q}_{i,j} - \mathbf{d}_{i,j}\|_2^2, \tag{3.2}$$

with the corresponding gradient $\mathbf{g}_{\text{red}}(\mathbf{m})$ given by

$$\mathbf{g}_{\text{red}}(\mathbf{m}) = \sum_{i=1}^{n_{\text{src}}}\sum_{j=1}^{n_{\text{freq}}} \omega_i^2 \text{diag}\Big(\text{conj}\big(\mathbf{u}_{i,j}\big)\Big)\mathbf{v}_{i,j}, \tag{3.3}$$

where the forward wavefield $\mathbf{u}_{i,j}$ and adjoint wavefield $\mathbf{v}_{i,j}$ have the following expressions:

$$\begin{aligned} \mathbf{u}_{i,j} &= \mathbf{A}_j^{-1}\mathbf{q}_{i,j}, \\ \mathbf{v}_{i,j} &= -\mathbf{A}_j^{-\top}\mathbf{P}_i^{\top}(\mathbf{P}_i\mathbf{u}_{i,j} - \mathbf{d}_{i,j}). \end{aligned} \tag{3.4}$$

Here, the symbol $^\top$ denotes the (complex-conjugate) transpose, and the term $\text{diag}\Big(\text{conj}\big(\mathbf{u}_{i,j}\big)\Big)$ represents a diagonal matrix with the elements of the complex conjugate of the vector $\mathbf{u}_{i,j}$ on the diagonal. In the equation 3.2, the dependence of the objective function $f_{\text{red}}(\mathbf{m})$ on $\mathbf{m}$ runs through the nonlinear operator $\mathbf{A}_j^{-1}\mathbf{q}_{i,j}$, instead of via the linear operator $\mathbf{A}_j\mathbf{u}_{i,j}$. As a result, the price we pay is that the objective function $f_{\text{red}}(\mathbf{m})$ becomes highly nonlinear in $\mathbf{m}$. The gains, of course, are that we no longer have to optimize over the wavefields and that we can compute the forward and adjoint wavefields independently in parallel. However, these gains are perhaps a bit short-lived, because the reduced objective function $f_{\text{red}}(\mathbf{m})$ may contain local minima (Warner et al., 2013a), which introduces more difficulties in the search for the best medium parameters using the local derivative information only.

3.3 Wavefield-reconstruction inversion

To free FWI from these parasitic minima, van Leeuwen and Herrmann (2013b) and van Leeuwen and Herrmann (2015) proposed WRI—a penalty formulation of FWI that reads

$$\underset{\mathbf{m},\,\mathbf{u}}{\text{minimize}}\, f_{\text{pen}}(\mathbf{m},\mathbf{u}) = \frac{1}{2}\sum_{i=1}^{n_{\text{src}}}\sum_{j=1}^{n_{\text{freq}}} \left(\|\mathbf{P}_i\mathbf{u}_{i,j} - \mathbf{d}_{i,j}\|_2^2 + \lambda^2\|\mathbf{A}_j\mathbf{u}_{i,j} - \mathbf{q}_{i,j}\|_2^2\right). \tag{3.5}$$

In this optimization problem, we keep the wavefields $\mathbf{u}$ as unknown variables instead of eliminating them as in FWI, i.e., we replace the PDE constraint by an ℓ_2-norm penalty term. The scalar parameter λ controls the balance between the data and the PDE misfits. As λ increases, the wavefield is more tightly constrained by the wave equation, and the objective function $f_{\text{pen}}(\mathbf{m}, \mathbf{u})$ in equation 3.5 converges to the objective function of the reduced problem in equation 3.2 (van Leeuwen and Herrmann, 2015). Different from the nonlinear objective function of the reduced problem in equation 3.2, the WRI objective $f_{\text{pen}}(\mathbf{m}, \mathbf{u})$ is a bi-quadratic function with respect to $\mathbf{m}$ and $\mathbf{u}$, i.e., for fixed $\mathbf{m}$, the objective $f_{\text{pen}}(\mathbf{m}, \mathbf{u})$ is quadratic with respect to $\mathbf{u}$ and vice versa. In addition, van Leeuwen and Herrmann (2013b) and van Leeuwen and Herrmann (2015) indicated that WRI explores a larger search space by not satisfying the PDE-constraints exactly, which results in that WRI is less “nonlinear” and may contain less local minima compared to the conventional adjoint-state method.

3.4 Variable projection method

In the optimization problem of equation 3.5, both the wavefields $\mathbf{u}$ and medium parameters $\mathbf{m}$ are unknown. Searching for both $\mathbf{u}$ and $\mathbf{m}$ by methods like gradient-descent and limited-memory Broyden–Fletcher–Goldfarb–Shanno (l-BFGS) method (Nocedal and Wright, 2006) requires storing the two unknown variables. However, the memory cost of storing $\mathbf{u}$ can be extremely large, because $\mathbf{u} \in \mathbb{C}^{n_{\text{grid}} \times n_{\text{freq}} \times n_{\text{src}}}$. To mitigate the storage cost, van Leeuwen and Herrmann (2013b) and van Leeuwen and Herrmann (2015) applied the variable projection method (Golub and Pereyra, 2003) to solve equation 3.5 using the fact that $f_{\text{pen}}(\mathbf{m}, \mathbf{u})$ is bi-quadratic with respect to $\mathbf{m}$ and $\mathbf{u}$. The variable projection method is designed to solve a general category of problems named separable nonlinear least-squares (SNLLS), which permits the following standard expression:

$$\underset{\theta, \mathbf{x}}{\text{minimize}}\, f(\theta, \mathbf{x}) = \frac{1}{2}\|\Phi(\theta)\mathbf{x} - \mathbf{y}\|^2, \tag{3.6}$$

where the matrix $\Phi(\theta)$ varies (nonlinearly) with respect to the variable vector θ. The vectors $\mathbf{x}$ and $\mathbf{y}$ represent a linear variable and the observations, respectively. More specifically, we can observe that the WRI problem in equation 3.5 is a special case of SNLLS problems by setting

$$\theta = \mathbf{m},\ \mathbf{x}_{i,j} = \mathbf{u}_{i,j},\ \mathbf{y}_{i,j} = \begin{bmatrix} \lambda \mathbf{q}_{i,j} \\ \mathbf{d}_{i,j} \end{bmatrix},\ \text{and}\ \Phi_{i,j}(\theta) = \begin{bmatrix} \lambda \mathbf{A}_j \\ \mathbf{P}_i \end{bmatrix}. \tag{3.7}$$

The vectors $\mathbf{x}$ and $\mathbf{y}$ are block vectors containing all $\{\mathbf{x}_{i,j}\}_{1\leq i\leq n_{\text{src}},1\leq j\leq n_{\text{freq}}}$ and $\{\mathbf{y}_{i,j}\}_{1\leq i\leq n_{\text{src}},1\leq j\leq n_{\text{freq}}}$. Likewise, $\Phi(\theta)$ is a block-diagonal matrix.

When introducing the variable projection method, it is often useful to recognize that for fixed θ, the objective function $f(\theta,\mathbf{x})$ becomes quadratic with respect to $\mathbf{x}$. As a consequence, the variable $\mathbf{x}$ has a closed-form optimal least-squares solution $\overline{\mathbf{x}}(\theta)$ that minimizes the objective $f(\theta,\mathbf{x})$:

$$\overline{\mathbf{x}}(\theta) = \big(\Phi(\theta)^\top\Phi(\theta)\big)^{-1}\Phi(\theta)^\top\mathbf{y}, \tag{3.8}$$

For each pair of (i,j), we now have

$$\overline{\mathbf{x}}_{i,j}(\theta) = \big(\Phi_{i,j}(\theta)^\top\Phi_{i,j}(\theta)\big)^{-1}\Phi_{i,j}(\theta)^\top\mathbf{y}_{i,j}. \tag{3.9}$$

By substituting equation 3.8 into equation 3.6, we project out the variable $\mathbf{x}$ by the optimal solution $\overline{\mathbf{x}}(\theta)$ and arrive at a reduced optimization problem over θ alone, i.e., we have

$$\underset{\theta}{\text{minimize}}\,\overline{f}(\theta) = f\big(\theta,\overline{\mathbf{x}}(\theta)\big) = \|\Big(\mathbf{I}-\Phi(\theta)\big(\Phi(\theta)^\top\Phi(\theta)\big)^{-1}\Phi(\theta)^\top\Big)\mathbf{y}\|^2. \tag{3.10}$$

The gradient of the reduced objective function $\overline{f}(\theta)$ can be computed via the chain rule, yielding

$$\begin{aligned}\overline{\mathbf{g}}(\theta) = \nabla_\theta\overline{f}(\theta) &= \nabla_\theta f\big(\theta,\overline{\mathbf{x}}(\theta)\big)\\ &= \nabla_\theta f(\theta,\mathbf{x})|_{\mathbf{x}=\overline{\mathbf{x}}(\theta)} + \nabla_{\mathbf{x}} f(\theta,\mathbf{x})|_{\mathbf{x}=\overline{\mathbf{x}}(\theta)}\nabla_\theta\mathbf{x}.\end{aligned} \tag{3.11}$$

Note that from the definition of $\overline{\mathbf{x}}(\theta)$, we have

$$\nabla_{\mathbf{x}} f(\theta,\mathbf{x})|_{\mathbf{x}=\overline{\mathbf{x}}(\theta)} = 0. \tag{3.12}$$

With this equality, the gradient in equation 3.11 can be written as:

$$\overline{\mathbf{g}}(\theta) = \nabla_\theta f(\theta,\mathbf{x})|_{\mathbf{x}=\overline{\mathbf{x}}(\theta)}. \tag{3.13}$$

Equation 3.13 implies that although the optimal solution $\overline{\mathbf{x}}(\theta)$ is a function of θ, the construction of the gradient $\overline{\mathbf{g}}(\theta)$ does not need to include the cross-derivative term $\nabla_{\mathbf{x}} f(\theta,\mathbf{x})|_{\mathbf{x}=\overline{\mathbf{x}}(\theta)}\nabla_\theta\mathbf{x}$.

As shown by van Leeuwen and Herrmann (2013b) and van Leeuwen and Herrmann (2015), we can through substituting equation 3.7 into equations 3.8, 3.10, and 3.13 successfully project out the wavefields $\mathbf{u}$ by computing the optimal wavefield $\overline{\mathbf{u}}_{i,j}(\mathbf{m})$ for each source and frequency:

$$\overline{\mathbf{u}}_{i,j}(\mathbf{m}) = \left(\lambda^2 \mathbf{A}_j^\top \mathbf{A}_j + \mathbf{P}_i^\top \mathbf{P}_i\right)^{-1} \left(\lambda^2 \mathbf{A}_j^\top \mathbf{q}_{i,j} + \mathbf{P}_i^\top \mathbf{d}_{i,j}\right), \tag{3.14}$$

and arrive at minimizing the following reduced objective:

$$\begin{aligned} \underset{\mathbf{m}}{\text{minimize}} \overline{f}_{\text{pen}}(\mathbf{m}) & \\ &= f_{\text{pen}}\big(\mathbf{m}, \overline{\mathbf{u}}(\mathbf{m})\big) \\ &= \frac{1}{2} \sum_{i=1}^{n_{\text{src}}} \sum_{j=1}^{n_{\text{freq}}} \left(\|\mathbf{P}_i \overline{\mathbf{u}}_{i,j}(\mathbf{m}) - \mathbf{d}_{i,j}\|_2^2 + \lambda^2 \|\mathbf{A}_j \overline{\mathbf{u}}_{i,j}(\mathbf{m}) - \mathbf{q}_{i,j}\|_2^2\right), \end{aligned} \tag{3.15}$$

with a gradient given by

$$\overline{\mathbf{g}}_{\text{pen}}(\mathbf{m}) = \sum_{i=1}^{n_{\text{src}}} \sum_{j=1}^{n_{\text{freq}}} \lambda^2 \omega_i^2 \text{diag}\Big(\text{conj}\big(\overline{\mathbf{u}}_{i,j}(\mathbf{m})\big)\Big) \big(\mathbf{A}_j \overline{\mathbf{u}}_{i,j}(\mathbf{m}) - \mathbf{q}_{i,j}\big). \tag{3.16}$$

With the projection of the wavefields $\mathbf{u}$ in equation 3.14, the new reduced objective $\overline{f}_{\text{pen}}(\mathbf{m})$ varies only with respect to the variable $\mathbf{m}$. As a result, we do not need to store the $\mathbf{u}$'s reducing the storage costs from $\mathcal{O}(n_{\text{grid}} \times (n_{\text{freq}} \times n_{\text{src}} + 1))$ to $\mathcal{O}(n_{\text{grid}})$.

3.5 FWI v.s. WRI

To illustrate the potential advantage of WRI over FWI, we conduct a simple numerical experiment, adopted from van Leeuwen and Herrmann (2013b), to compare the objective functions of FWI and WRI. We work with a 2D velocity model that is parameterized by a single varying parameter v_0 as follows:

$$v(z) = v_0 + 0.75z \,\text{m/s}, \tag{3.17}$$

where the parameter z increases with the vertical depth. We simulate data with a single source for $v_0 = 2500\,\text{m/s}$ using a grid spacing of 50 m and a single frequency of 5 Hz. The data do not contain free-surface related multiples. We place the source at $(z,\, x)$ coordinates (50 m, 50 m) and record the data at 200 receivers located at the depth of 50 m with a sampling interval of 50 m. In Figure 3.1, we compare the objective functions corresponding to FWI and WRI as a function of v_0 for various values of λ, i.e., $\lambda = 10^2$, 10^3, 10^4, and 10^5. Clearly, the global minima of WRI for different values of λ coincide with the one of FWI. Meanwhile, the objective function of WRI for a

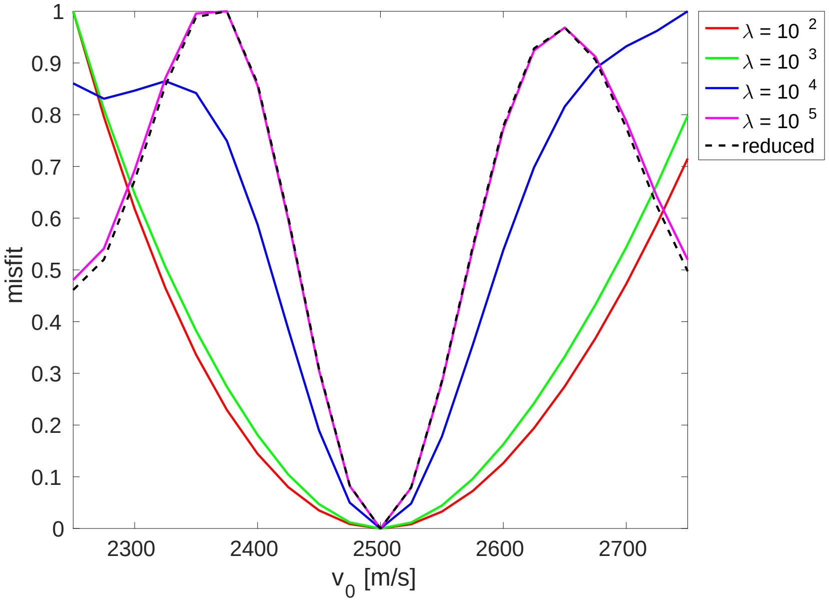


Figure 3.1: Comparison of objective functions of FWI and WRI as a function of v_0. For WRI, we select $\lambda = 10^2$, 10^3, 10^4, and 10^5. Clearly, as λ increases the objective function of WRI converges to that of the reduced formulation. When λ is small, there are no local minima in the objective function of WRI.

small λ exhibits no local minima, while local minima exist in the objective functions corresponding to FWI and WRI for a large λ. Furthermore, the behavior of the objective function for WRI converges to the one of FWI as λ increases. This example implies that through expanding the search space and carefully selecting the penalty parameter, WRI is potentially less prone to the local minima compared to the conventional FWI.

3.6 Conclusions

This chapter has discussed the theoretical formulations of full-waveform inversion and wavefield-reconstruction inversion in the frequency domain. We derived the optimization problem associated with FWI and the corresponding gradient. We also derived the optimization problem of WRI and introduced how to use the variable projection method to solve WRI. After introducing these two approaches, we illustrated the potential advantage of

WRI over FWI on mitigating local minima by a simple 2D example. We observed that the objective function of WRI converges to that of the reduced formulation as λ increases, while the objective functions of WRI exhibit no local minima for small selections of λ. In the next three chapters, we will extensively use the application of wavefield-reconstruction inversion based on the mathematical formulation described in this chapter.

Chapter 4

Source estimation for wavefield-reconstruction inversion

4.1 Introduction

Source information is essential to all wave-equation-based seismic inversions, including full-waveform inversion (FWI) and wavefield-reconstruction inversion (WRI). When the source function is inaccurate, errors will propagate into the predicted data and introduce additional data misfit. As a consequence, inversion results that minimize this data misfit may become erroneous. To mitigate the errors introduced by the incorrect and pre-estimated sources, an embedded procedure that updates sources along with medium parameters is necessary for the inversion.

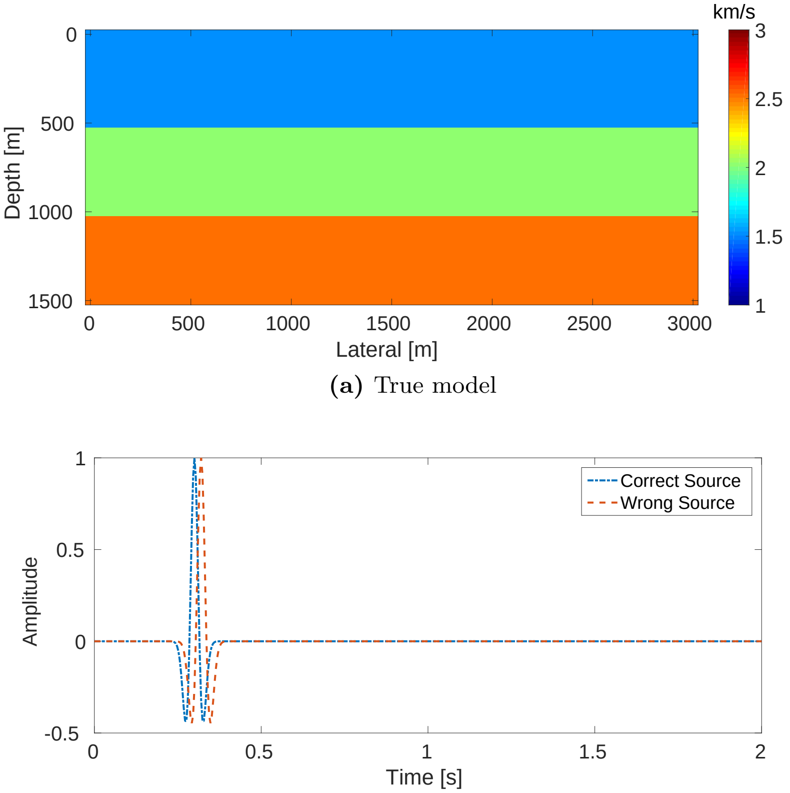
Many source estimation approaches have been developed for the conventional FWI, also known as the adjoint-state method (Plessix, 2006; Virieux and Operto, 2009). Liu et al. (2008) proposed to treat source functions as unknown variables and update them with medium parameters simultaneously. However, the amplitudes of source functions and medium parameters are different in scale, and so are the amplitudes of their corresponding gradients. As a result, the gradient-based optimization algorithms would tend to primarily update the one at the larger scale. Pratt (1999) proposed another approach to estimate source functions during each iteration of the adjoint-state method. In this approach, the author estimates the source functions

A version of this chapter has been submitted.

after each update of the medium parameters by solving a least-squares problem for each frequency. The solution of the least-squares problem is a single complex scalar that minimizes the difference between the observed and predicted data computed with the updated medium parameters. Because the updates of the source functions and medium parameters are in two separate steps, this method is not impacted by the different amplitude scales of the gradients. Indeed, Aravkin and van Leeuwen (2012) and van Leeuwen et al. (2014) pointed out that the problem of FWI with source estimation falls into a general category called separable nonlinear least-squares (SNLLS) problems (Golub and Pereyra, 2003), which can be solved by the so-called variable projection method. The method proposed by Pratt (1999) falls into that category. By virtue of these variable projections, the gradient with respect to the source functions becomes zero, because the estimated source is the optimal solution that minimizes the misfit between the observed and predicted data given the current medium parameters (Aravkin and van Leeuwen, 2012). For this reason, the gradient with respect to the medium parameters no longer contains the non-zero cross derivative with respect to the source functions, and the inverse problem becomes source-independent. In fact, because the problem of FWI with source estimation is a special case of the SNLLS problem, the minimization of this source-independent objective converges in fewer iterations than that of the original source-dependent objective theoretically (Golub and Pereyra, 2003). Furthermore, Li et al. (2013) presented empirical examples to illustrate that the variable projection method is more robust to noise compared to the method that performs alternating gradient descent steps on two variables separately (Zhou and Greenhalgh, 2003), as well as to the one that minimizes the two variables simultaneously (Liu et al., 2008).

Similar to the conventional FWI, WRI does, as illustrated in Figures 4.1 and 4.2, also require accurate information on the source functions. For instance, a small difference in the origin time of the source can lead to a significant deterioration of the inversion result (cf. Figures 4.1 and 4.2). However, up to now, there is still missing an on-the-fly source estimation procedure in the context of WRI .

Recognizing this sensitivity to source functions, in this chapter, we propose a modified WRI framework with source estimation integrated, making the method applicable to field data. In this new formulation, we continue to solve an optimization problem but now one that has the source functions, medium parameters, and wavefields all as unknowns. For fixed medium parameters, we observe that the objective function of this optimization problem is quadratic with respect to both wavefields and source functions. As



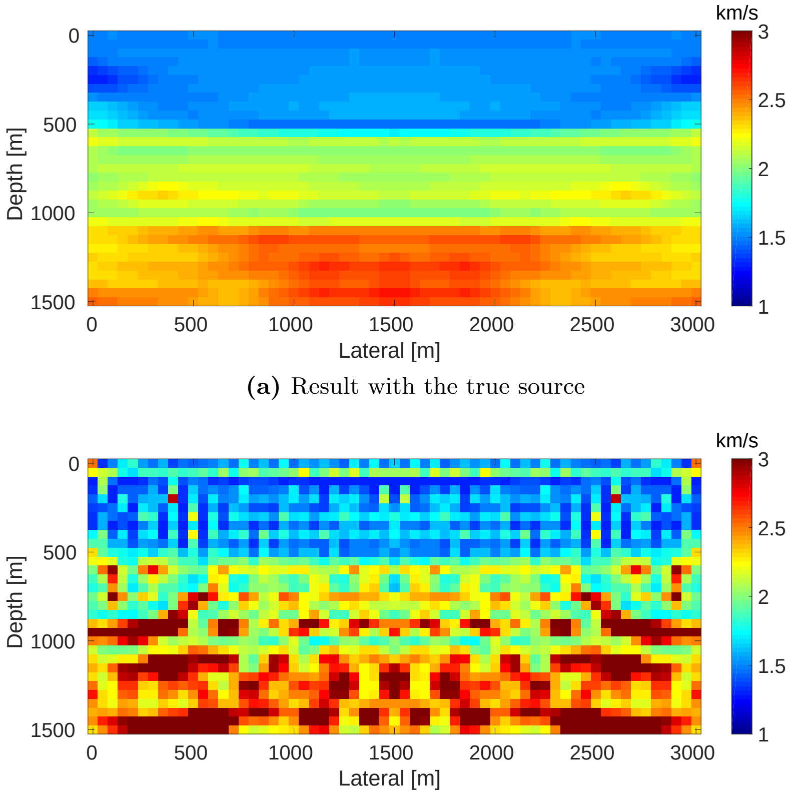
(a) True model

(b) Correct source v.s. wrong source

Figure 4.1: (a) The true model. (b) Comparison of the correct (blue) and wrong (red) source functions.

a result, if we collect the wavefield and source function for each source experiment into one single unknown vector variable, the optimization problem becomes an SNLLS problem with respect to the medium parameters and this new unknown variable. As before, we propose the variable projection method to solve the optimization problem by projecting out this new unknown variable during each iteration. After these least-squares problems, we compute single model updates either with the gradient descent or limited-memory Broyden–Fletcher–Goldfarb–Shanno (l-BFGS) method (Nocedal and Wright, 2006).

This chapter is organized as follows. First, we present the optimization problem of the source estimation for WRI. This optimization problem can be formulated as an SNLLS problem by collecting the wavefield and source



(a) Result with the true source

(b) Result with the wrong source

Figure 4.2: (a) Inversion result with the true source function. (b) Inversion result with the wrong source function.

function for each source experiment into one single unknown vector variable. Then, we apply the variable projection method to solve this optimization problem through simultaneously projecting out the wavefield and source function during each iteration. We conclude by presenting a computationally efficient implementation to compute the projection, which we evaluate on several synthetic examples to highlight the benefits of our approach.

4.2 WRI with source estimation

While spatial locations of the sources are often known quite accurately, the source function is generally unknown. If we ignore spatial directivity of the sources, we can handle this situation in the context of WRI by incorporating the source function through setting $\mathbf{q}_{i,j} = \alpha_{i,j}\mathbf{e}_{i,j}$ in the frequency-domain, where $\alpha_{i,j}$ is a complex number representing the frequency-dependent source weight for the i^{th} source experiment at the j^{th} frequency. In this expression, the symbol $\mathbf{e}_{i,j}$ represents a unit vector that equals one at the source location and zero elsewhere. When we substitute $\mathbf{q}_{i,j} = \alpha_{i,j}\mathbf{e}_{i,j}$ into the objective function in equation 3.5, we arrive at an optimization problem with three unknown variables, i.e., we have

$$\begin{aligned}\underset{\mathbf{m},\,\mathbf{u},\,\alpha}{\text{minimize}} & f_{\text{pen}}(\mathbf{m},\,\mathbf{u},\,\alpha)\\ &= \frac{1}{2}\sum_{i=1}^{n_{\text{src}}}\sum_{j=1}^{n_{\text{freq}}}\left(\|\mathbf{P}_i\mathbf{u}_{i,j} - \mathbf{d}_{i,j}\|_2^2 + \lambda^2\|\mathbf{A}_j\mathbf{u}_{i,j} - \alpha_{i,j}\mathbf{e}_{i,j}\|_2^2\right),\end{aligned} \tag{4.1}$$

where the vector $\alpha = \{\alpha_{i,j}\}_{1\leq i\leq n_{\text{src}},1\leq j\leq n_{\text{freq}}}$ contains all the source weights.

Because this new objective $f_{\text{pen}}(\mathbf{m},\,\mathbf{u},\,\alpha)$ contains more than two unknowns, it no longer corresponds to a standard SNLLS problem, so we can not directly apply the variable projection method to solve it. One simple approach that comes to mind to convert the problem in equation 4.1 to a standard SNLLS problem would be to reduce the number of unknown variables by collecting two of the three variables into a new variable. Since for fixed $\mathbf{m}$ and α, $f_{\text{pen}}(\mathbf{m},\,\mathbf{u},\,\alpha)$ is still quadratic with respect to $\mathbf{u}$, the most straightforward approach would be to lump the medium parameters $\mathbf{m}$ and the source weights α together. Following the notation of the standard SNLLS problems, for each pair of (i,j), we now have

$$\theta = \begin{bmatrix}\mathbf{m}\\ \alpha\end{bmatrix},\ \mathbf{x}_{i,j} = \begin{bmatrix}\mathbf{u}_{i,j}\\ 1\end{bmatrix},\ \mathbf{y}_{i,j} = \begin{bmatrix}\mathbf{0}\\ \mathbf{d}_{i,j}\end{bmatrix},\ \text{and}\ \Phi_{i,j}(\theta) = \begin{bmatrix}\lambda\mathbf{A}_j & -\lambda\alpha_{i,j}\mathbf{e}_{i,j}\\ \mathbf{P}_i & \mathbf{0}\end{bmatrix}. \tag{4.2}$$

Through the substitution of equation 4.2 into equations 3.9 and 3.10, we can again apply the variable projection method to project out the wavefields $\mathbf{u}$ via

$$\overline{\mathbf{u}}_{i,j}(\mathbf{m},\alpha_{i,j}) = \left(\lambda^2\mathbf{A}_j^\top\mathbf{A}_j + \mathbf{P}_i^\top\mathbf{P}_i\right)^{-1}\left(\lambda^2\alpha_{i,j}\mathbf{A}_j^\top\mathbf{e}_{i,j} + \mathbf{P}_i^\top\mathbf{d}_{i,j}\right). \tag{4.3}$$

In this way, we arrive at the following reduced problem:

$$\underset{\mathbf{m},\,\alpha}{\text{minimize}}\ \tilde{f}(\mathbf{m},\alpha) = f_{\text{pen}}\big(\mathbf{m},\,\overline{\mathbf{u}}(\mathbf{m},\alpha),\,\alpha\big). \tag{4.4}$$

Similarly, the computation of the gradient $\tilde{\mathbf{g}}(\mathbf{m},\alpha)$ of the new objective $\tilde{f}(\mathbf{m},\alpha)$ is straightforward if we substitute equation 4.2 into equation 3.13, which yields

$$\tilde{\mathbf{g}}(\mathbf{m},\alpha) = \begin{bmatrix} \nabla_{\mathbf{m}}\tilde{f}(\mathbf{m},\alpha) \\ \nabla_{\alpha}\tilde{f}(\mathbf{m},\alpha) \end{bmatrix} = \begin{bmatrix} \nabla_{\mathbf{m}} f_{\text{pen}}(\mathbf{m},\mathbf{u},\alpha)|_{\mathbf{u}=\bar{\mathbf{u}}(\mathbf{m},\alpha)} \\ \nabla_{\alpha} f_{\text{pen}}(\mathbf{m},\mathbf{u},\alpha)|_{\mathbf{u}=\bar{\mathbf{u}}(\mathbf{m},\alpha)} \end{bmatrix}, \tag{4.5}$$

where

$$\begin{aligned} &\nabla_{\mathbf{m}} f_{\text{pen}}(\mathbf{m},\mathbf{u},\alpha)|_{\mathbf{u}=\bar{\mathbf{u}}(\mathbf{m},\alpha)} \\ =&\sum_{i=1}^{n_{\text{src}}}\sum_{j=1}^{n_{\text{freq}}} \lambda^2\omega_i^2 \text{diag}\Big(\text{conj}\big(\bar{\mathbf{u}}_{i,j}(\mathbf{m},\alpha_{i,j})\big)\Big)\big(\mathbf{A}_j\bar{\mathbf{u}}_{i,j}(\mathbf{m},\alpha_{i,j}) - \alpha_{i,j}\mathbf{e}_{i,j}\big) \end{aligned} \tag{4.6}$$

is the gradient with respect to $\mathbf{m}$ and

$$\begin{aligned} &\nabla_{\alpha} f_{\text{pen}}(\mathbf{m},\mathbf{u},\alpha)|_{\mathbf{u}=\bar{\mathbf{u}}(\mathbf{m},\alpha)} \\ =&\{\lambda^2\mathbf{e}_{i,j}^{\top}\big(\alpha_{i,j}\mathbf{e}_{i,j} - \mathbf{A}_j\bar{\mathbf{u}}_{i,j}\left(\mathbf{m},\alpha_{i,j}\right)\big)\}_{1\leq i\leq n_{\text{src}},1\leq j\leq n_{\text{freq}}} \end{aligned} \tag{4.7}$$

is the gradient with respect to the source weights.

While the gradients in equations 4.6 and 4.7 seemingly provide all information that we would need to drive the inversion, their amplitudes may significantly differ in scale, which is a common problem in optimization problems with multiple unknowns (Sambridge, 1990; Kennett et al., 1988; Rawlinson et al., 2006). An undesired consequence of this mismatch between the gradients in equations 4.6 and 4.7 is that the optimization may tend to primarily update the variable with the larger gradient contribution resulting in small updates for the other variable. This behavior may slow down the convergence, and the small updates may lead to errors that end up in the other variable resulting in a poor solution.

To mitigate the challenge arising from the different amplitude scales of $\mathbf{m}$ and α and their gradients, we propose another strategy to reduce the original problem in equation 4.1. Instead of collecting $\mathbf{m}$ and α together, we lump $\mathbf{u}$ and α into a single vector $\mathbf{x}$. To understand the potential advantage of this alternative strategy, let us rewrite the optimization problem in equation 4.1

as

$$\begin{aligned}\underset{\mathbf{m},\mathbf{u},\alpha}{\text{minimize}}\, f_{\text{pen}}(\mathbf{m},\mathbf{u},\alpha) &= \frac{1}{2}\sum_{i=1}^{n_{\text{src}}}\sum_{j=1}^{n_{\text{freq}}} \| \begin{bmatrix}\lambda\mathbf{A}_j & -\lambda\mathbf{e}_{i,j}\\ \mathbf{P}_i & 0\end{bmatrix}\begin{bmatrix}\mathbf{u}_{i,j}\\ \alpha_{i,j}\end{bmatrix} - \begin{bmatrix}0\\ \mathbf{d}_{i,j}\end{bmatrix} \|^2\\ &= \frac{1}{2}\sum_{i=1}^{n_{\text{src}}}\sum_{j=1}^{n_{\text{freq}}} \| \begin{bmatrix}\lambda\mathbf{A}_j & -\lambda\mathbf{e}_{i,j}\\ \mathbf{P}_i & 0\end{bmatrix}\mathbf{x}_{i,j} - \begin{bmatrix}0\\ \mathbf{d}_{i,j}\end{bmatrix} \|^2\\ &:= \hat{f}_{\text{pen}}(\mathbf{m},\mathbf{x}).\end{aligned} \tag{4.8}$$

For fixed $\mathbf{m}$, the objective function $\hat{f}_{\text{pen}}(\mathbf{m},\mathbf{x})$ continues to be quadratic in the new variable $\mathbf{x}$, where $\mathbf{u}$ and α are grouped together. As a result, we can rewrite the optimization problem in equation 4.8 into the standard SNLLS form (see equation 3.6) when we use the following notations:

$$\theta = \mathbf{m},\ \mathbf{x}_{i,j} = \begin{bmatrix}\mathbf{u}_{i,j}\\ \alpha_{i,j}\end{bmatrix},\ \mathbf{y}_{i,j} = \begin{bmatrix}\mathbf{0}\\ \mathbf{d}_{i,j}\end{bmatrix},\ \text{and}\ \Phi_{i,j}(\theta) = \begin{bmatrix}\lambda\mathbf{A}_j & -\lambda\mathbf{e}_{i,j}\\ \mathbf{P}_i & \mathbf{0}\end{bmatrix}. \tag{4.9}$$

Now by substituting the notations in equation 4.9 into equation 3.9, we can apply the variable projection method again to project out $\mathbf{x}$ from the objective function $\hat{f}_{\text{pen}}(\mathbf{m},\mathbf{x})$ for fixed $\mathbf{m}$ by computing the optimal solution

$$\overline{\mathbf{x}}_{i,j}(\mathbf{m}) = \begin{bmatrix}\overline{\mathbf{u}}_{i,j}(\mathbf{m})\\ \overline{\alpha}_{i,j}(\mathbf{m})\end{bmatrix} = \begin{bmatrix}\lambda^2\mathbf{A}_j^\top\mathbf{A}_j + \mathbf{P}_i^\top\mathbf{P}_i & -\lambda^2\mathbf{A}_j^\top\mathbf{e}_{i,j}\\ -\lambda^2\mathbf{e}_{i,j}^\top\mathbf{A}_j & \lambda^2\mathbf{e}_{i,j}^\top\mathbf{e}_{i,j}\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{P}_i^\top\mathbf{d}_{i,j}\\ \mathbf{0}\end{bmatrix}. \tag{4.10}$$

The linear system in equation 4.10 is solvable if and only if the rank of the matrix $\Phi_{i,j}(\theta)$ in equation 4.9 is $n_{\text{grid}} + 1$, which requires

$$\mathbf{P}_i\mathbf{A}_j^{-1}\mathbf{e}_{i,j} \neq 0. \tag{4.11}$$

In fact, this means that the wavefields corresponding to a point source do not vanish at all the receiver locations simultaneously, which holds in almost all seismic situations. Indeed, even in the extremely rare situation where a source is made nearly "silent" or non-radiating due to its surrounding geological structures, we can simply discard this source in the inversion to avoid inverting a rank deficient matrix. After the projection in equation 4.10, we obtain an objective function that only depends on $\mathbf{m}$:

$$\hat{f}(\mathbf{m}) = f_{\text{pen}}\big(\mathbf{m},\overline{\mathbf{u}}(\mathbf{m}),\overline{\alpha}(\mathbf{m})\big), \tag{4.12}$$

whose gradient $\hat{\mathbf{g}}(\mathbf{m})$ can be derived as

$$\begin{aligned}\hat{\mathbf{g}}(\mathbf{m}) &= \nabla_{\mathbf{m}}\hat{f}(\mathbf{m}) = \nabla_{\mathbf{m}}f_{\text{pen}}(\mathbf{m},\mathbf{u},\alpha)|_{\mathbf{u}=\overline{\mathbf{u}}(\mathbf{m}),\alpha=\overline{\alpha}(\mathbf{m})} \\ &= \sum_{i=1}^{n_{\text{src}}}\sum_{j=1}^{n_{\text{freq}}}\lambda^2\omega_i^2\text{diag}\Big(\text{conj}\big(\overline{\mathbf{u}}_{i,j}(\mathbf{m})\big)\Big)\big(\mathbf{A}_j\overline{\mathbf{u}}_{i,j}(\mathbf{m}) - \overline{\alpha}_{i,j}(\mathbf{m})\mathbf{e}_{i,j}\big).\end{aligned} \tag{4.13}$$

Compared to the objective function $\tilde{f}(\mathbf{m},\alpha)$, which we obtained by projecting out $\mathbf{u}$ alone, the objective function $\hat{f}(\mathbf{m})$ only depends on $\mathbf{m}$. As a consequence, we avoid having to optimize over two different variables that differ in scale. Moreover, because the objective $\hat{f}(\mathbf{m})$ is source-independent, the optimization with it does not require any initial guesses of the source weights, while the optimization with the objective $\tilde{f}(\mathbf{m},\alpha)$ does need a sufficiently accurate initial guess to ensure reaching the correct solution.

4.3 Optimization scheme

With the gradient $\hat{\mathbf{g}}(\mathbf{m})$ we derived in the previous section we can compute updates for the medium parameters via the steepest-descent method (Nocedal and Wright, 2006):

$$\mathbf{m}_{k+1} = \mathbf{m}_k - \beta_k\hat{\mathbf{g}}(\mathbf{m}_k). \tag{4.14}$$

Here, the value β_k is an appropriately chosen step length at the k^{th} iteration, which requires a line search to determine. Even with an optimized step length, as a first-order method, the steepest-descent method can be slow. To speed up the convergence, second-order methods including Newton method, Gauss-Newton method and quasi-Newton method may be more desirable (Pratt, 1999; Akcelik, 2002; Askan et al., 2007). During each iteration, Newton method and Gauss-Newton method use the full or Gauss-Newton Hessian to compute the update direction. However, the construction and inversion of the full or Gauss-Newton Hessian for both FWI (Pratt, 1999) and WRI (van Leeuwen and Herrmann, 2015) involve a large amount of additional PDE solves, which makes these two methods less attractive in this context. On the other hand, the quasi-Newton method, especially the l-BFGS method (Nocedal and Wright, 2006; Brossier et al., 2009; van Leeuwen and Herrmann, 2013a), utilizes the gradient at the current iteration k and a few previous gradients (typically, 3-20 iterations) to construct an approximation of the inverse of the Hessian. Hence, except for the increased memory, the l-BFGS method constructs the approximation of the inverse of the Hessian for free. For this reason, we use the l-BFGS method as our optimization scheme and denote the l-BFGS update as

$$\mathbf{m}_{k+1} = \mathbf{m}_k - \beta_k \mathbf{H}(\mathbf{m}_k)^{-1}\hat{\mathbf{g}}(\mathbf{m}_k), \tag{4.15}$$

where the matrix $\mathbf{H}(\mathbf{m}_k)$ is the l-BFGS Hessian and the step length β_k is determined by a weak Wolfe line search. We give a detailed description of the l-BFGS method for the proposed WRI framework with source estimation in Algorithm 3.

Algorithm 3 l-BFGS algorithm for WRI with source estimation

```
Initialization with an initial model $\mathbf{m}_0$
for $k = 1 \to n_{\text{iter}}$
    for $j = 1 \to n_{\text{freq}}$
        for $i = 1 \to n_{\text{src}}$
            Compute $\overline{\mathbf{u}}_{i,j}(\mathbf{m}_k)$ and $\overline{\alpha}_{i,j}(\mathbf{m}_k)$ by equation 4.10
        end
    end
    Compute $\hat{f}_k$ and $\hat{\mathbf{g}}_k$ by equations 4.12 and 4.13
    Apply l-BFGS Hessian with history size $n_{\text{his}}$
    $\mathbf{p}_k = \text{lbfgs}(-\hat{\mathbf{g}}_k, \{\mathbf{t}_l\}_{l=k-n_{\text{his}}}^{k}, \{\mathbf{s}_l\}_{l=k-n_{\text{his}}}^{k})$
    $\{\mathbf{m}_{k+1}, \hat{f}_{k+1}, \hat{\mathbf{g}}_{k+1}\} = \text{line search}(\hat{f}_k, \hat{\mathbf{g}}_k, \mathbf{p}_k)$
    $\mathbf{t}_{k+1} = \mathbf{m}_{k+1} - \mathbf{m}_k$
    $\mathbf{s}_{k+1} = \hat{\mathbf{g}}_{k+1} - \hat{\mathbf{g}}_k$
end
```

4.4 Fast solver for WRI with source estimation

Most of the computational burden for the objective function and its gradient in equations 4.12 and 4.13 lies within inverting the matrix

$$\mathbf{C}_{i,j}(\mathbf{m}) = \Phi_{i,j}(\mathbf{m})^\top \Phi_{i,j}(\mathbf{m}) = \begin{bmatrix} \lambda^2 \mathbf{A}_j^\top \mathbf{A}_j + \mathbf{P}_i^\top \mathbf{P}_i & -\lambda^2 \mathbf{A}_j^\top \mathbf{e}_{i,j} \\ -\lambda^2 \mathbf{e}_{i,j}^\top \mathbf{A}_j & \lambda^2 \mathbf{e}_{i,j}^\top \mathbf{e}_{i,j} \end{bmatrix}. \tag{4.16}$$

For 2D problems, we can invert the matrix $\mathbf{C}_{i,j}(\mathbf{m})$ by direct solvers such as the Cholesky method, while for 3D problems, we may need an iterative solver equipped with a proper preconditioner. In either case, following computational challenges arise from the fact that the matrix $\mathbf{C}_{i,j}(\mathbf{m})$ differs from source to source and frequency to frequency. First, to apply the Cholesky method, we need to calculate the Cholesky factorization for each

i and j, i.e., for each source and frequency. As a result, the computational cost arrives at $\mathcal{O}\big(n_{\text{src}} \times n_{\text{freq}} \times (n_{\text{grid}}^3 + n_{\text{grid}}^2)\big)$, which prohibits the application to problems with a large number of sources. Secondly, for iterative solvers, because the matrix $\mathbf{C}_{i,j}(\mathbf{m})$ varies with respect to i and j, the corresponding preconditioner may as well as depend on i and j. Therefore, we may have to design $n_{\text{src}} \times n_{\text{freq}}$ different preconditioners, which can be computationally difficult and intractable. Moreover, the additional terms $-\lambda^2\mathbf{A}_j^\top\mathbf{e}_{i,j}$, $-\lambda^2\mathbf{e}_{i,j}^\top\mathbf{A}_j$, and $\lambda^2\mathbf{e}_{i,j}^\top\mathbf{e}_{i,j}$ may lead to the condition number of the matrix $\mathbf{C}_{i,j}$ worse than that of the matrix $\lambda^2\mathbf{A}_j^\top\mathbf{A}_j + \mathbf{P}_i^\top\mathbf{P}_i$. As a result, without a good preconditioner, the projection procedure in the framework of WRI with source estimation may be much slower than that without source estimation.

To lighten the computational costs of inverting the matrix $\mathbf{C}_{i,j}(\mathbf{m})$, we describe a new inversion scheme to implement the algorithm when the projection operator $\mathbf{P}$ remains the same for all sources. This situation is common in ocean bottom node acquisition and the land acquisition where all sources “see” the same receivers. When we replace $\mathbf{P}_i$ by a single $\mathbf{P}$ in equation 4.16, the matrix $\mathbf{C}_{i,j}(\mathbf{m})$ can be simplified as

$$\mathbf{C}_{i,j}(\mathbf{m}) = \begin{bmatrix} \lambda^2\mathbf{A}_j^\top\mathbf{A}_j + \mathbf{P}^\top\mathbf{P} & -\lambda^2\mathbf{A}_j^\top\mathbf{e}_{i,j} \\ -\lambda^2\mathbf{e}_{i,j}^\top\mathbf{A}_j & \lambda^2\mathbf{e}_{i,j}^\top\mathbf{e}_{i,j} \end{bmatrix}. \tag{4.17}$$

Unfortunately, this matrix $\mathbf{C}_{i,j}(\mathbf{m})$ still depends on the source and frequency indices, and a straightforward inversion still faces the aforementioned computational challenges. However, this form, equation 4.17, allows us to use an alternative block matrix formula to invert $\mathbf{C}_{i,j}(\mathbf{m})$. To arrive at this result, let us first write $\mathbf{C}_{i,j}(\mathbf{m})$ in a simpler form

$$\mathbf{C}_{i,j}(\mathbf{m}) = \begin{bmatrix} \mathbf{M}_1 & \mathbf{M}_2 \\ \mathbf{M}_3 & \mathbf{M}_4 \end{bmatrix}, \tag{4.18}$$

where the matrix $\mathbf{M}_1 = \lambda^2\mathbf{A}_j^\top\mathbf{A}_j + \mathbf{P}^\top\mathbf{P}$, the vector $\mathbf{M}_2 = \mathbf{M}_3^\top = -\lambda^2\mathbf{A}_j^\top\mathbf{e}_{i,j}$, and the scalar $\mathbf{M}_4 = \lambda^2\mathbf{e}_{i,j}^\top\mathbf{e}_{i,j}$. Now if we apply the block matrix inversion formula (Bernstein, 2005) to compute $\mathbf{C}_{i,j}^{-1}(\mathbf{m})$, we arrive at the following closed form analytical expression

$$\begin{aligned}
\mathbf{C}_{i,j}^{-1}(\mathbf{m}) &= \begin{bmatrix} \tilde{\mathbf{M}}_1 & \tilde{\mathbf{M}}_2 \\ \tilde{\mathbf{M}}_3 & \tilde{\mathbf{M}}_4 \end{bmatrix}, \text{ with} \\
&\tilde{\mathbf{M}}_1 = (\mathbf{I} + \mathbf{M}_1^{-1}\mathbf{M}_2(\mathbf{M}_4 - \mathbf{M}_3\mathbf{M}_1^{-1}\mathbf{M}_2)^{-1}\mathbf{M}_3)\mathbf{M}_1^{-1}, \\
&\tilde{\mathbf{M}}_2 = -\mathbf{M}_1^{-1}\mathbf{M}_2(\mathbf{M}_4 - \mathbf{M}_3\mathbf{M}_1^{-1}\mathbf{M}_2)^{-1}, \\
&\tilde{\mathbf{M}}_3 = -(\mathbf{M}_4 - \mathbf{M}_3\mathbf{M}_1^{-1}\mathbf{M}_2)^{-1}\mathbf{M}_3\mathbf{M}_1^{-1}, \text{ and} \\
&\tilde{\mathbf{M}}_4 = (\mathbf{M}_4 - \mathbf{M}_3\mathbf{M}_1^{-1}\mathbf{M}_2)^{-1}.
\end{aligned} \tag{4.19}$$

While complicated the right-hand side of this equation only involves the inversion of the source-independent matrix $\mathbf{M}_1$, all other terms are all scalar inversions that can be evaluated at negligible costs. Substituting equation 4.19 into equation 4.10, we obtain an analytical solution for the optimal variable

$$\begin{aligned}
\overline{\mathbf{x}}_{i,j}(\mathbf{m}) &= \mathbf{C}_{i,j}^{-1}(\mathbf{m}) \begin{bmatrix} \mathbf{P}^\top \mathbf{d}_{i,j} \\ 0 \end{bmatrix} \\
&= \begin{bmatrix} (\mathbf{I} + \mathbf{M}_1^{-1}\mathbf{M}_2(\mathbf{M}_4 - \mathbf{M}_3\mathbf{M}_1^{-1}\mathbf{M}_2)^{-1}\mathbf{M}_3)\mathbf{M}_1^{-1}\mathbf{P}^\top \mathbf{d}_{i,j} \\ -(\mathbf{M}_4 - \mathbf{M}_3\mathbf{M}_1^{-1}\mathbf{M}_2)^{-1}\mathbf{M}_3\mathbf{M}_1^{-1}\mathbf{P}^\top \mathbf{d}_{i,j} \end{bmatrix}.
\end{aligned} \tag{4.20}$$

In equation 4.20, we have to invert two terms, i.e., $\mathbf{M}_1$ and $\mathbf{M}_4 - \mathbf{M}_3\mathbf{M}_1^{-1}\mathbf{M}_2$. Because $\mathbf{M}_4$ is a scalar, the term $(\mathbf{M}_4 - \mathbf{M}_3\mathbf{M}_1^{-1}\mathbf{M}_2)^{-1}$ is a scalar inversion, whose computational cost is negligible. Therefore, to construct $\overline{\mathbf{x}}_{i,j}(\mathbf{m})$, we only need to compute the following two vectors:

$$\begin{aligned}
\mathbf{w}_1 &= \mathbf{M}_1^{-1}\mathbf{M}_2 \\
&= -\lambda^2(\lambda^2\mathbf{A}_j^\top\mathbf{A}_j + \mathbf{P}^\top\mathbf{P})^{-1}\mathbf{A}_j^\top\mathbf{e}_{i,j},
\end{aligned} \tag{4.21}$$

and

$$\begin{aligned}
\mathbf{w}_2 &= \mathbf{M}_1^{-1}\mathbf{P}^\top\mathbf{d}_{i,j} \\
&= (\lambda^2\mathbf{A}_j^\top\mathbf{A}_j + \mathbf{P}^\top\mathbf{P})^{-1}\mathbf{P}^\top\mathbf{d}_{i,j}.
\end{aligned} \tag{4.22}$$

From these two expressions, we can observe that the computation of the vectors $\mathbf{w}_1$ and $\mathbf{w}_2$ for each source is independent from that of other sources, therefore, we can sequentially compute them from source to source, yielding a negligible requirement of storing only 2 additional wavefields during the inversion. For each frequency, because the matrix $\mathbf{M}_1$ is source-independent, we only need one Cholesky factorization, whose computational cost is $\mathcal{O}(n_{\text{grid}}^3)$. With the pre-computed Cholesky factors, for each source,

	nr. of Cholesky factorizations per frequency	nr. of inversions w/ Cholesky factors per frequency
w/o SE	1	2 per source
w/ SE – old form	n_{src}	2 per source
w/ SE – new form	1	4 per source

Table 4.1: Comparison of the computational costs for different algorithms.

solving $\mathbf{w}_1$ and $\mathbf{w}_2$ by equations 4.21 and 4.22 requires inverting the matrix $\mathbf{M}_1$ twice with a computational cost of $\mathcal{O}(n_{\text{grid}}^2)$. Consequently, we reduce the total cost from $\mathcal{O}\big(n_{\text{src}} \times n_{\text{freq}} \times (n_{\text{grid}}^3 + n_{\text{grid}}^2)\big)$ to $\mathcal{O}\big(n_{\text{freq}} \times (n_{\text{grid}}^3 + 2 \times n_{\text{src}} \times n_{\text{grid}}^2)\big)$. As a reference, the computational complexity of WRI without source estimation is of a close order $\mathcal{O}\big(n_{\text{freq}} \times (n_{\text{grid}}^3 + n_{\text{src}} \times n_{\text{grid}}^2)\big)$ (van Leeuwen and Herrmann, 2015). A comparison of the computational complexity of WRI with and without source estimation (SE) using the Cholesky method is shown in Table 4.1. In this table, we refer to the scheme that directly inverts the matrix $\mathbf{C}_{i,j}(\mathbf{m})$ as the old form, and refer to the proposed scheme in equation 4.20 as the new form. Compared to the old form, instead of requiring n_{src} Cholesky factorizations, the proposed new form only requires 1 Cholesky factorization for each frequency, which significantly reduces the computational cost. Furthermore, besides reducing the cost for direct solvers, the proposed inversion scheme can also benefit iterative solvers. For each frequency, since all sources only need to invert the same matrix $\mathbf{M}_1$, the proposed new form avoids inverting the potentially ill-conditioned matrix $\mathbf{C}_{i,j}$ directly and only requires one preconditioner instead of n_{src}, which significantly simplifies the projection procedure. In the following section of numerical experiments, we will only show the computational gain for direct solvers.

4.5 Numerical experiments

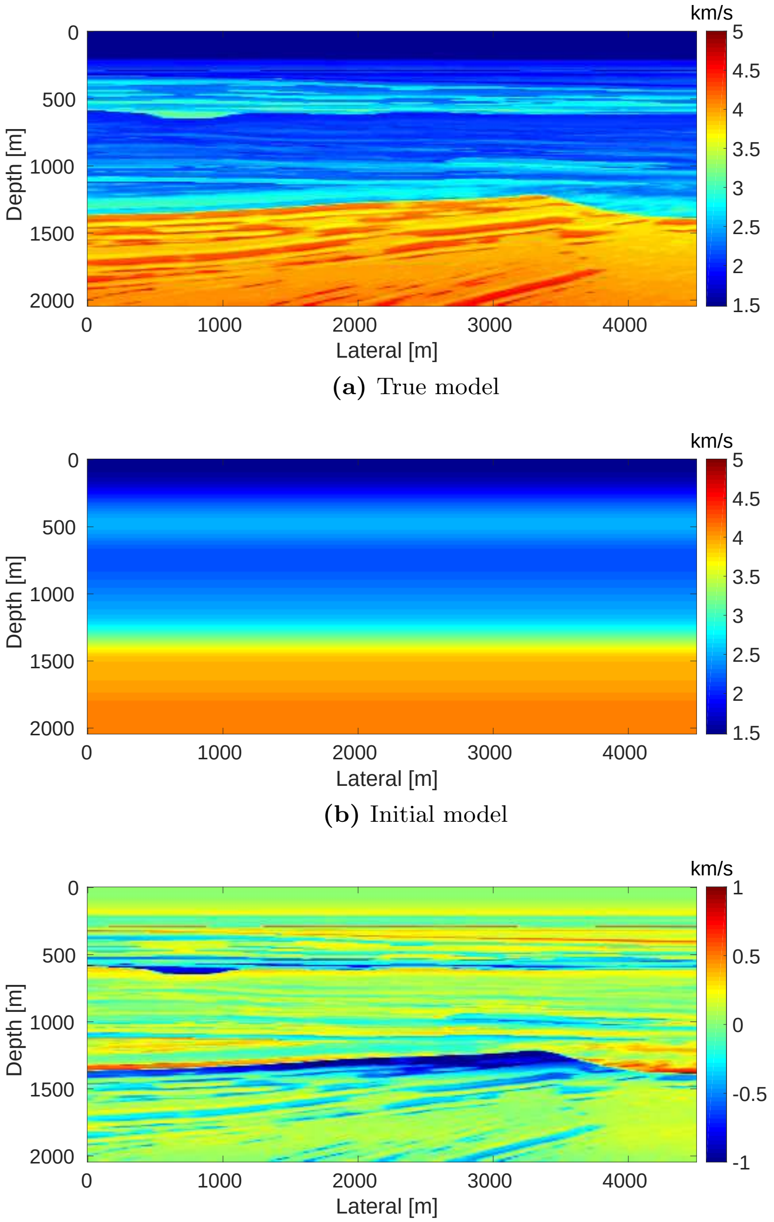
4.5.1 BG model with frequency-domain data

To compare the performance of the two source estimation methods for WRI described in the previous section, we conduct numerical experiments on a part of the BG compass model $\mathbf{m}_{\mathrm{t}}$ shown in Figure 4.3a, which is a geologically realistic model created by BG Group and has been widely used to evaluate performances of different FWI methods (Li et al., 2012; van Leeuwen and Herrmann, 2013a). We will refer to the method that combines $\mathbf{m}$ and α as WRI-SE-MS and the one that combines $\mathbf{u}$ and α as WRI-SE-WS. For our discretization, we use an optimal 9-point finite-difference frequency-domain forward modeling code (Chen et al., 2013). Sources and receivers are positioned at the depth of $z = 20\,\mathrm{m}$ with a source distance of 50 m and a receiver distance of 10 m, resulting a maximum offset of 4.5 km. Data is generated with a source function given by a Ricker wavelet centered at 20 Hz with a time shift of 0.5 s. We do not model the free-surface, so the data contain no free-surface related multiples. As is commonly practiced (see e.g. Pratt (1999)), we perform frequency continuation using frequency bands ranging from $\{2, 2.5, 3, 3.5, 4, 4.5\}$ Hz to $\{27, 27.5, 28, 28.5, 29, 29.5\}$ Hz with an overlap of one frequency between every two consecutive bands. During the inversion, the result of the current frequency band serves as a warm starter for the inversion of the next frequency band. During each iteration, we also apply a point-wise bound constraint (Peters and Herrmann, 2016) to the update to ensure that each gridpoint of the model update remains within a geologically reasonable interval. In this experiment, because the lowest and highest velocities of the true model are 1480 m/s and 4500 m/s, we set the interval of the bound constraint to be [1300, 5000] m/s. The initial model $\mathbf{m}_0$ (Figure 4.3b) is generated by smoothing the original model, followed by an average along the horizontal direction. The difference between the initial and true models is shown in Figure 4.3c. We use this initial model and apply the source estimation method proposed by Pratt (1999) to obtain an initial guess of the source weights for WRI-SE-MS. Because this guess minimizes the difference between the observed and predicted data with the initial model, it can be considered as the best choice, when there is no more additional information. For WRI-SE-WS, as described in the previous sections, it does not need an initial guess of the source weights. To control the computational load, we fix the maximum number of l-BFGS iterations to be 20 for each frequency band. We select the penalty parameter $\lambda = 1e4$

according to the selection criteria in van Leeuwen and Herrmann (2015) that optimizes the performance of WRI.

As a benchmark, we depict the inversion result obtained with the true source weights in Figure 4.4a. Figures 4.4b and 4.4c show the inversion results obtained by WRI-SE-MS and WRI-SE-WS, respectively. Figure 4.5a shows the difference between the results in Figures 4.4a and 4.4b, and Figure 4.5b displays the difference between the results in Figures 4.4a and 4.4c. We observe that while both the results obtained by WRI-SE-MS and WRI-SE-WS are close to the result obtained with the true source weights, the difference shown in Figure 4.5a is almost 10 times larger than that shown in Figure 4.5b. In addition, we compare the true source weights and the estimated source weights obtained with WRI-SE-MS and WRI-SE-WS in Figures 4.6a (phase) and 4.6b (amplitude). From these two figures, we observe that both methods can provide reasonable estimates of the source weights while the estimates with WRI-SE-WS contain smaller errors. These errors found in the source weights may result in the large model differences shown in Figure 4.5a. In Figure 4.7, we display the relative model errors $\frac{\|\mathbf{m}_{\mathrm{t}}-\mathbf{m}\|}{\|\mathbf{m}_{\mathrm{t}}-\mathbf{m}_0\|}$ versus the number of iterations during the three inversions. The dashed, solid, and dotted curves correspond to inversions with the true source weights and those estimates using WRI-SE-MS and WRI-SE-WS. The relative model errors of WRI-SE-WS are almost the same as those using the true source weights, while the errors of WRI-SE-MS are clearly larger than the other ones. Figure 4.8 depicts a comparison of the data residuals corresponding to the inversion results of WRI-SE-MS and WRI-SE-WS. Clearly, the data residual of WRI-SE-WS is much smaller than that of WRI-SE-MS. In Table 4.2, we quantitatively compare the inversion results obtained by WRI-SE-MS and WRI-SE-WS in terms of the final model errors, source errors, and data residuals, which also illustrates the outperformances of WRI-SE-WS over WRI-SE-MS. These observations imply that compared to iteratively updating α, projecting out α and $\mathbf{u}$ together during each iteration can provide more accurate estimates of source weights, which further benefits the inversions of medium parameters.

To evaluate the computational performance of the fast inversion scheme of WRI-SE-WS, we compare the time spent in evaluating one objective for three cases, namely (i) WRI without source estimation; (ii) WRI-SE-WS with the old form, equation 4.10; and (iii) WRI-SE-WS with the new form, equation 4.20. The computational time for each of the three cases is 94 s, 2962 s, and 190 s, respectively. As expected, case (iii) spends twice the amount of time as case (i), because of the additional PDE solves for each frequency and each source. As only one Cholesky factorization is required

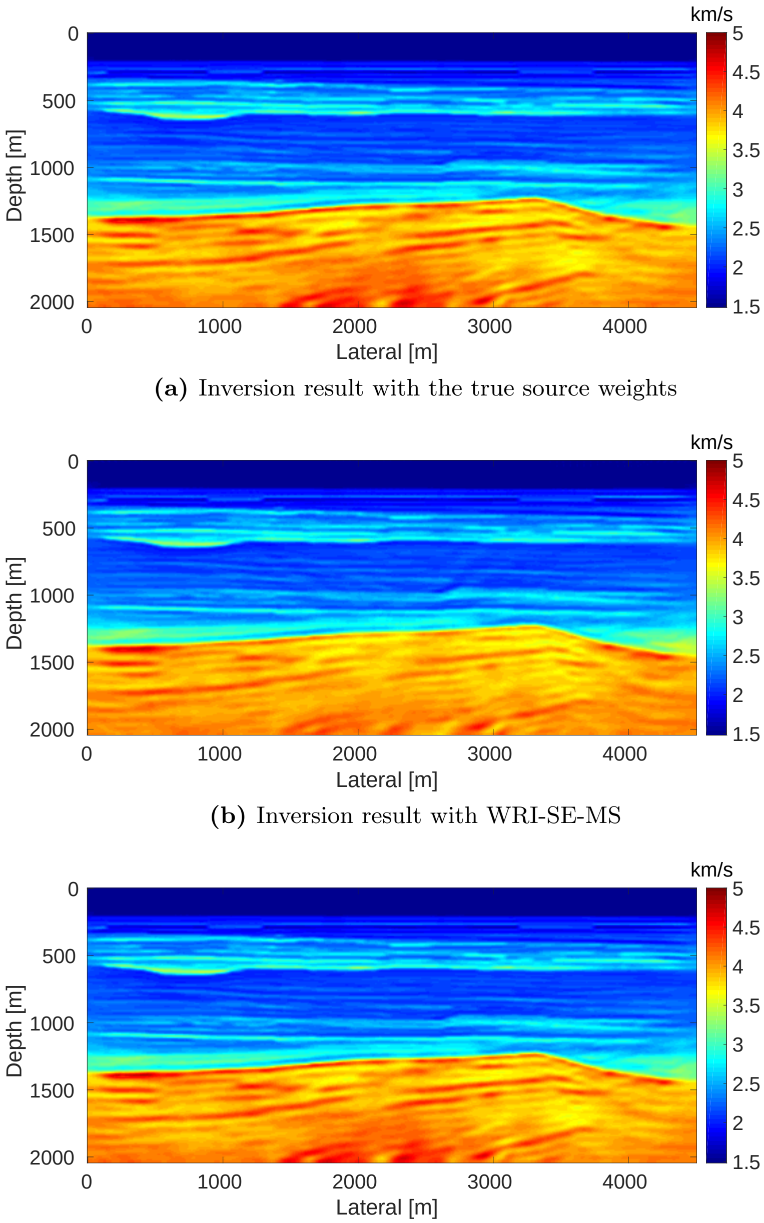


(a) True model

(b) Initial model

(c) Difference between the initial and true models

Figure 4.3: (a) True model, (b) initial model, and (c) the difference between them.

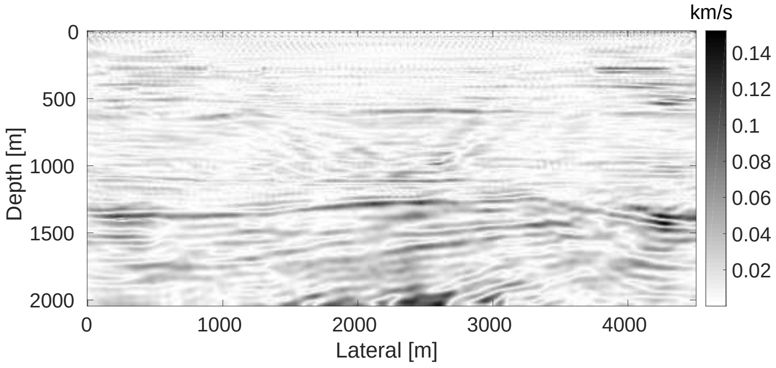


(a) Inversion result with the true source weights

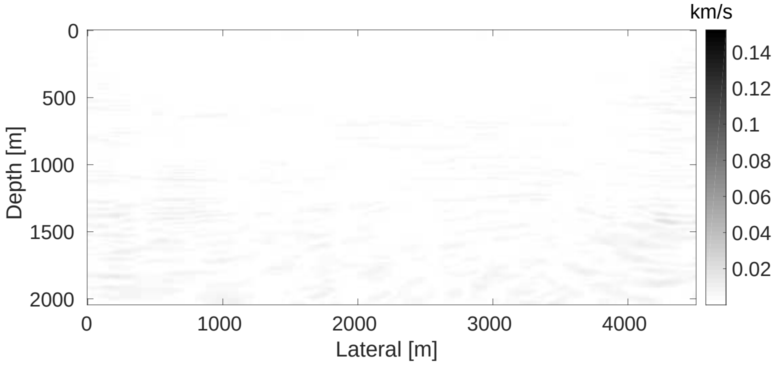
(b) Inversion result with WRI-SE-MS

(c) Inversion result with WRI-SE-WS

Figure 4.4: (a) Inversion result with the true source function. (b) Inversion result with WRI-SE-MS. (c) Inversion result with WRI-SE-WS.



(a) Model difference for WRI-SE-MS

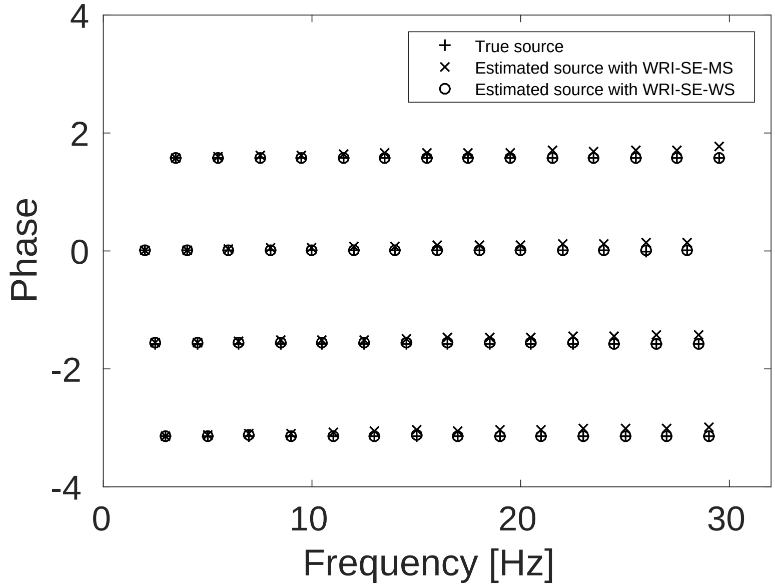


(b) Model difference for WRI-SE-WS

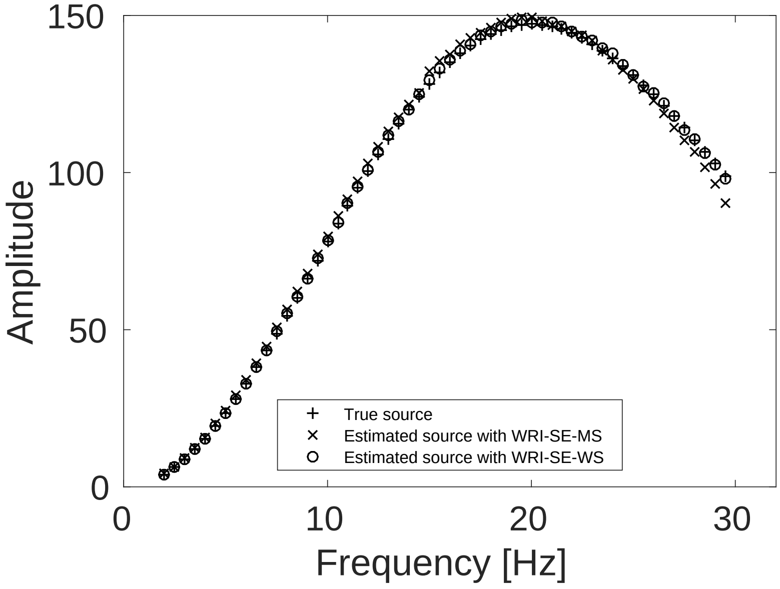
Figure 4.5: (a) Difference between the inversion results using the true source function and the estimated source function by WRI-SE-MS. (b) Difference between the inversion results using the true source function and the estimated source function by WRI-SE-WS.

	WRI-SE-MS	WRI-SE-WS
$\|\mathbf{m}_{\mathrm{t}} - \mathbf{m}_{\mathrm{f}}\|_2$	48	45
$\|\alpha_{\mathrm{t}} - \alpha_{\mathrm{f}}\|_2$	$9.3e2$	$8e1$
$\|\mathbf{d}_{\mathrm{obs}} - \mathbf{d}_{\mathrm{pred}}\|_2$	$6.6e3$	$1.5e3$

Table 4.2: Comparisons between inversion results obtained by WRI-SE-MS and WRI-SE-WS in terms of the final model errors, source errors, and data residuals.



(a) Phase comparison



(b) Amplitude comparison

Figure 4.6: Comparison of the true source function (+) and the estimated source function with WRI-SE-MS (×) and WRI-SE-WS (∘). (a) Phase comparison. (b) Amplitude comparison.

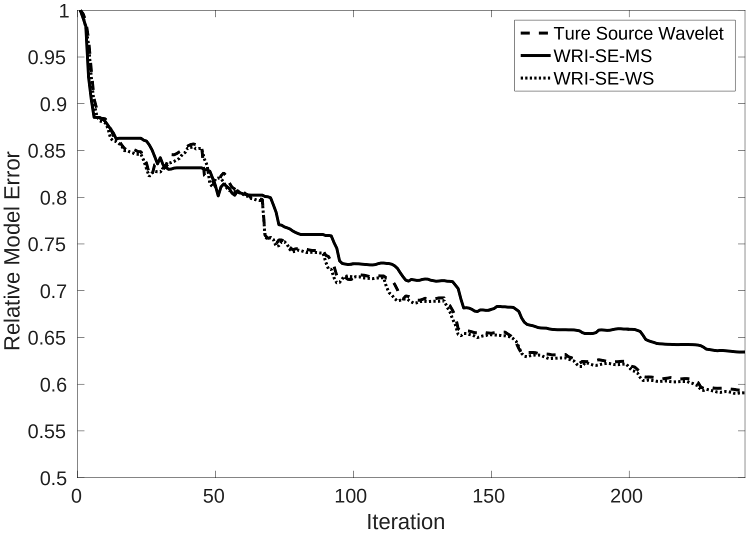


Figure 4.7: Comparison of the relative model errors.

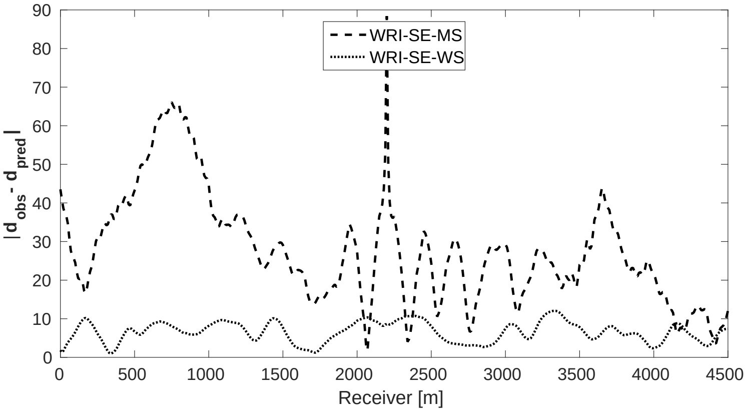


Figure 4.8: Comparison of the data residuals corresponding to the inversion results of WRI-SE-MS (dashed line) and WRI-SE-WS (dotted line).

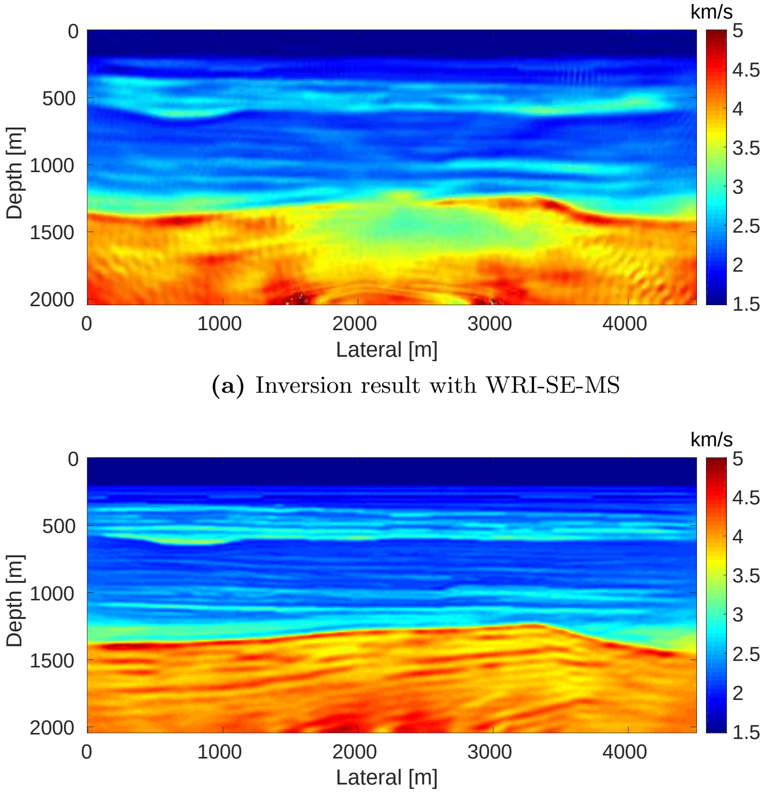
for case (iii), the computational time is only 1/15 of that for case (ii). This result illustrates that the proposed approach can estimate source weights in the context of WRI with a small computational overhead.

4.5.2 BG model with time-domain data

To test the robustness of the two source estimation techniques in less ideal situations, we perform the inversion tests on non-inverse-crime data. For this purpose, we generate time-domain data with a recording time of 4 seconds using iWAVE (Symes et al., 2011) and transform the data from the time domain to the frequency domain for the inversions. The data are generated on uniform grids with a grid size of 6 m, while the inversions are carried out on uniform grids with a coarser grid of 10 m. As a result, modeling errors arise from the differences between the modeling kernels and the grid sizes. All other experimental settings in this example are the same as example 1. Figures 4.9a and 4.9b show the inversion results with WRI-SE-MS and WRI-SE-WS, respectively. From Figure 4.9a, we can observe that method WRI-SE-MS fails to converge to a reasonable solution, as there is a large low-velocity block at the depth of $z = 1500\,\mathrm{m}$, which does not exist in the true model. On the other hand, the result of WRI-SE-WS (Figure 4.9b) is more consistent with the true model. The average model errors presented in the inversion results of WRI-SE-MS and WRI-SE-WS are 0.18 km/s and 0.11 km/s, respectively. Moreover, compared to the true source weights, the amplitudes of the estimated weights obtained with WRI-SE-MS contain very large errors, while the results obtained with WRI-SE-WS are almost identical to the true source weights (see Figure 4.10). Additionally, the residual between the observed and predicted data computed by the inversion results of WRI-SE-WS is much smaller than that of WRI-SE-MS (see Figure 4.11). We can also obtain similar observations from the quantitative comparison presented in Table 4.3. These results imply that compared to WRI-SE-MS, WRI-SE-WS is more robust with respect to the modeling errors.

4.5.3 BG model with noisy data

To test the stability of the proposed two techniques with respect to measurement noise in data, we add 40% Gaussian noise into the data used in example 2. Again, other experimental settings remain the same as in example 2. As expected, due to the noise in the data, both inverted models (Figures 4.12a and 4.12b) contain more noise than the inverted models in example 2. Similar to example 2, the result of WRI-SE-MS still contains a large incorrect



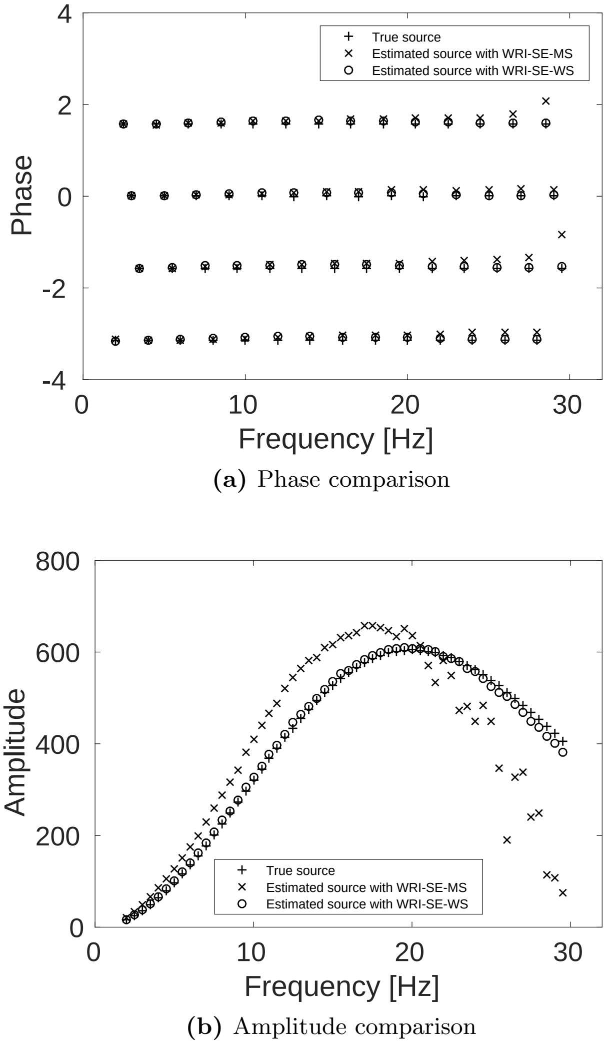
(a) Inversion result with WRI-SE-MS

(b) Inversion result with WRI-SE-WS

Figure 4.9: Inversion results with (a) WRI-SE-MS and (b) WRI-SE-WS.

	WRI-SE-MS	WRI-SE-WS
$\lVert\mathbf{m}_{\rm t}-\mathbf{m}_{\rm f}\rVert_2$	80	53
$\lVert\alpha_{\rm t}-\alpha_{\rm f}\rVert_2$	$1.1e3$	$2.1e2$
$\lVert\mathbf{d}_{\rm obs}-\mathbf{d}_{\rm pred}\rVert_2$	$6.8e3$	$2.4e3$

Table 4.3: Comparisons between inversion results obtained by WRI-SE-MS and WRI-SE-WS in terms of the final model errors, source errors, and data residuals.



(a) Phase comparison

(b) Amplitude comparison

Figure 4.10: Comparison of the true source function (+) and the estimated source function with WRI-SE-MS (×) and WRI-SE-WS (∘). (a) Phase comparison. (b) Amplitude comparison.

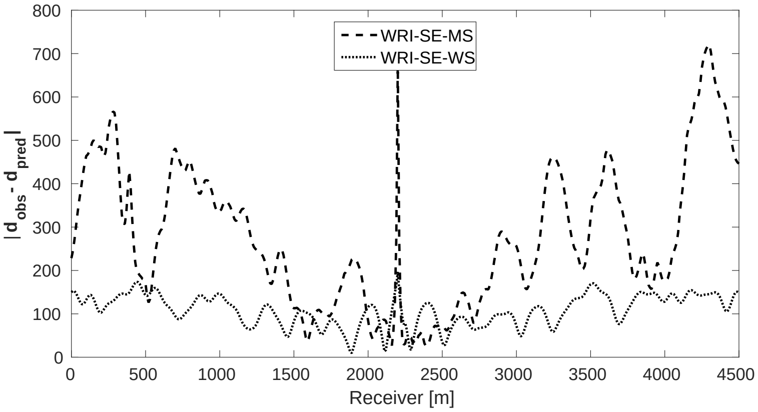
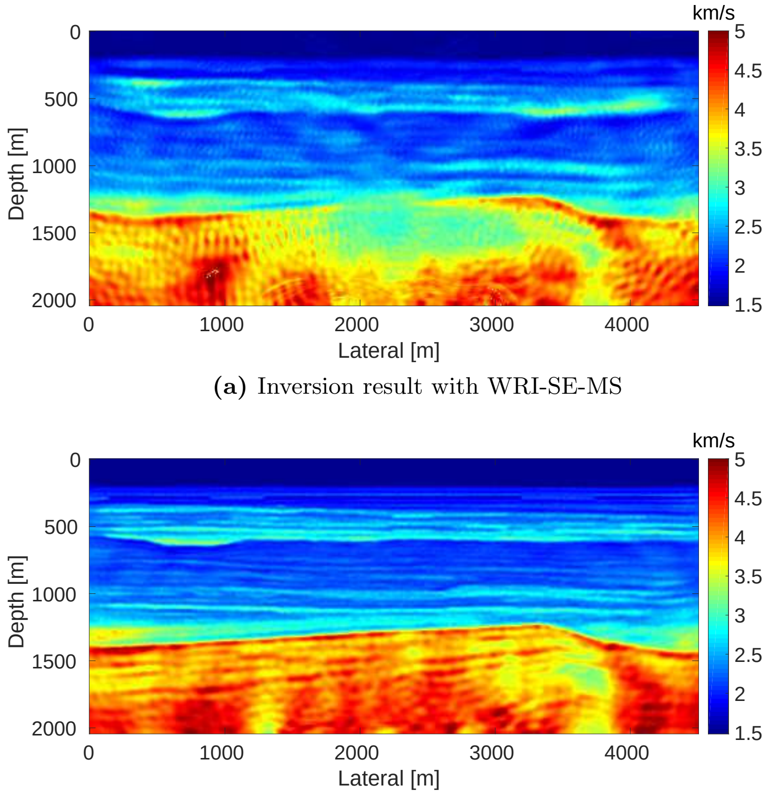


Figure 4.11: Comparison of the data residuals corresponding to the inversion results of WRI-SE-MS (dashed line) and WRI-SE-WS (dotted line).

low-velocity block at the depth of $z = 1500\,\mathrm{m}$, which we do not find in the result of WRI-SE-WS. The final average model errors of WRI-SE-MS and WRI-SE-WS are $0.22\,\mathrm{km/s}$ and $0.16,\mathrm{km/s}$, respectively. A comparison of the true source weights (+), estimated source weights obtained with WRI-SE-MS (×) and WRI-SE-WS (∘) is depicted in Figure 4.13. The estimated source weights with WRI-SE-WS agree with the true source weights much better than those obtained with WRI-SE-MS. We also compare the data residuals corresponding to the inversion results of WRI-SE-WS and WRI-SE-MS in Figure 4.14. Clearly, WRI-SE-MS produces larger data residuals than WRI-SE-WS. Moreover, the quantitative comparison in Table 4.4 illustrates that the inversion results of WRI-SE-WS exhibit smaller model errors, source errors, and data residuals when compared to that of WRI-SE-MS. These observations imply that compared to WRI-SE-MS, WRI-SE-WS is more robust and stable with respect to measurement noise.

4.5.4 Comparison with FWI

Finally, we intend to compare the performances of FWI with source estimation and WRI-SE-WS under bad initial models. We use the same experimental settings as example 1 except with frequency bands ranging from $\{7, 7.5, 8, 8.5, 9, 9.5\}$ Hz to $\{27, 27.5, 28, 28.5, 29, 29.5\}$ Hz and the selection of the penalty parameter $\lambda = 1e0$. During the inversions, we use the initial model displayed in Figure 4.15a. This model is difficult for both FWI



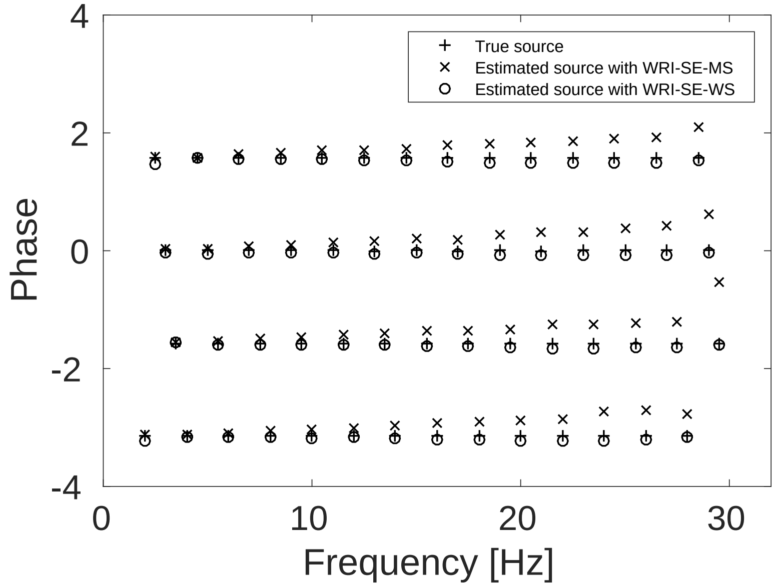
(a) Inversion result with WRI-SE-MS

(b) Inversion result with WRI-SE-WS

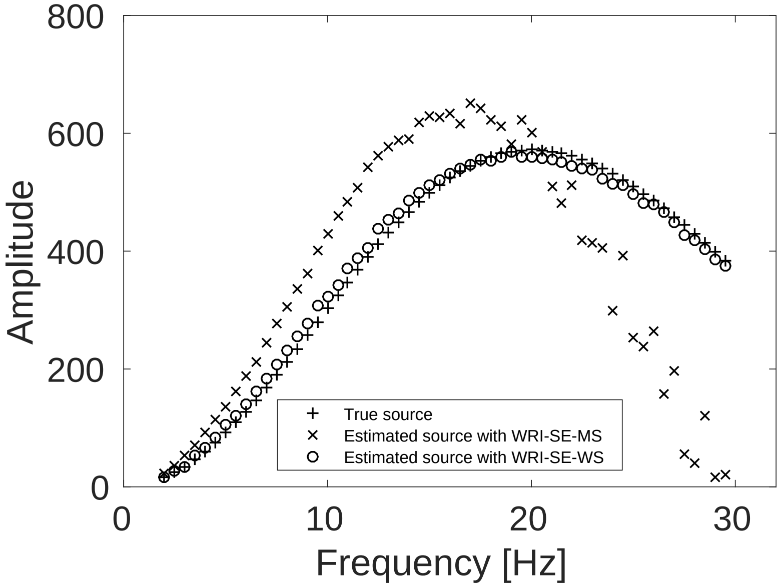
Figure 4.12: Inversion results with (a) WRI-SE-MS and (b) WRI-SE-WS.

	WRI-SE-MS	WRI-SE-WS
$\|\mathbf{m}_{\mathrm{t}} - \mathbf{m}_{\mathrm{f}}\|_2$	96	76
$\|\alpha_{\mathrm{t}} - \alpha_{\mathrm{f}}\|_2$	$1.5e3$	$2.3e2$
$\|\mathbf{d}_{\mathrm{obs}} - \mathbf{d}_{\mathrm{pred}}\|_2$	$2.1e4$	$9.8e3$

Table 4.4: Comparisons between inversion results obtained by WRI-SE-MS and WRI-SE-WS in terms of the final model errors, source errors, and data residuals.



(a) Phase comparison



(b) Amplitude comparison

Figure 4.13: Comparison of the true source function (+) and the estimated source function with WRI-SE-MS (×) and WRI-SE-WS (∘). (a) Phase comparison. (b) Amplitude comparison.

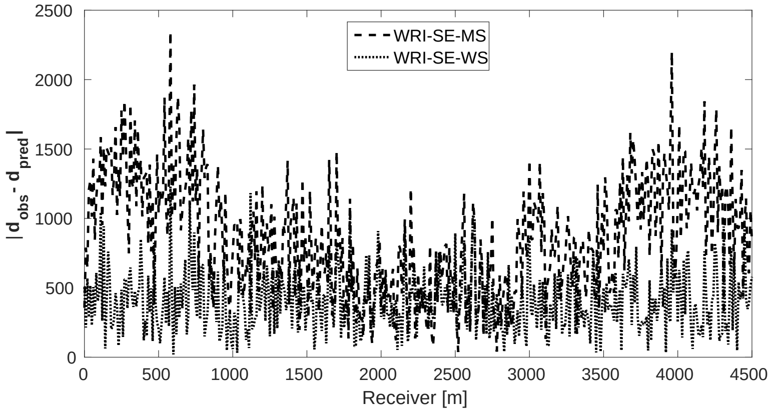
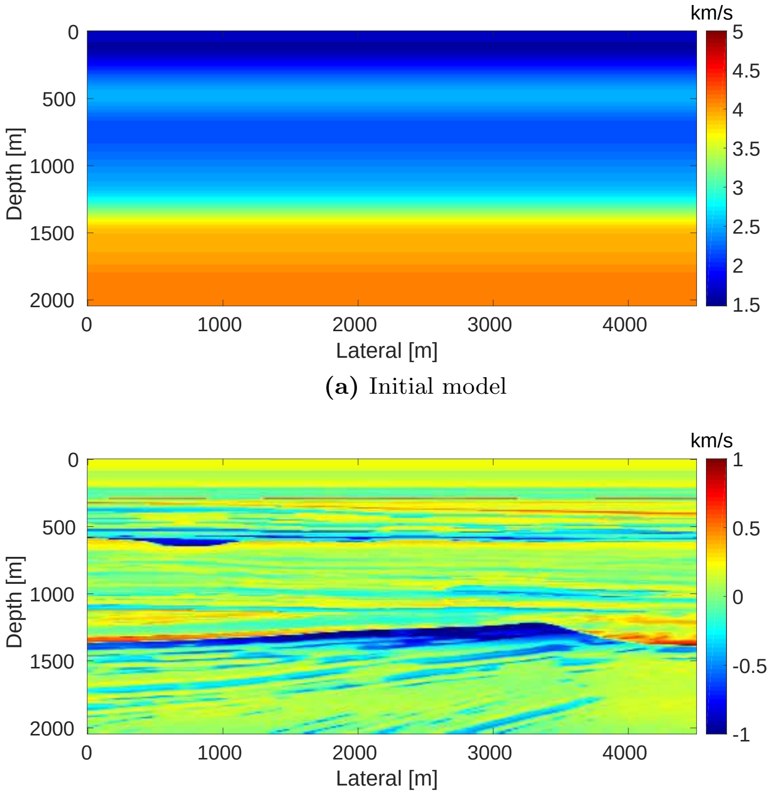


Figure 4.14: Comparison of the data residuals corresponding to the inversion results of WRI-SE-MS (dashed line) and WRI-SE-WS (dotted line).

and WRI with the given frequency range, because in the shallow part of the model, i.e., from the depth of 0 m to 120 m, the velocity of the initial model is 0.2 km/s higher than the true one (shown in Figure 4.15b), which can produce cycle-skipped predicted data shown in Figure 4.16. Moreover, as the maximum offset is 4.5 km, the transmitted waves that the conventional FWI uses to build up long-wavelength structures can only reach the depth of 1.5 km (Virieux and Operto, 2009). When the transmitted data are cycle-skipped, the resulting long-wavelength velocity structures would be erroneous, which would further adversely affect the reconstruction of the short-wavelength velocity structures, especially those below 1.5 km.

Figures 4.17a and 4.17b show the inversion results obtained by WRI and FWI, respectively. As expected, due to the cycle-skipped data and absence of low-frequency data, the conventional FWI fails to correctly invert the velocity in the shallow area where the transmitted waves arrive, i.e., $z \leq 1.5$ km, which subsequently yields larger errors within the inverted velocity in the deep part of the model, i.e., $z > 1.5$ km. On the other hand, WRI mitigates the negative effects of the cycle-skipped data to the inversion and correctly reconstructs the velocity in both areas that can and cannot be reached by the transmitted turning waves. The final average model error of WRI is 0.1 km/s, which is much smaller than that of FWI—0.22 km/s. Moreover, as shown in Table 4.5, the ℓ_2-norm model error of FWI is twice as large as that of WRI. This implies that besides the transmitted waves

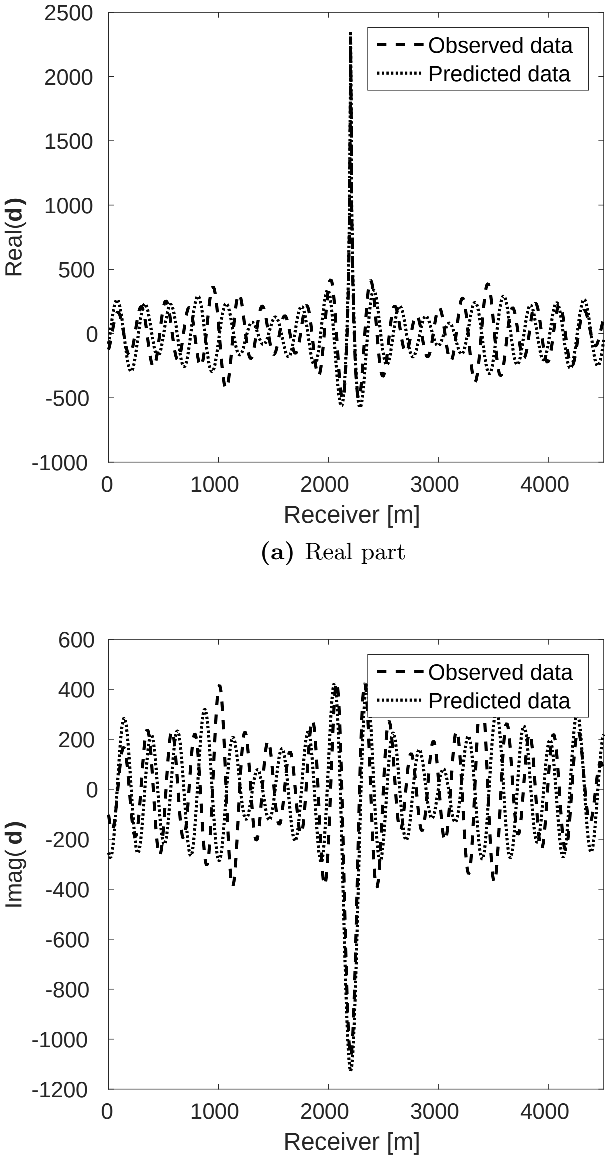


(a) Initial model

(b) Difference between the initial and true models

Figure 4.15: (a)Initial model and (b) difference between the initial model and true models.

WRI also uses the reflection waves to invert the velocity model. Clearly, the main features and interfaces in the model are reconstructed correctly by WRI. Figure 4.18 displays three vertical cross-sections through the model. Compared to FWI, WRI provides a result that matches the true model much better. Figure 4.19 shows the comparison of the estimated source weights obtained by FWI and WRI. From the comparisons in Figures 4.19a and 4.19b, we observe that compared to FWI, WRI provides estimates of source weights with mildly better phases and significantly better amplitudes. The ℓ_2-norm source error of FWI is much larger than that of WRI as shown in Table 4.5. This is not surprising as the quality of the source estimation and the inverted velocity model are closely tied to each other. Figure 4.20 shows the comparisons between the observed data and predicted data com-



(a) Real part

(b) Imaginary part

Figure 4.16: Comparisons between the observed (dashed line) and predicted data (dotted line) computed with the initial model for a source located at the center of the model. (a) Real part comparison. (b) Imaginary part comparison.

	FWI	WRI
$\|\mathbf{m}_{\mathrm{t}} - \mathbf{m}_{\mathrm{f}}\|_2$	100	50
$\|\alpha_{\mathrm{t}} - \alpha_{\mathrm{f}}\|_2$	$1.3e3$	$6e1$
$\|\mathbf{d}_{\mathrm{obs}} - \mathbf{d}_{\mathrm{pred}}\|_2$	$3.8e4$	$1.7e3$

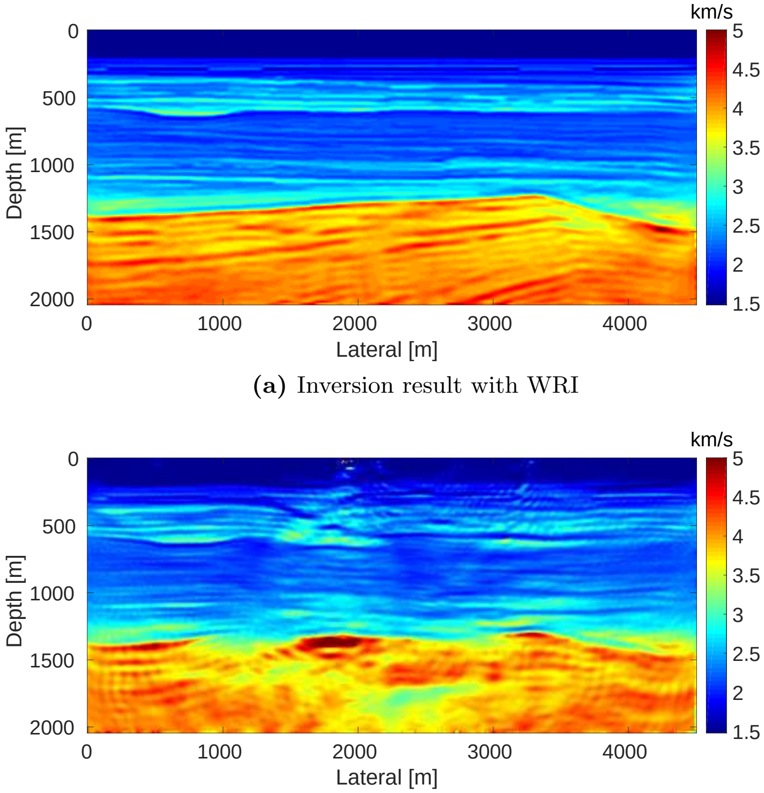
Table 4.5: Comparisons between inversion results obtained by FWI and WRI in terms of the final model errors, source errors, and data residuals.

puted with the final results of WRI and FWI. The predicted data computed from the WRI inversion (Figures 4.20a and 4.20b) almost matches the observed data, while the mismatch of that from the FWI inversion is much larger (Figures 4.20c and 4.20d), which can also be found in the quantitative comparison in Table 4.5. All these results imply that besides providing a better velocity reconstruction with less model errors compared to the true model, WRI also produces a better estimate of the source weights with much less errors compared to the true source weights.

4.6 Discussions

Source functions are essential for wave-equation-based inversion methods to conduct a successful reconstruction of the subsurface structures. Because the correct source functions are typically unavailable in practical applications, an on-the-fly source estimation procedure is necessary for the inversion. In this study, we proposed an on-the-fly source estimation technique for the recently developed WRI by adapting its variable projection procedure for the wavefields to include the source functions. Through simultaneously projecting out the wavefields and source functions, we obtained a reduced objective function that only depends on the medium parameters. As a result, we are able to accomplish the inversion without prior information of the source functions.

The main computational cost of the proposed source estimation method lies within the procedure of projecting out the wavefields and source weights, in which we have to invert a potentially ill-conditioned source-dependent matrix. For ocean bottom node and land acquisitions where all sources “see” the same receivers, we applied the block matrix inversion formula and arrived at a new inversion scheme that only involves inverting a source-independent matrix. This new scheme brings benefits to both direct and iterative solvers

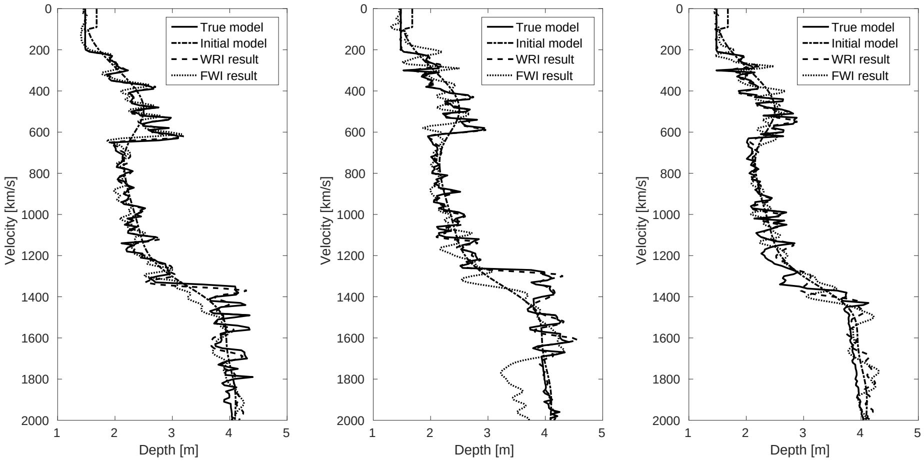


(a) Inversion result with WRI

(b) Inversion result with FWI

Figure 4.17: Inversion results with (a) WRI and (b) FWI.

when compared to the direct inversion of the original source-dependent matrix. For 2D problems, we illustrated that without losing accuracy, the inversion of the proposed scheme only needs one Cholesky factorization for each frequency, while the direct inversion requires n_{src}. Indeed, this benefit does not only apply to the Cholesky method, but also to other faster direct solvers including the multifrontal massively parallel sparse direct solver (MUMPS, Amestoy et al. (2016)), which can further speed up the projection procedure. When solving 3D problems with iterative solvers, the proposed inversion scheme reduces the number of preconditioners for each frequency from n_{src} to one; therefore, it significantly reduces the computational complexity. Furthermore, the source-independent matrix required to invert is the same one as in the conventional WRI without source estimation. Thus, we avoid iteratively inverting the potentially ill-conditioned

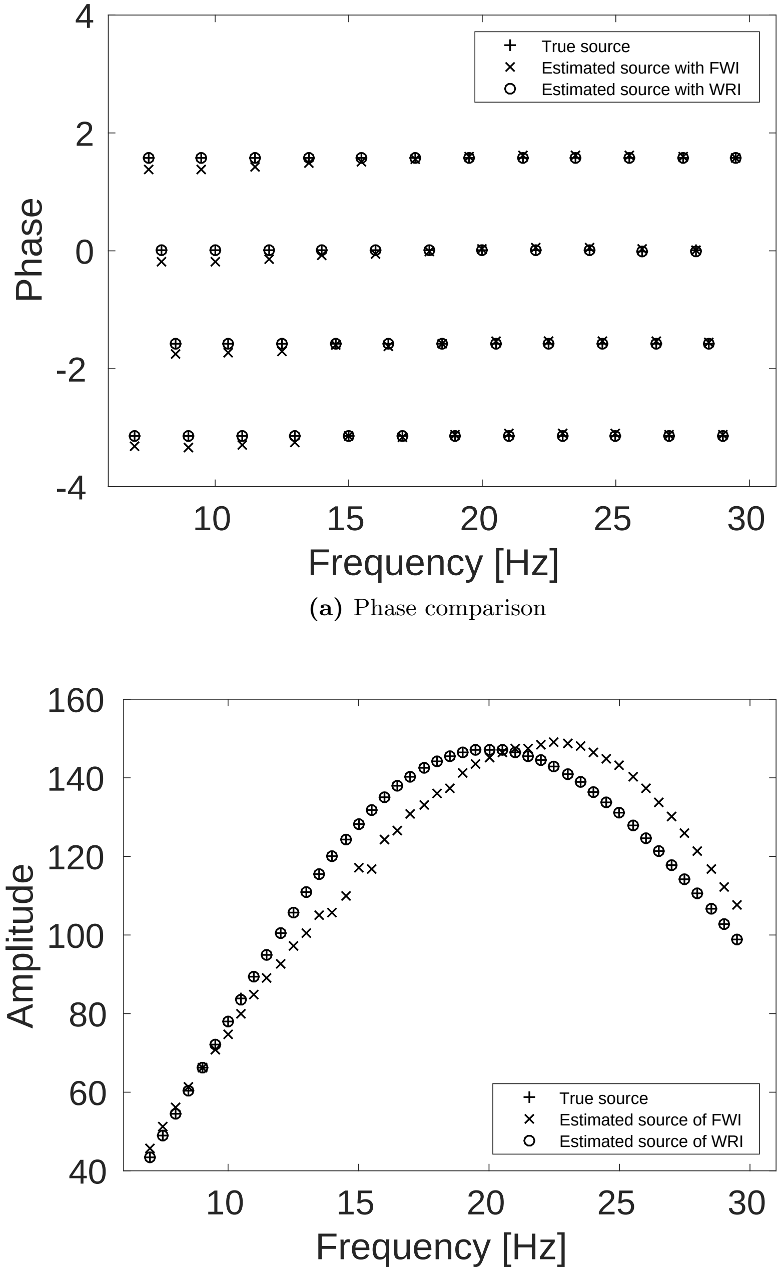


(a) $x = 1000\,\mathrm{m}$ **(b)** $x = 2500\,\mathrm{m}$ **(c)** $x = 4000\,\mathrm{m}$

Figure 4.18: Vertical profiles of the true model (solid line), initial model (dashdot line), and the inversion results with WRI (dashed line) and FWI (dotted line) at (a) $x = 1000\,\mathrm{m}$, (b) $x = 2500\,\mathrm{m}$, and (c) $x = 4000\,\mathrm{m}$.

source-dependent matrix and can straightforwardly apply any efficient and scalable methods designed for the conventional WRI (Peters et al., 2015) to the proposed approach, which allows us to extend the application of this approach to 3D frequency-domain waveform inversions.

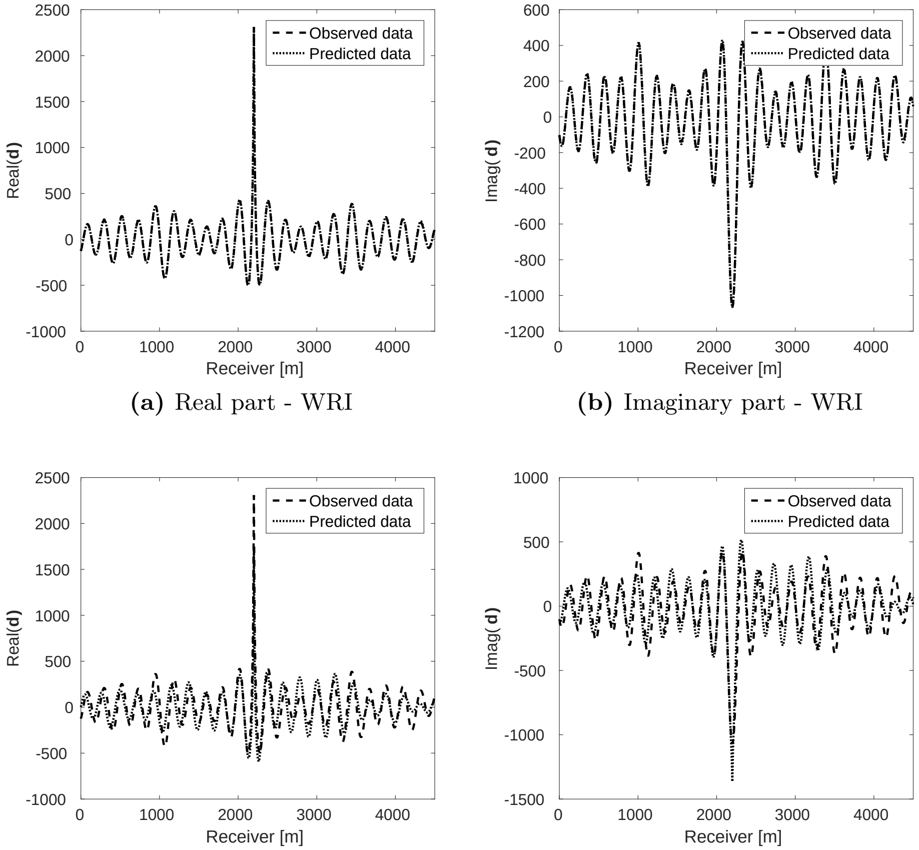
Aside from benefiting inversions with ocean bottom node and land acquisitions, the proposed scheme may also benefit inversions with conventional 3D marine towed-streamer acquisitions where all sources “see” different receivers. In this case, we may lose the benefit of reducing the number of preconditioners due to the acquisition geometry, however, the benefit that avoids the direct inversion of the potentially ill-conditioned source-dependent matrix remains. As a result, we still can straightforwardly apply any preconditioners designed for the conventional WRI without source estimation to solve the proposed scheme. With carefully designed preconditioners, the computational cost of the proposed WRI with source estimation can be roughly equal to that of standard 3D frequency-domain FWI using iterative solvers, because the total numbers of the iterative PDE solves required by the two approaches are approximately the same. As a result, it should



(a) Phase comparison

(b) Amplitude comparison

Figure 4.19: Comparison of the true source function (+) and the estimated source function with FWI (×) and WRI (∘). (a) Phase comparison. (b) Amplitude comparison.



(a) Real part - WRI

(b) Imaginary part - WRI

(c) Real part - FWI

(d) Imaginary part - FWI

Figure 4.20: Comparisons between the observed (dashed line) and predicted data (dotted line) computed with the final results of WRI and FWI. (a) Real part and (b) imaginary part of WRI result. (c) Real part and (d) imaginary part of FWI result.

be viable to apply the proposed WRI with source estimation to inversions with the conventional 3D marine towed-streamer acquisitions.

While the presented examples illustrated the feasibility of the proposed approach—WRI with source estimation—for frequency-domain inversions, its extension to the time-domain inversions is not trivial. The main computational challenge lies in the fact that in the time domain, the projection procedure requires us to solve a very large linear system, whose size is n_{time} (number of time slices) times larger than that in the frequency domain. As a result, the computational and storage cost in the time domain is larger than that in the frequency domain. Moreover, the time-domain source function is a time series rather than a complex number in the frequency domain. In order to estimate the time series successfully, we may need additional constraints, such as the duration of the source function, to regularize the unknown source functions. Due to the two facts, the extension of the proposed approach to the time-domain inversions requires further investigations.

In the examples of this chapter, we use a single λ for all the frequencies. However, as the amplitudes of data and the Helmholtz matrices differ from frequency to frequency, we may use different λ's for different frequencies to conduct better inversions. Moreover, as introduced by Esser et al. (2018), it would be a good strategy to conduct the inversion with multiple cycles of warm-started WRI during which the penalty parameter λ increases. We will discuss the strategy of selecting λ in the following chapters.

4.7 Conclusions

In this chapter, we showed that the recently proposed wavefield-reconstruction inversion method can be modified to handle unknown source situations. In the proposed modification of wavefield-reconstruction inversion, we considered the source functions as unknown variables and updated them jointly with the medium and wavefields parameters. To update the three unknowns, we proposed an optimization strategy based on the variable projection method. During each iteration of the inversion, for fixed medium parameters, the objective function is quadratic with respect to both wavefields and source functions. This fact motivates us to use the variable projection method to first project out the wavefields and source functions simultaneously, and then to update the medium parameters. As a result, we obtained an objective that does not depend on the wavefields and the source functions. Numerical experiments illustrated that equipped with the proposed source estimation method wavefield-reconstruction inversion can produce accurate inversion results without prior knowledge of the true source weights. Moreover, this method can also produce reliable results when the observations contain noise.

We also compared the proposed on-the-fly source estimation technique for wavefield-reconstruction inversion to the conventional source estimation technique for full-waveform inversion. The numerical result illustrated that by expanding the search space and the inexact PDE-fitting, the proposed source estimation technique for wavefield-reconstruction inversion is less sensitive to inaccurate initial models and can start the inversion with data without low-frequency components.

Chapter 5

Wavefield reconstruction inversion with source estimation and convex constraints — application to the 2D Chevron 2014 synthetic blind-test dataset

5.1 Introduction

In the previous chapter, we have proposed a modified wavefield-reconstruction inversion approach that is equipped with an on-the-fly source estimation technique (WRI-SE). In this chapter, we aim to investigate its effectiveness for realistic problems. For this purpose, we apply the WRI-SE approach to a highly realistic 2D marine elastic dataset, which is for full-waveform inversion blind tests and designed by Chevron in 2014 (Qin et al., 2015). When inverting this dataset, several challenges remain: (1) frequency-band-limited data without available low frequencies below 4–5 Hz; (2) unclear relations between P-velocity and S-velocity; and (3) bad starting model that is far away from the true model. To address these challenges, we extend the WRI-SE approach by involving multiple pieces of prior information about the Earth, such as limits on the velocity or minimum smoothness conditions on the model, into the optimization framework.

Wave-equation-based inversions including full-waveform inversion (FWI) and wavefield-reconstruction inversion (WRI) may require some forms of regularization to yield solutions, which conform with our knowledge about the Earth at the survey location. We propose a constrained optimization framework that incorporates the WRI-SE approach with multiple pieces of prior information about the geology. To achieve this task, we first mathematically represent different prior information of the model parameters by different convex sets. Then we constrain the model parameters in the optimization problem of WRI-SE by projection onto intersections of these convex sets, which is similar to Baumstein (2013). This results in a constrained optimization problem that does not involve quadratic penalties in the objective function. We illustrate that this constrained optimization framework provides us with the flexibility to straightforwardly incorporate several preexisting regularization ideas, such as Tikhonov regularization and gradient filtering (Brenders and Pratt, 2007), into one framework without significant additional computation.

This chapter is organized as follows. First, we present the constrained optimization problem that is incorporated with the objective function of WRI-SE. Then we introduce the optimization algorithm that solves the constrained optimization problem. Finally, we show the results on the Chevron 2014 data set to demonstrate the effectiveness of the proposed method. To highlight the effect of the proposed approach, we apply no data preprocessing.

5.2 WRI-SE with regularization by projection onto the intersection of convex sets

In search of a flexible and easy to use framework, which can incorporate most existing ideas in the geophysical community to regularize the waveform inversion problem, we propose to solve constrained optimization problems of the form

$$\underset{\mathbf{m}}{\text{minimize}}\, f(\mathbf{m}) \quad \text{s.t.} \quad \mathbf{m} \in \mathcal{C}_1 \bigcap \mathcal{C}_2, \tag{5.1}$$

where the objective function $f(\mathbf{m})$ corresponds to the aforementioned reduced objective function for WRI with source estimation in equation 4.12:

$$\begin{aligned} f(\mathbf{m}) = \frac{1}{2} \sum_{i=1}^{n_{\text{src}}} \sum_{j=1}^{n_{\text{freq}}} \big(&\|\mathbf{P}_i \overline{\mathbf{u}}_{i,j}(\mathbf{m}) - \mathbf{d}_{i,j}\|_2^2 \\ &+ \lambda^2 \|\mathbf{A}_j(\mathbf{m}) \overline{\mathbf{u}}_{i,j}(\mathbf{m}) - \overline{\alpha}_{i,j}(\mathbf{m}) \mathbf{e}_{i,j}\|_2^2 \big), \end{aligned} \tag{5.2}$$

with

$$\begin{bmatrix}\overline{\mathbf{u}}_{i,j}(\mathbf{m})\\ \overline{\alpha}_{i,j}(\mathbf{m})\end{bmatrix} = \begin{bmatrix}\lambda^2\mathbf{A}_j(\mathbf{m})^\top\mathbf{A}_j(\mathbf{m}) + \mathbf{P}_i^\top\mathbf{P}_i & -\lambda^2\mathbf{A}_j(\mathbf{m})^\top\mathbf{e}_{i,j}\\ -\lambda^2\mathbf{e}_{i,j}^\top\mathbf{A}_j(\mathbf{m}) & \lambda^2\mathbf{e}_{i,j}^\top\mathbf{e}_{i,j}\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{P}_i\mathbf{d}_{i,j}\\ \mathbf{0}\end{bmatrix}. \tag{5.3}$$

$\mathcal{C}_1$ and $\mathcal{C}_2$ are convex sets (more than 2 sets can be used). Intuitively, the line connecting any two points in a convex set is also in the set—i.e., for all $\mathbf{x}$,$\mathbf{y} \in \mathcal{C}$ the following holds:

$$c\mathbf{x} + (1-c)\mathbf{y} \in \mathcal{C},\ 0 \leq c \leq 1. \tag{5.4}$$

Each set represents a known piece of information about the Earth, yielding a constraint that the estimated model has to satisfy. The model is required to be in the intersection of the two sets, denoted by $\mathcal{C}_1 \bigcap \mathcal{C}_2$, which is also convex. To summarize, solving Problem 5.1 means that we try to minimize the objective function while satisfying the constraints, which represent the known information about the geology. Below are two useful convex sets that are also used in the examples. Bound-constraints can be defined elementwise as

$$\mathcal{C}_1 \equiv \{\mathbf{m} \mid \mathbf{b}_l \leq \mathbf{m} \leq \mathbf{b}_u\}. \tag{5.5}$$

If a model is in this set, then the value of every model parameter is between a lower bound $\mathbf{b}_l$ and upper bound $\mathbf{b}_u$. These bounds can vary per model parameter and can, therefore, incorporate spatially varying information about the bounds. Bound constraints can also incorporate a reference model by limiting the maximum deviation from this reference model as: $\mathbf{b}_l = \mathbf{m}_{\text{ref}} - \delta\mathbf{m}$. The projection onto $\mathcal{C}_1$, denoted as $\mathcal{P}_{\mathcal{C}_1}$, is the elementwise median

$$\mathcal{P}_{\mathcal{C}_1}(\mathbf{m}) = \text{median}\{\mathbf{b}_l, \mathbf{m}, \mathbf{b}_u\}. \tag{5.6}$$

The second convex set used here is a subset of the wavenumber domain representation of the model (spatial Fourier transformed). This convex set can be defined as

$$\mathcal{C}_2 \equiv \{\mathbf{m} \mid \mathbf{E}^*\mathbf{F}^*(\mathbf{I} - \mathbf{S})\mathbf{F}\mathbf{E}\mathbf{m} = 0\}, \tag{5.7}$$

which is defined by a series of matrix multiplications. First, we apply the mirror-extension operator $\mathbf{E}$ to the model $\mathbf{m}$ to prevent wrap-around effects due to applying linear operations in the wavenumber domain. Next, the mirror extended model is spatially Fourier transformed using $\mathbf{F}$. In this domain, we require the wavenumber content outside a certain range to be zero. This is enforced by the diagonal (windowing) matrix $\mathbf{S}$. This represents information that the model should have a certain minimum smoothness. In case we expect an approximately horizontally layered medium without sharp contrasts, we can include this information by requiring the wavenumber content of the model to vanish outside of an elliptical zone around the zero wavenumber point (see also Brenders and Pratt (2007)). In other words, more smoothness in one direction is required than in the other direction. The adjoint of $\mathbf{F}$, ($\mathbf{F}^*$) transforms back from the wavenumber to the physical domain, and $\mathbf{E}^*$ goes from mirror-extended domain to the original model domain. The projection onto $\mathcal{C}_2$, denoted as $\mathcal{P}_{\mathcal{C}_2}$, is given by

$$\mathcal{P}_{\mathcal{C}_2}(\mathbf{m}) = \mathbf{E}^*\mathbf{F}^*\mathbf{S}\mathbf{F}\mathbf{E}\mathbf{m}. \tag{5.8}$$

A basic way to solve Problem 5.1 is to use the projected-gradient algorithm, which has the main step

$$\mathbf{m}_{k+1} = \mathcal{P}_{\mathcal{C}}(\mathbf{m}_k - \gamma \nabla_{\mathbf{m}} f(\mathbf{m})), \tag{5.9}$$

at the k^{th} iteration with a step length γ. $\mathcal{P}_{\mathcal{C}} = \mathcal{P}_{\mathcal{C}_1 \bigcap \mathcal{C}_2}$ is the projection onto the intersection of all convex sets that the model is required to be in. "The projection onto" means that we need to find the point, which is in both sets and closest to the currently proposed model estimate generated by $(\mathbf{m}_k - \gamma \nabla_{\mathbf{m}} f(\mathbf{m}))$. To find such a point, defined mathematically as

$$\mathcal{P}_{\mathcal{C}}(\mathbf{m}) = \arg\min_{\mathbf{x}} \|\mathbf{x} - \mathbf{m}\|_2 \quad \text{s.t.} \quad \mathbf{x} \in \mathcal{C}_1 \bigcap \mathcal{C}_2, \tag{5.10}$$

we can use a splitting method such as the Dykstra's projection algorithm (Bauschke and Borwein, 1994). This algorithm finds the projection onto the intersection of the two sets by projecting onto each set separately. This is a cheap and simple algorithm and enables projections onto complicated intersections of sets. The algorithm is given by Algorithm 4.

Algorithm 4 Dykstra's projection algorithm.

```
$\mathbf{x}_0 = \mathbf{m}$, $\mathbf{p}_0 = \mathbf{0}$, $\mathbf{q}_0 = \mathbf{0}$
For $l = 0, 1, \ldots$
    $\mathbf{y}_l = \mathcal{P}_{\mathcal{C}_1}(\mathbf{x}_l + \mathbf{p}_l)$
    $\mathbf{p}_{l+1} = \mathbf{x}_l + \mathbf{p}_l - \mathbf{y}_l$
    $\mathbf{x}_{l+1} = \mathcal{P}_{\mathcal{C}_2}(\mathbf{y}_l + \mathbf{q}_l)$
    $\mathbf{q}_{l+1} = \mathbf{y}_l + \mathbf{q}_l - \mathbf{x}_{l+1}$
End
```

While gradient-projections is a solid approach, it can also be relatively

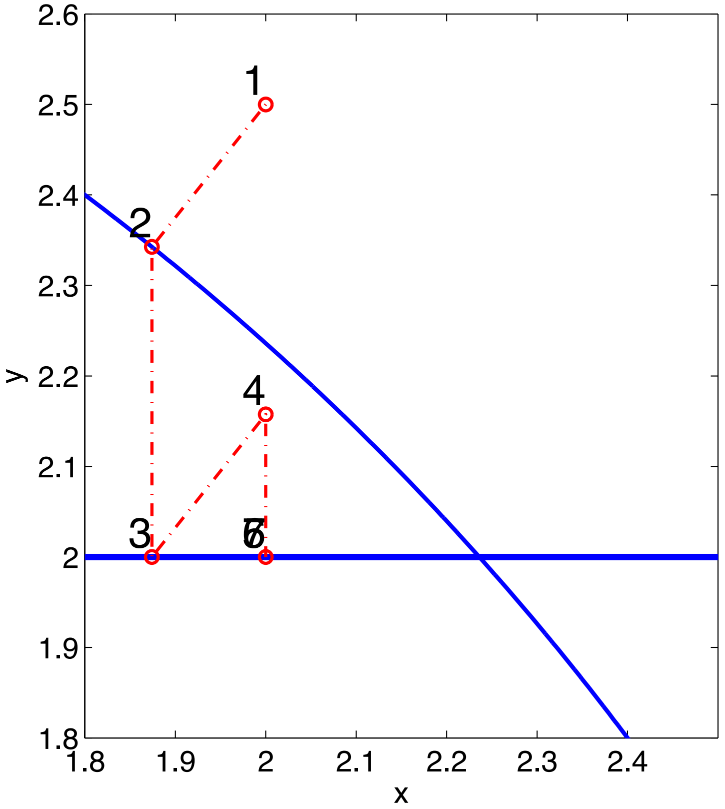


Figure 5.1: The trajectory of Dykstra's algorithm for a toy example with constraints $y \leq 2$ and $x^2 + y^2 \leq 3$. Iterates 5, 6 and 7 coincide; the algorithm converged to the point closest to point number 1 and satisfying both constraints.

slowly converging. A potentially much faster algorithm is the class of projected Newton-type algorithms (PNT) (see for example Schmidt et al., 2012). Standard Newton-type methods iterate,

$$\mathbf{m}_{k+1} = \mathbf{m}_k - \gamma \mathbf{B}^{-1} \nabla_{\mathbf{m}} f(\mathbf{m}), \tag{5.11}$$

with $\mathbf{B}$ an approximation of the Hessian. At every iteration, the PNT method finds the new model by solving the constrained quadratic problem

$$\mathbf{m}_{k+1} = \underset{\mathbf{m} \in \mathcal{C}_1 \bigcap \mathcal{C}_2}{\arg\min} \; Q(\mathbf{m}), \tag{5.12}$$

where the local quadratic model is given by

$$Q(\mathbf{m}) = f(\mathbf{m}_k) + (\mathbf{m} - \mathbf{m}_k)^* \nabla_{\mathbf{m}} f(\mathbf{m}_k) + (\mathbf{m} - \mathbf{m}_k)^* \mathbf{B}_k (\mathbf{m} - \mathbf{m}_k). \quad (5.13)$$

The solution of Problem 5.12 does not involve additional PDE solves and is different from a simple projection of the Newton-type update. The PNT algorithm needs the objective function value, its gradient, approximation to the Hessian, and an algorithm to solve the constrained quadratic sub-problem (Problem 5.12), which in turn needs an algorithm (Dykstra) to solve the projection problem (Problem 5.10). The PNT sub-problem can be solved in many ways. We select the alternating direction method of multipliers (ADMM) algorithm (Eckstein and Yao, 2012), which is given by Algorithm 5.

Algorithm 5 ADMM.

```
1. $\mathbf{x}_0 = \mathbf{m}_k$, $\mathbf{z}_0 = \mathbf{x}_0$, $\mathbf{y}_0 = \mathbf{0}$
2. For $l = 0, 1, \ldots$
3.     $\mathbf{x}_{l+1} = \arg\min_{\mathbf{x}} \left( Q(\mathbf{x}) + (\kappa/2) \|\mathbf{x} - \mathbf{z}_l + \mathbf{y}_l\|_2^2 \right)$
4.     $\mathbf{z}_{l+1} = \mathcal{P}_{\mathcal{C}_1 \bigcap \mathcal{C}_2}(\mathbf{x}_{l+1} + \mathbf{y}_l)$ by Dykstra
       $\mathbf{y}_{l+1} = \mathbf{y}_l + \mathbf{x}_{l+1} - \mathbf{z}_{l+1}$
5. End
```

In Algorithm 5, the scalar κ is a penalty parameter. ADMM can solve Problem 5.12 very efficiently once provided an approximate Hessian $\mathbf{B}$ that is easy to invert. In this algorithm, we use the diagonal approximate of the Gauss-Newton Hessian given by van Leeuwen and Herrmann (2013b) to construct the matrix $\mathbf{B}$, which is much easier to invert than a discretized Helmholtz system. The Dykstra-splitting algorithm solves a sub-problem inside the ADMM algorithm, which is also a splitting algorithm.

PNT has a similar computational cost as the standard projected-gradient algorithm—the same number of Helmholtz systems need to be solved. The additional cost is low because the projections are cheap to compute and the inversion of the diagonal approximate Gauss-Hessian matrix is also cheap.

The final optimization algorithm is given by Algorithm 6 and can be best described as Dykstra's algorithm within ADMM within a projected-Newton type method. During each iteration, the constraints are satisfied exactly.

Algorithm 6 Projected-Newton type based waveform inversion with projections onto convex sets.

```
1. Define convex sets $\mathcal{C}_1$, $\mathcal{C}_2$, ...
2. Set up Dykstra's algorithm to solve Problem 5.10
3. Set up ADMM to solve Problem 5.12
4. $\text{minimize}_{\mathbf{m}} f(\mathbf{m}) \quad \text{s.t.} \quad \mathbf{m} \in \mathcal{C}_1 \bigcap \mathcal{C}_2$ using PNT
```

5.3 Practical considerations

The above projected Newton-type approach with convex constraints provides the formal optimization framework with which we will carry out our experiments. Before demonstrating the performance of our approach, we discuss a number of practical considerations to make the inversion successful.

Frequency continuation. We divide the frequency spectrum into overlapping frequency bands of increasing frequency. This multiscale continuation makes the approach more robust against cycle skips.

Selection of the smoothing parameter. Although there is no penalty parameter to select in our method, we do need to define the convex sets onto which we project. The minimum-smoothness constraints require two parameters—one that indicates how much smoothness is required and one that indicates the difference between horizontal and vertical smoothness. At the start of the first frequency batch, these parameters are chosen such that the minimum smoothness constraints correspond to the smoothness of the initial guess and we add some “room” to update the model to less smooth model. This is the only user intervention of our regularization framework. Subsequent frequency batches multiply the selected parameters by the empirical formula $d\frac{f_{\max}}{v_{\min}}$, where d is the grid-point spacing, $f_{\max}$ the maximum frequency of the current batch, and $v_{\min}$ the minimum velocity in the current model. This automatically adjusts the selected parameters for changing grids, frequency, and velocity.

Data fit for Wavefield Reconstruction Inversion. Our inversions are carried out with WRI-SE, which solves for wavefields that fit both the data and the wave equation for the current model iterate in a least-squares sense. Since the data are noisy, we chose a relatively large λ so that we do not fit noise that lives at high source-coordinate wave numbers. In this way, the equation at the current model acts as a regularization term that does prevent fitting the noise at the low frequencies.

5.4 Application to the 2D Chevron 2014 blind-test dataset

While wave-equation-based inversion techniques have received a lot of attention lately, their performance on realistic blind synthetics exposes a number of challenges. The best results so far have been obtained by experienced practitioners from industry and academia who developed labor intensive workflows to obtain satisfactory results from noisy synthetic data sets that are representative of complex geological areas. Our aim is to create versatile and easy to use workflows that reduce user input by using projections on convex sets.

Specifically, we will demonstrate the benefits from a combination of the convex bound and minimal smoothness constraints that require minimal parameter selection. Moreover, we will investigate the effectiveness of our on-the-fly source estimation technique for WRI when applying to more realistic data. We invert the 2D Chevron 2014 marine isotropic elastic synthetic blind-test dataset with two different settings: with and without the minimum smoothness constraints. The dataset contains 1600 shots located from 1 km to 40 km. The maximum offset is 8 km. Both source and receiver depths are 15 m. We use 16 frequency bands ranging from 3 Hz to 19 Hz with overlap. Each frequency band contains three different frequencies. For each frequency band, we perform 5 Projected Newton-type iterations. For each iteration, we randomly jittered selected 360 sources from the 1600 sources to calculate the misfit, gradient, and Hessian approximation. The lower and upper bound of velocity are 1510 m/s and 4500 m/s, respectively.

Figures 5.2, 5.3a, and 5.3b correspond to the initial model, inversion result without the minimum smoothness constraint, and inversion result with the minimum smoothness constraint, respectively. At the depth of 1.5 km, both inversion results have the gas clouds at $x = 10$ km, 20 km, and 24 km. Both results also obtain the low-velocity layer from the depth of 2 km to 3 km. This low velocity layer is difficult for conventional FWI to reconstruct from the initial model, because the turning waves that conventional FWI uses to reconstruct the long wavelength structures rarely reach the depth below 2 km for a maximal offset of 8 km (Virieux and Operto, 2009). Comparing the two results, the inversion without the smoothness constraint has many vertical artifacts due to the data noise and subsampled sources, while with the smoothness constraint, there are fewer vertical artifacts in Figure 5.3b. Figure 5.4 depicts a comparison of the well-log velocity (red), initial velocity (blue), and inverted velocities with (black) and without (green) the smoothness constraints. Clearly, the velocity inverted with

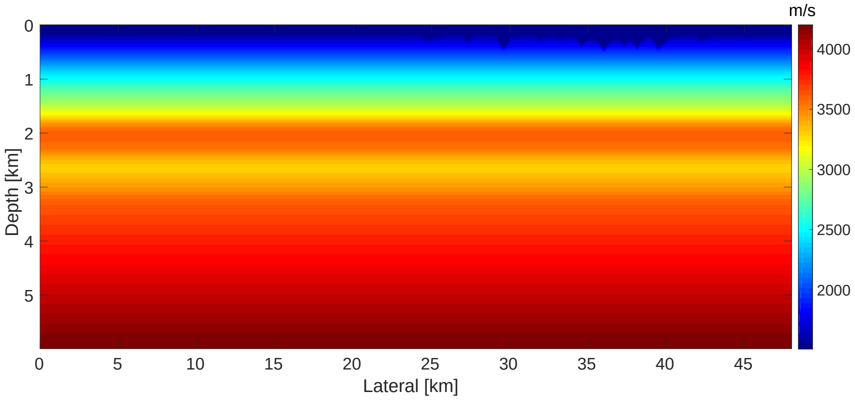


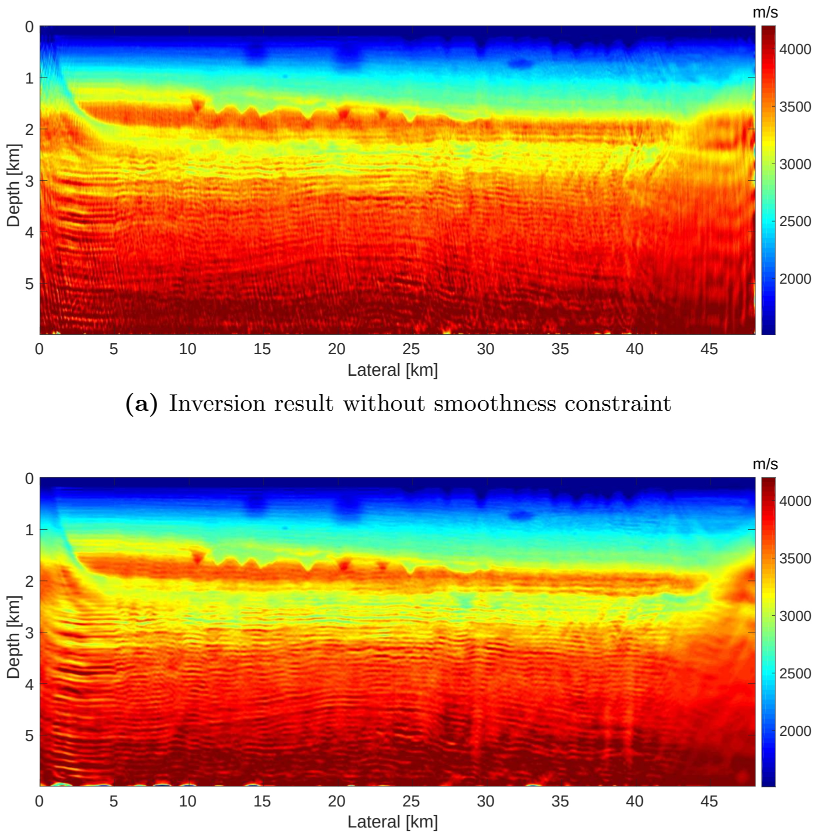
Figure 5.2: Initial model

the smoothness constraint matches with the well-log velocity better than that inverted without the constraint. Both results demonstrate that the smoothness constraint is able to remove artifacts and improve the quality of the inversion result.

Figure 5.5 shows the comparison between the estimated source wavelet and the true source wavelet provided by Chevron. Aside from an obvious amplitude scaling, both the frequency-dependent phase and amplitude characteristics are well recovered from this complex elastic data set. This is remarkable because our modeling engine is the scalar constant-density wave equation and this result is a clear illustration of the benefits of having a formulation where the physics and data are fitted jointly during the inversion. Considering both the difference between the inversion results and the true model as well as errors from the discretization, these slight differences between the true and estimated wavelet are acceptable.

5.5 Discussions

Conceptually, the approach presented is related to the work of Baumstein (2013). There too, projections onto convex sets are used to regularize the waveform inversion problem. The optimization, computation of projections, as well as the sets are different. However, in that approach convergence may not be guaranteed because the projections are not guaranteed to be onto the intersections and because they only project at the end and do not solve constrained sub-problems. This is a problem when incorporating second-order information.



(a) Inversion result without smoothness constraint

(b) Inversion result with smoothness constraint

Figure 5.3: Comparison of inversion result with and without smoothness constraint. (a) Inversion result without smoothness constraint. (b) Inversion result with smoothness constraint.

Tikhonov regularization is a well-known regularization technique and adds quadratic penalties to objectives such as

$$\underset{\mathbf{m}}{\text{minimize}}\, f(\mathbf{m}) + \frac{\beta_1}{2}\|\mathbf{R}_1\mathbf{m}\|_2^2 + \frac{\beta_2}{2}\|\mathbf{R}_2\mathbf{m}\|_2^2, \tag{5.14}$$

where two regularizers are used. The constants β_1 and β_2 are the scalar weights, determining the importance of each regularizer. The matrices $\mathbf{R}_1$ and $\mathbf{R}_2$ represent properties of the model, which we would like to penalize. This method should in principle be able to achieve similar results as the projection approach, but it has a few disadvantages. The first one is that

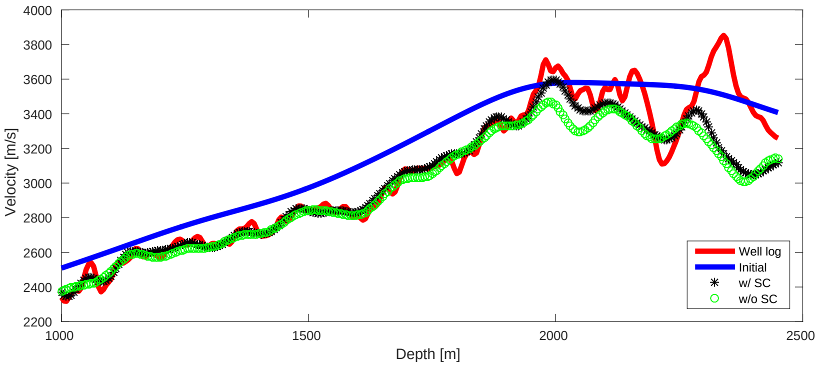


Figure 5.4: Comparison of the well-log velocity (red), initial velocity (blue), and inverted velocities with (black) and without (green) the smoothness constraint.

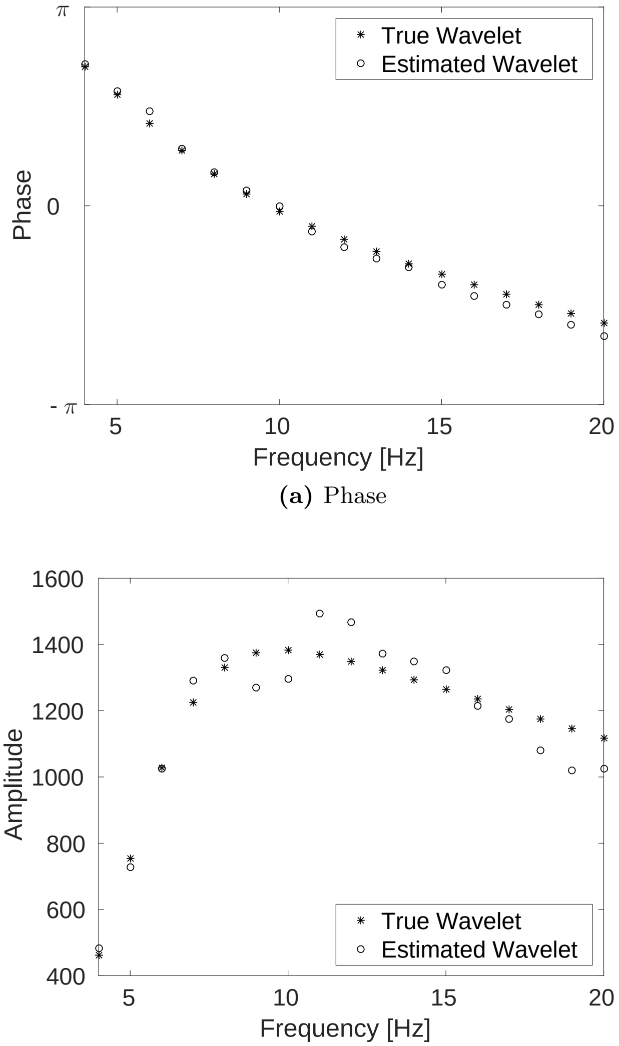
it may be difficult or costly to determine the weights β_1 and β_2. Multiple techniques for finding a "good" set of weights are available, including the L-curve, cross-validation, and the discrepancy principle, see Sen and Roy (2003) for an overview in a geophysical setting. A second issue with the penalty approach is the effect of the penalty terms on the condition number of the Hessian of the optimization problem, see Nocedal and Wright (2000). This means that certain choices of β_1, β_2, $\mathbf{R}_1$, and $\mathbf{R}_2$ may lead to an optimization problem that is unfeasibly expensive to solve. A third issue is related to the two-norm squared, $\| \cdot \|_2^2$ not being a so-called "exact" penalty function Nocedal and Wright (2000). This means that constraints can be approximated by the penalty functions, but generally not satisfied exactly. Note that the projection and quadratic penalty strategies for regularization can be easily combined as

$$\underset{\mathbf{m}}{\text{minimize}}\, f(\mathbf{m}) + \frac{\beta_1}{2} \|\mathbf{R}_1 \mathbf{m}\|_2^2 + \frac{\beta_2}{2} \|\mathbf{R}_2 \mathbf{m}\|_2^2, \text{ s.t. } \mathbf{m} \in \mathcal{C}_1 \bigcap \mathcal{C}_2. \tag{5.15}$$

Multiple researchers, for example Brenders and Pratt (2007), use the concept of filtering the gradient. A gradient filter can be represented as applying a linear operation to the gradient step as

$$\mathbf{m}_{k+1} = \mathbf{m}_k - \gamma \tilde{\mathbf{F}} \nabla_{\mathbf{m}} f(\mathbf{m}), \tag{5.16}$$

where the operator $\tilde{\mathbf{F}}$ filters the gradient. Brenders and Pratt (2007) apply a low-pass filter to the gradient to prevent high-wavenumber updates to the



(a) Phase

(b) Amplitude

Figure 5.5: Phase (a) and amplitude (b) comparisons between the final estimated source wavelet (○) and the true source wavelet (∗)

model while inverting low-frequency seismic data. A similar effect is achieved by the proposed framework and using the set defined in equation 5.7. The projection-based approach has the advantage that it generalizes to multiple constraints on the model in multiple domains, whereas gradient filtering does not. Restricting the model to be in a certain space is quite similar to reparametrization of the model.

5.6 Conclusions

To arrive at a flexible and versatile constrained algorithm for WRI-SE, we combine Dykstra's algorithm with a projected-Newton type algorithm capable of merging several regularization strategies including Tikhonov regularization, gradient filtering, and reparameterization. While our approach could in certain cases yield equivalent results yielded by standard types of regularization, there are no guarantees for the latter. In addition, the parameter selections become easier because the parameters are directly related to the properties of the model itself rather than to balancing data fits and regularization. Application of WRI-SE to the blind-test Chevron 2014 dataset show excellent and competitive results that benefit from imposing the smoothing constraint. While there is still a lot of room for improvement, these results are encouraging and were obtained without data preprocessing, excessive handholding, and cosmetic "tricks". The fact that we recover the wavelet accurately using an acoustic constant density only modeling kernel also shows the relative robustness of the presented method.

Chapter 6

Uncertainty quantification for inverse problems with weak PDE constraints and its application to seismic wave-equation-based inversions

6.1 Introduction

Inverse problems with partial-differential-equation (PDE) constraints arise in many applications of geophysics (Tarantola and Valette, 1982a; Pratt, 1999; Haber et al., 2000; Epanomeritakis et al., 2008; Virieux and Operto, 2009). The goal of these problems is to infer the values of unknown spatial distributions of physical parameters (e.g., sound speed, density, or electrical conductivity) from indirectly measured data, where the underlying physical model is described by a PDE (e.g., the Helmholtz equation or Maxwell's equations). The most challenging aspects of these problems arise from the fact that they are typically multimodal, with many spurious local minima (Biegler et al., 2012), which can inhibit gradient-based optimization algorithms from reaching the global optimal solution successfully.

A version of this chapter has been submitted.

This multimodality stems in part from the fact that the observed data are measured on a small subset of the entire boundary of the domain (Bui-Thanh et al., 2013) and the nonlinear parameter-to-data forward map (van Leeuwen and Herrmann, 2013b, 2015). One approach to dealing with the multimodality is to formulate the inverse problem as a deterministic optimization problem that aims at minimizing the misfit between the predicted and observed data in an appropriate norm, while also adding a regularization term that may eliminate the nonconvexity in certain situations (Virieux and Operto, 2009; Martin et al., 2012). The result of this deterministic approach is an estimate of the model parameters that is consistent with the observed data and contains few unwanted features. Since observed data typically contain measurement noise and modeling errors, we are not only interested in an estimate that best fits the data, but also in a complete statistical description of the unknown model parameters (Tarantola and Valette, 1982b; Osypov et al., 2013b). To that end, statistical approaches, and in particular the Bayesian inference method, are desirable and necessary. Unlike in the deterministic case, the solution produced by Bayesian inference is a posterior probability density function (PDF), which incorporates uncertainties in the observed data, the forward modeling map, and one's prior knowledge of the parameters. Once we can tractably compute the posterior distribution, we can extract various statistical properties of the unknown parameters.

Bayesian inference methods have been applied to a number of PDE-constrained geophysical statistical inverse problems (Tarantola, 1987, 2005; Aster et al., 2011; Martin et al., 2012; Bui-Thanh et al., 2013; Zhu et al., 2016; Ely et al., 2017). In these reported works, the PDE is typically treated as a strict constraint when formulating the posterior PDF, i.e., the field variables should always exactly satisfy the PDE. This leads to the so-called reduced or adjoint-state method (Plessix, 2006; Hinze et al., 2008) that eliminates the field variables by solving the PDE, resulting in a posterior PDF with multiple modes. To study the posterior PDF, Markov chain Monte Carlo (McMC) type methods, including the Metropolis-Hasting based methods (Haario et al., 2006; Stuart et al., 2016; Ely et al., 2017), the stochastic Newton-type method (Martin et al., 2012), and the randomize-then-optimize (RTO) method (Bardsley et al., 2015) sample the posterior PDF by drawing samples from a proposal distribution followed by an accept or reject step. To compute the accept/reject ratio, these methods have to evaluate the posterior PDF for each sample, which leads to solving a large number of computationally expensive PDEs. Moreover, according to the scaling analysis by Roberts et al. (2001), McMC type methods require a significantly larger number of samples to reach a status of convergence for large-scale problems in comparison with small-scale problems, which is well known as the curse of dimensionality. These difficulties preclude the straightforward applications of these methods to large-scale problems with more than 10^6 unknown parameters.

In this chapter, we present a new formulation of the posterior distribution that subsumes the conventional reduced formulation as a special case. Instead of treating the PDE as a strict constraint and eliminating the field variables by solving the PDE, we relax the PDE-constraint by introducing the field variables as auxiliary variables. This idea is similar to the method of van Leeuwen and Herrmann (2015) applied to deterministic PDE-constrained optimization problems, in which the PDE misfit is treated as a penalty term in the misfit function and weighted by a penalty parameter. Moreover, the idea of relaxing the PDE-constraint is also widely-used in weather forecasting applications (Fisher et al., 2005) for the sake of improving the stability of results. In the field of seismic exploration, Fang et al. (2015) and Fang et al. (2016) first introduced a method to construct a posterior PDF with weak acoustic wave-equation constraints. In the following study by Lim (2017), the authors showed that for small-scale problems, this posterior PDF can be sampled by the randomized maximum likelihood-McMC (RML-McMC) method. We demonstrate that the conventional reduced posterior PDF is a special case of our new formulation. By exploiting the structure of the new posterior PDF, we show that, with an appropriate penalty parameter, the new posterior PDF can be approximated by a Gaussian PDF, which is centered at the maximum a posteriori (MAP) estimate that maximizes the posterior PDF. To construct this Gaussian approximation, we exploit the local derivative information of the posterior PDF and formulate the covariance matrix as a PDE-free operator, which allows us to compute the matrix-vector product without the requirement of computing a large number of additional PDEs. By avoiding an explicit formulation of the covariance matrix, which would be impractical to compute and store, we can apply a recently proposed bootstrap type method (Efron, 1981, 1992)—the so-called randomize-then-optimize method (Bardsley et al., 2014)—to affordably draw samples from this surrogate distribution.

We apply our new computational framework to several seismic wave-equation-based inverse problems ranging in size and complexity of the underlying parameters. Our first example compares our sampling method with a benchmark method—the randomize maximum likelihood (RML) method (Chen and Oliver, 2012)—to validate our Gaussian approximation on a simple model with 1800 unknown parameters. Next, we apply our computational framework to a more complex model with 92,455 unknown parameters to test the feasibility of the approach to more realistically sized problems.

This chapter is organized into three major sections. The first introduces the derivation of the posterior PDF and the corresponding sampling method in a general setting. The second section introduces each component in the general framework when applied to the full-waveform inversion type problems. The final section presents the results of the application of our framework to several numerical inverse problems for velocity models with different size and complexity.

6.2 Bayesian framework for inverse problems with a weak PDE-constraint

In a PDE-constrained inverse problem, the goal is to infer the unknown discretized n_{grid}-dimensional physical model parameters $\mathbf{m} \in \mathbb{R}^{n_{\text{grid}}}$ from n_{data}-dimensional noisy observed data $\mathbf{d} \in \mathbb{C}^{n_{\text{data}}}$. As the noisy data are stochastic in nature, so are the inversion results obtained from them. Bayesian inference is a widely-used approach that seeks to estimate the posterior PDF of the unknown parameters $\mathbf{m}$ by incorporating the statistics of the measurement and modeling error and one's prior knowledge of the underlying model. Mathematically, Bayesian inference applies Bayes' law to formulate the posterior PDF $\rho_{\text{post}}(\mathbf{m}|\mathbf{d})$ of the model parameters $\mathbf{m}$ given the observed data $\mathbf{d}$ by combining a likelihood PDF and a prior PDF as

$$\rho_{\text{post}}(\mathbf{m}|\mathbf{d}) \propto \rho_{\text{like}}(\mathbf{d}|\mathbf{m})\rho_{\text{prior}}(\mathbf{m}), \tag{6.1}$$

where the likelihood PDF $\rho_{\text{like}}(\mathbf{d}|\mathbf{m})$ describes the probability of observing the data $\mathbf{d}$ given the model parameters $\mathbf{m}$, and the prior PDF $\rho_{\text{prior}}(\mathbf{m})$ describes one's prior beliefs in the unknown model parameters. The proportionality constant depends on the observed data $\mathbf{d}$, which are fixed. Once we have a computationally tractable approach to estimating the posterior PDF, we can apply certain sampling methods to draw samples from the posterior PDF, which can then be used to compute statistical properties of interest such as the MAP estimate

$$\mathbf{m}_{*} = \arg\max_{\mathbf{m}} \rho_{\text{post}}(\mathbf{m}|\mathbf{d}), \tag{6.2}$$

the mean value

$$E[\mathbf{m}|\mathbf{d}] = \int \mathbf{m}\rho_{\text{post}}(\mathbf{m}|\mathbf{d})\mathrm{d}\mathbf{m}, \tag{6.3}$$

the model covariance matrix

$$\Gamma(m_k, m_l|\mathbf{d}) = \int m_k m_l \rho_{\text{post}}(\mathbf{m}|\mathbf{d}) d\mathbf{m} - E[m_k|\mathbf{d}]E[m_l|\mathbf{d}], \tag{6.4}$$

the model standard deviation (STD)

$$\text{STD}(m_k) = \sqrt{\Gamma(m_k, m_k|\mathbf{d})}, \tag{6.5}$$

and the marginal distributions of $\mathbf{m}$

$$\rho_{\text{M}}(m_k|\mathbf{d}) = \int \cdots \int \rho_{\text{post}}(\mathbf{m}|\mathbf{d}) \prod_{l=1, l\neq k}^{n_{\text{grid}}} dm_l, \tag{6.6}$$

where the values m_k and m_l denote the k^{th} and l^{th} elements of the vector $\mathbf{m}$ (Kaipio and Somersalo, 2006; Matheron, 2012). If samples from the posterior PDF are available, then the quantities listed above can be easily computed through ensemble averages. The primary issue for statisticians is to construct the posterior PDF and design methods that can efficiently draw samples from it, which are used in the approximation of the integrals above.

6.2.1 The posterior PDF

To motivate the derivation of the new posterior PDF, it is helpful to start with the conventional formulation of the posterior PDF for PDE-constrained inverse problems. The so-called reduced approach eliminates the PDE-constraint by solving the PDE, which leads to the following nonlinear forward modeling map $F(\mathbf{m})$:

$$F(\mathbf{m}) = \mathbf{P}\mathbf{A}(\mathbf{m})^{-1}\mathbf{q}. \tag{6.7}$$

Here the vector $\mathbf{q} \in \mathbb{C}^{n_{\text{grid}}}$ represents the discretized (known) source term. The matrix $\mathbf{A}(\mathbf{m}) \in \mathbb{C}^{n_{\text{grid}} \times n_{\text{grid}}}$ denotes the discretized PDE operator and the operator $\mathbf{P} \in \mathbb{R}^{n_{\text{data}} \times n_{\text{grid}}}$ samples the data $\mathbf{d}$ from the vector of field variables $\mathbf{u}$, which is the solution of the PDE $\mathbf{u} = \mathbf{A}(\mathbf{m})^{-1}\mathbf{q}$. In real-world seismic applications, the observed data always contain correlated measurement noise arising from environmental disturbances and modeling errors, which are difficult to precisely quantify and model. One popular approach to simplifying the problem is to assume that the combined measurement and modeling noise $\epsilon \in \mathbb{C}^{n_{\text{data}}}$ is additive and is drawn from a Gaussian distribution with zero mean and covariance matrix Γ_{noise} (Bui-Thanh et al., 2013; Osypov et al., 2013b; Bardsley et al., 2015), i.e., $\epsilon \sim \mathcal{N}(0, \Gamma_{\text{noise}})$, independent of $\mathbf{m}$. The assumption of Gaussianity results in distributions that are relatively easy to model and sample, thereby providing a rich source of tractable examples (Kaipio and Somersalo, 2006). With this assumption in mind and the additional assumption that the prior distribution of the model $\mathbf{m}$ is also Gaussian with the mean model parameters $\tilde{\mathbf{m}}$ and the covariance matrix Γ_{prior}, we arrive at the following posterior distribution:

$$\rho_{\text{post}}(\mathbf{m}|\mathbf{d}) \propto \exp\left(-\frac{1}{2}\|F(\mathbf{m}) - \mathbf{d}\|^2_{\Gamma^{-1}_{\text{noise}}} - \frac{1}{2}\|\mathbf{m} - \tilde{\mathbf{m}}\|^2_{\Gamma^{-1}_{\text{prior}}}\right), \tag{6.8}$$

where the symbol $\|\boldsymbol{.}\|_{\Gamma^{-1}_{\text{noise}}}$ denotes the weighted ℓ_2-norm with the weighting matrix $\Gamma^{-1}_{\text{noise}}$. There are several challenges in computing various quantities associated with the posterior distribution in equation 6.8. In order to obtain the MAP estimate $\mathbf{m}_*$, we need to solve the following deterministic optimization problem:

$$\begin{aligned}
\mathbf{m}_* &= \arg\max_{\mathbf{m}} \rho_{\text{post}}(\mathbf{m}|\mathbf{d}) = \arg\min_{\mathbf{m}} -\log \rho_{\text{post}}(\mathbf{m}|\mathbf{d}) \\
&= \arg\min_{\mathbf{m}} \frac{1}{2}\|F(\mathbf{m}) - \mathbf{d}\|^2_{\Gamma^{-1}_{\text{noise}}} + \frac{1}{2}\|\mathbf{m} - \tilde{\mathbf{m}}\|^2_{\Gamma^{-1}_{\text{prior}}},
\end{aligned} \tag{6.9}$$

which can be solved by the so-called adjoint-state method. As noted in van Leeuwen and Herrmann (2013b), the nonlinear forward modeling map $F(\mathbf{m})$ results in the objective function $-\log \rho_{\text{post}}(\mathbf{m}|\mathbf{d})$ being highly oscillatory with respect to the model parameters $\mathbf{m}$, which yields many local minima. To find the globally optimal solution, a sufficiently close initial model is necessary, which may be difficult to obtain in real-world scenarios. As mentioned previously, the nonlinear parameter-to-data map also results in computational difficulties when sampling the posterior distribution in equation 6.8. Specifically, the tradeoff between designing a proposal distribution that is well tuned to the true posterior distribution and one that is computationally cheap to sample is not a straightforward choice to make. As one models a proposal that is easier to sample, typically the price to pay is having to draw more samples until convergence is reached.

These challenges result from the nonlinear forward modeling map $F(\mathbf{m})$ induced by the strict PDE-constraint in the optimization problem in equation 6.9. To overcome these difficulties, van Leeuwen and Herrmann (2013b) and van Leeuwen and Herrmann (2015) proposed a penalty formulation to solve deterministic PDE-constrained optimization problems, wherein they relax the strict PDE-constraint by penalizing the data misfit function by a weighted PDE misfit with a penalty parameter λ. This results in the following joint optimization problem with respect to both the model parameters $\mathbf{m}$ and the field variables collected in the vector $\mathbf{u}$:

$$\underset{\mathbf{m},\mathbf{u}}{\arg\min}\, f_{\text{pen}}(\mathbf{m},\mathbf{u}) = \frac{1}{2}\|\mathbf{P}\mathbf{u}-\mathbf{d}\|^2_{\Gamma^{-1}_{\text{noise}}} + \frac{\lambda^2}{2}\|\mathbf{A}(\mathbf{m})\mathbf{u}-\mathbf{q}\|^2 + \frac{1}{2}\|\mathbf{m}-\tilde{\mathbf{m}}\|^2_{\Gamma^{-1}_{\text{prior}}}. \tag{6.10}$$

The authors note that the problem in equation 6.10 is a separable nonlinear least-squares problem, in which the optimization with respect to $\mathbf{u}$ is a linear data-fitting problem when $\mathbf{m}$ is fixed. In van Leeuwen and Herrmann (2015), the authors eliminate the field variables $\mathbf{u}$ by the variable projection method (Golub and Pereyra, 2003) in order to avoid the high memory costs involved in storing a unique field variable for each individual source. The variable projection method also eliminates the dependence of the objective function in equation 6.10 on $\mathbf{u}$. As for each input parameter $\mathbf{m}$, there is a unique $\overline{\mathbf{u}}(\mathbf{m})$ satisfying $\nabla_{\mathbf{u}} f_{\text{pen}}(\mathbf{m},\mathbf{u})|_{\mathbf{u}=\overline{\mathbf{u}}(\mathbf{m})} = 0$, which has the closed form solution

$$\overline{\mathbf{u}}(\mathbf{m}) = \left(\lambda^2\mathbf{A}(\mathbf{m})^\top\mathbf{A}(\mathbf{m}) + \mathbf{P}^\top\Gamma^{-1}_{\text{noise}}\mathbf{P}\right)^{-1}\left(\lambda^2\mathbf{A}(\mathbf{m})^\top\mathbf{q} + \mathbf{P}^\top\Gamma^{-1}_{\text{noise}}\mathbf{d}\right), \tag{6.11}$$

where the symbol $\top$ denotes the (complex-conjugate) transpose. As noted by van Leeuwen and Herrmann (2013b) and van Leeuwen and Herrmann (2015), minimizing the objective function (equation 6.10) with a carefully selected λ is less prone to being trapped in suboptimal local minima, because the inversion is carried out over a larger search space (implicitly through $\overline{\mathbf{u}}(\mathbf{m})$) and it is therefore easier to fit the observed data for poor starting models compared to the conventional reduced formulation in equation 6.9.

Motivated by the penalty approach to solving the deterministic inverse problems, we propose a more generic posterior PDF for statistical PDE-constrained inverse problems. As before, we relax the PDE-constraint by introducing the field variables $\mathbf{u}$ as auxiliary variables, i.e., we have

$$\rho_{\text{post}}(\mathbf{u},\mathbf{m}|\mathbf{d}) \propto \rho(\mathbf{d}|\mathbf{u},\mathbf{m})\rho(\mathbf{u},\mathbf{m}), \tag{6.12}$$

where the conditional PDF $\rho(\mathbf{d}|\mathbf{u},\mathbf{m})$ now describes the probability of observing the data $\mathbf{d}$ given the field variables $\mathbf{u}$ and model parameters $\mathbf{m}$. To formulate the joint PDF $\rho(\mathbf{u},\mathbf{m})$, we apply the standard conditional decomposition (Sambridge et al., 2006)

$$\rho(\mathbf{u},\mathbf{m}) = \rho(\mathbf{u}|\mathbf{m})\rho_{\text{prior}}(\mathbf{m}). \tag{6.13}$$

Hence,

$$\rho_{\text{post}}(\mathbf{u},\mathbf{m}|\mathbf{d}) \propto \rho(\mathbf{d}|\mathbf{u},\mathbf{m})\rho(\mathbf{u}|\mathbf{m})\rho_{\text{prior}}(\mathbf{m}). \tag{6.14}$$

The implication of the reduced formulation is that the field variables $\mathbf{u}$ satisfy the PDE strictly—i.e., $\mathbf{u} = \mathbf{A}(\mathbf{m})^{-1}\mathbf{q}$. This adherence to the PDE corresponds to solutions $\mathbf{u}$ that satisfy the PDE $\mathbf{A}(\mathbf{m})\mathbf{u} = \mathbf{q}$ with the probability density $\rho(\mathbf{u}|\mathbf{m}) = \delta\big(\mathbf{A}(\mathbf{m})\mathbf{u} - \mathbf{q}\big)$, where $\delta()$ denotes the Dirac delta function (Dummit and Foote, 2004). Conversely, replacing the constraint by a quadratic penalty term allows for a Gaussian error with zero mean and covariance matrix $\lambda^{-2}\mathbf{I}$ in the PDE misfit $\mathbf{A}(\mathbf{m})\mathbf{u} - \mathbf{q}$. This yields

$$\rho(\mathbf{u}|\mathbf{m}) = (2\pi)^{-\frac{n_{\text{grid}}}{2}} \det\left(\lambda^2\mathbf{A}(\mathbf{m})^\top\mathbf{A}(\mathbf{m})\right)^{\frac{1}{2}} \exp\left(-\frac{\lambda^2}{2}\|\mathbf{A}(\mathbf{m})\mathbf{u} - \mathbf{q}\|^2\right). \tag{6.15}$$

Indeed, for a given $\mathbf{m}$, this distribution $\rho(\mathbf{u}|\mathbf{m})$ with respect to $\mathbf{u}$ is a Gaussian distribution with a mean of $\mathbf{A}(\mathbf{m})^{-1}\mathbf{q}$ and a covariance matrix of $\lambda^{-2}\mathbf{A}(\mathbf{m})^{-1}\mathbf{A}(\mathbf{m})^{-\top}$. The conditional probability $\rho_{\text{post}}(\mathbf{u},\mathbf{m}|\mathbf{d})$ is the joint PDF with respect to both the model parameters $\mathbf{m}$ and field variables $\mathbf{u}$. Because the control or model variables $\mathbf{m}$ are of primary interest, we eliminate the dependence of the joint PDF on the auxiliary variables $\mathbf{u}$ by marginalizing over $\mathbf{u}$:

$$\begin{aligned}
\rho_{\text{post}}(\mathbf{m}|\mathbf{d}) &= \int \rho_{\text{post}}(\mathbf{u},\mathbf{m}|\mathbf{d})\mathrm{d}\mathbf{u} \\
&\propto \int \rho(\mathbf{d}|\mathbf{u},\mathbf{m})\rho(\mathbf{u}|\mathbf{m})\rho_{\text{prior}}(\mathbf{m})\mathrm{d}\mathbf{u} \\
&= \int \rho(\mathbf{d}|\mathbf{u},\mathbf{m})\rho(\mathbf{u}|\mathbf{m})\mathrm{d}\mathbf{u} \times \rho_{\text{prior}}(\mathbf{m}) \\
&\propto \det\big(\mathbf{H}(\mathbf{m})\big)^{-\frac{1}{2}} \det\left(\lambda^2\mathbf{A}(\mathbf{m})^\top\mathbf{A}(\mathbf{m})\right)^{\frac{1}{2}} \\
&\quad \times \exp\Big(-\frac{1}{2}\big(\lambda^2\|\mathbf{A}(\mathbf{m})\overline{\mathbf{u}}(\mathbf{m}) - \mathbf{q}\|^2 + \|\mathbf{P}\overline{\mathbf{u}}(\mathbf{m}) - \mathbf{d}\|^2_{\Gamma^{-1}_{\text{noise}}} \\
&\qquad\qquad + \|\mathbf{m} - \tilde{\mathbf{m}}\|^2_{\Gamma^{-1}_{\text{prior}}}\big)\Big),
\end{aligned} \tag{6.16}$$

where

$$\begin{aligned}
\mathbf{H}(\mathbf{m}) &= -\nabla^2_{\mathbf{u}} \log \rho_{\text{post}}(\mathbf{u},\mathbf{m}|\mathbf{d})\big|_{\mathbf{u}=\overline{\mathbf{u}}(\mathbf{m})} \\
&= \lambda^2\mathbf{A}(\mathbf{m})^\top\mathbf{A}(\mathbf{m}) + \mathbf{P}^\top\Gamma^{-1}_{\text{noise}}\mathbf{P},
\end{aligned} \tag{6.17}$$

and $\overline{\mathbf{u}}(\mathbf{m})$ is given by equation 6.11. As in the deterministic case, the posterior PDF corresponding to the conventional reduced formulation (cf. equation 6.8) can also be derived from the marginal PDF $\rho_{\text{post}}(\mathbf{m}|\mathbf{d})$ in equation 6.16. Indeed, as $\lambda \to \infty$, we have

$$\begin{aligned}
&\lim_{\lambda\to\infty} \det\big(\mathbf{H}(\mathbf{m})\big)^{-\frac{1}{2}} \det\big(\lambda^2 \mathbf{A}(\mathbf{m})^\top \mathbf{A}(\mathbf{m})\big)^{\frac{1}{2}} \\
&= \lim_{\lambda\to\infty} \det\left(\mathbf{I} + \frac{1}{\lambda^2}\mathbf{A}(\mathbf{m})^{-\top}\mathbf{P}^\top \Gamma_{\text{noise}}^{-1}\mathbf{P}\mathbf{A}(\mathbf{m})^{-1}\right)^{-\frac{1}{2}} \\
&= \lim_{\lambda\to\infty} \det\left(\mathbf{I} + \frac{1}{\lambda^2}\Gamma_{\text{noise}}^{-\frac{1}{2}}\mathbf{P}\mathbf{A}(\mathbf{m})^{-1}\mathbf{A}(\mathbf{m})^{-\top}\mathbf{P}^\top \Gamma_{\text{noise}}^{-\frac{\top}{2}}\right)^{-\frac{1}{2}} \\
&= 1,
\end{aligned} \tag{6.18}$$

and

$$\lim_{\lambda\to\infty} \overline{\mathbf{u}}(\mathbf{m}) = \mathbf{A}(\mathbf{m})^{-1}\mathbf{q}. \tag{6.19}$$

This yields

$$\lim_{\lambda\to\infty} \rho_{\text{post}}(\mathbf{m}|\mathbf{d}) \propto \exp\big(-\frac{1}{2}\|\mathbf{P}\mathbf{A}(\mathbf{m})^{-1}\mathbf{q} - \mathbf{d}\|^2_{\Gamma_{\text{noise}}^{-1}} - \frac{1}{2}\|\mathbf{m} - \tilde{\mathbf{m}}\|^2_{\Gamma_{\text{prior}}^{-1}}\big), \tag{6.20}$$

which, as expected, is the posterior PDF corresponding to the conventional reduced formulation.

To illustrate the advantages of our penalty formulation for the posterior PDF (cf. equation 6.16) over the conventional reduced formulation (cf. equation 6.8), we conduct a simple experiment adopted from Esser et al. (2018). We invert for the sound speed given partial measurements of the pressure field and generate data with a known band-limited source. The data are recorded at four receivers (see Figure 6.1a) and is contaminated with Gaussian noise (see Figure 6.1b). As in Esser et al. (2018), we neglect the amplitude effects related to geometrical spreading and define the forward operator $F(m)$ as

$$F(m) = \mathbf{P}\mathbf{A}^{-1}(m)\mathbf{q} = \begin{pmatrix} \mathcal{F}^{-1}e^{i\omega m\|\mathbf{x}_1 - \mathbf{x}_s\|_2}\mathcal{F}\mathbf{q} \\ \mathcal{F}^{-1}e^{i\omega m\|\mathbf{x}_2 - \mathbf{x}_s\|_2}\mathcal{F}\mathbf{q} \\ \mathcal{F}^{-1}e^{i\omega m\|\mathbf{x}_3 - \mathbf{x}_s\|_2}\mathcal{F}\mathbf{q} \\ \mathcal{F}^{-1}e^{i\omega m\|\mathbf{x}_4 - \mathbf{x}_s\|_2}\mathcal{F}\mathbf{q} \end{pmatrix}, \tag{6.21}$$

where the operator $\mathcal{F}$ denotes the temporal Fourier transform and ω denotes the angular frequency. The vectors $\mathbf{x}_i$, $i = 1, ..., 4$ and $\mathbf{x}_s$ denote the receiver and source locations, respectively, and the scalar m represents the slowness of the medium, i.e., $m = \dfrac{1}{v}$ where v is the velocity. The source and receiver locations are in Figure 6.1a denoted by the symbols $(*)$ and

(∇). With the forward model defined in equation 6.21, we formulate the posterior distribution for the reduced formulation and the penalty formulation by choosing a Gaussian distribution with a mean of $5 \cdot 10^{-4}\,\mathrm{s/m}$ and a standard deviation of $1.2 \cdot 10^{-4}\,\mathrm{s/m}$ for the prior distribution $\rho_{\mathrm{prior}}(m)$. Finally, we add a Gaussian noise with a variance of 4×10^{-4} to the observed data, resulting in a noise-to-signal ratio of $\|\mathrm{noise}\|_2 / \|\mathrm{signal}\|_2 = 0.24$.

Figure 6.2 depicts the posterior PDFs of the reduced and penalty formulations for $\lambda =$ 10, 50, and 250. As expected, local maxima are present in the PDF of the reduced formulation and in the PDFs of the penalty formulation when the λ values become too large. As λ increases from $\lambda = 10$, where the posterior is unimodal, local maxima appear for larger $\lambda = 250$, only one of which has strong statistical significance. For $\lambda = 250$, the resulting PDF is close to the one yielded by the reduced formulation, which corresponds to $\lambda \to \infty$. From this stylized example, the strongly relaxed formulation appears unimodal and with low bias. As we will demonstrate in later sections, being less prone to local maxima reduces the computational cost associated with sampling these distributions.

6.2.2 Selection of λ

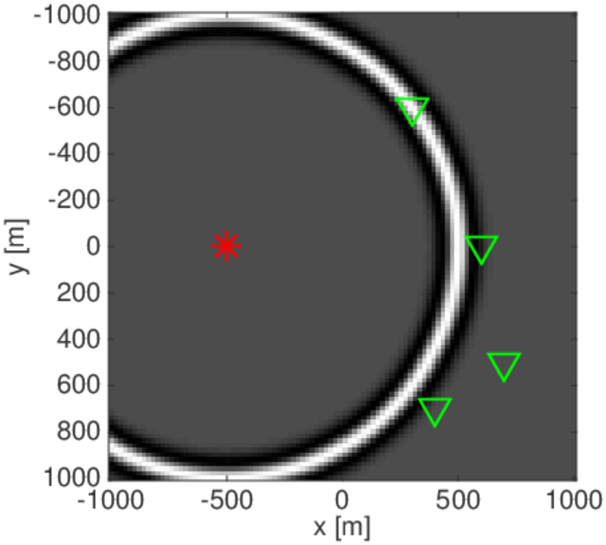
Before discussing computationally efficient sampling schemes, we first propose a method to select values for λ, which will balance the tradeoff between the unimodality of the distribution and its deviation from the reduced-formulation PDF.

To arrive at a scheme to select λ, we focus on two terms of the negative logarithm function of the posterior PDF in equation 6.16:

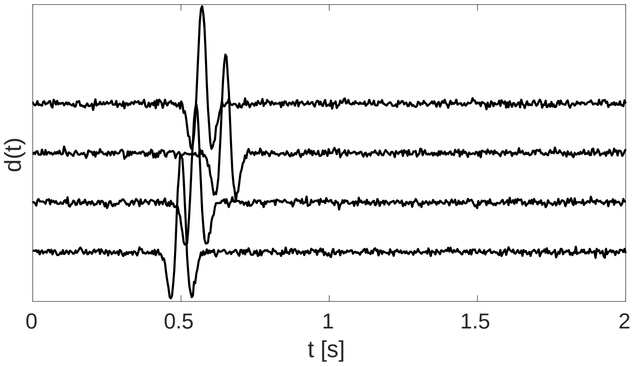
$$\begin{aligned}\phi(\mathbf{m}) =& -\log \rho_{\mathrm{post}}(\mathbf{m}|\mathbf{d}) \\ &= \frac{1}{2} \log \det \big(\mathbf{I} + \frac{1}{\lambda^2} \Gamma_{\mathrm{noise}}^{-\frac{1}{2}} \mathbf{PA}(\mathbf{m})^{-1} \mathbf{A}(\mathbf{m})^{-\top} \mathbf{P}^{\top} \Gamma_{\mathrm{noise}}^{-\frac{\top}{2}}\big) \\ &\quad + \frac{1}{2} \big(\lambda^2 \|\mathbf{A}(\mathbf{m})\bar{\mathbf{u}}(\mathbf{m}) - \mathbf{q}\|^2 + \|\mathbf{P}\bar{\mathbf{u}}(\mathbf{m}) - \mathbf{d}\|^2_{\Gamma_{\mathrm{noise}}^{-1}} + \|\mathbf{m} - \tilde{\mathbf{m}}\|^2_{\Gamma_{\mathrm{prior}}^{-1}}\big),\end{aligned} \tag{6.22}$$

namely the determinant

$$\phi_1(\mathbf{m}) = \frac{1}{2} \log \det \big(\mathbf{I} + \frac{1}{\lambda^2} \Gamma_{\mathrm{noise}}^{-\frac{1}{2}} \mathbf{PA}(\mathbf{m})^{-1} \mathbf{A}(\mathbf{m})^{-\top} \mathbf{P}^{\top} \Gamma_{\mathrm{noise}}^{-\frac{\top}{2}}\big), \tag{6.23}$$

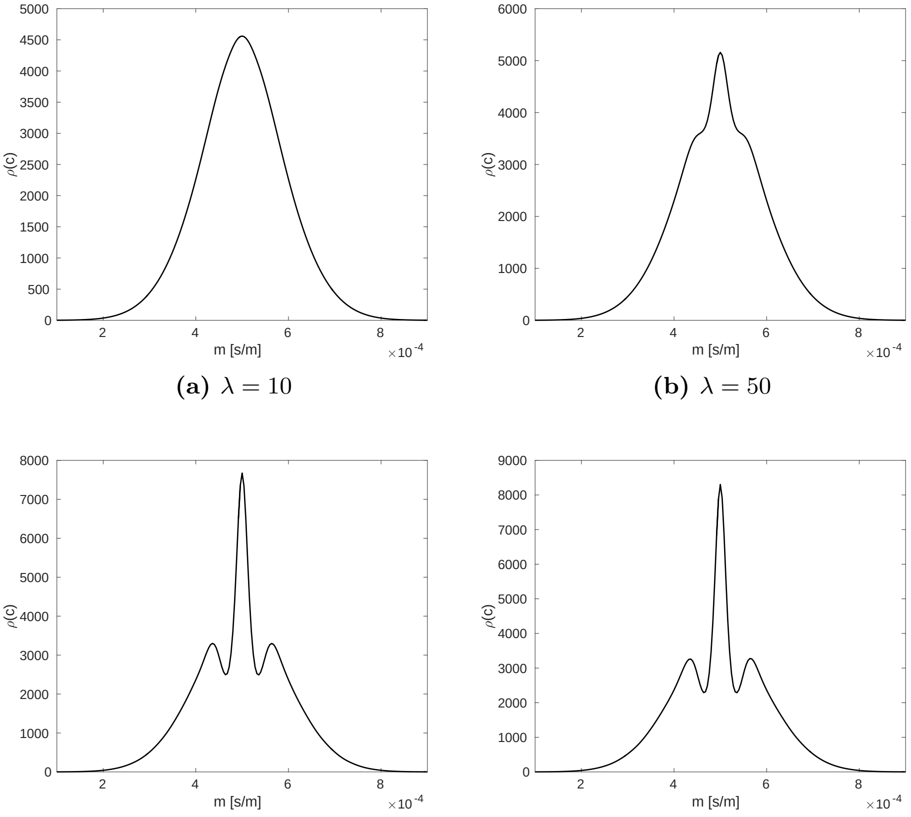


(a) Snap shot



(b) Data

Figure 6.1: (a) Snapshot of the time-domain wavefield generated by a single source ($*$); (b) recorded time-domain data at four receiver locations (∇).



(a) $\lambda = 10$

(b) $\lambda = 50$

(c) $\lambda = 250$

(d) Reduced

Figure 6.2: The four posterior PDFs corresponding to the penalty formulations with $\lambda = 10$ (a), 50 (b), and 250 (c), and the reduced formulation (d).

and the misfit

$$\phi_2(\mathbf{m}) = \frac{1}{2}\big(\lambda^2\|\mathbf{A}(\mathbf{m})\overline{\mathbf{u}}(\mathbf{m}) - \mathbf{q}\|^2 + \|\mathbf{P}\overline{\mathbf{u}}(\mathbf{m}) - \mathbf{d}\|^2_{\Gamma^{-1}_{\text{noise}}} + \|\mathbf{m} - \tilde{\mathbf{m}}\|^2_{\Gamma^{-1}_{\text{prior}}}\big). \tag{6.24}$$

These equations imply that we should avoid situations in which $\lambda \to 0$ and $\lambda \to \infty$. When $\lambda \to 0$, the optimal variable $\overline{\mathbf{u}}(\mathbf{m})$ tends to fit the data. As a result, $\phi_2(\mathbf{m}) \to \frac{1}{2}\|\mathbf{m}-\tilde{\mathbf{m}}\|^2_{\Gamma^{-1}_{\text{prior}}}$, which means that the observed data are not informative to the unknown parameters and have few contributions to the posterior distribution. Furthermore, as $\lambda \to 0$, the nonlinear determinant function $\phi_1(\mathbf{m}) \to \infty$ and will dominate the overall function $\phi(\mathbf{m})$, which results in a highly nonlinear mapping $\mathbf{m} \to \phi(\mathbf{m})$. On the other hand, when $\lambda \to \infty$, we find that $\phi_1(\mathbf{m}) \to 0$ and $\phi_2(\mathbf{m}) \to \frac{1}{2}\|\mathbf{PA}(\mathbf{m})^{-1}\mathbf{q} - \mathbf{d}\|^2_{\Gamma^{-1}_{\text{noise}}} + \frac{1}{2}\|\mathbf{m} - \tilde{\mathbf{m}}\|^2_{\Gamma^{-1}_{\text{prior}}}$ in which case the misfit $\phi(\mathbf{m})$ converges to the nonlinear reduced formulation. Considering both facts, we want to find an appropriate λ, so that $\phi_1(\mathbf{m})$ is relatively small compared to $\phi_2(\mathbf{m})$, thus ensuring enough information from the observed data, while $\phi_2(\mathbf{m})$ is still less likely to contain local minima.

Based on spectral arguments, van Leeuwen and Herrmann (2015) proposed a scaling for the penalty parameter according to the largest eigenvalue μ_1 of the matrix $\mathbf{A}(\mathbf{m})^{-\top}\mathbf{P}^{\top}\Gamma^{-1}_{\text{noise}}\mathbf{PA}(\mathbf{m})^{-1}$. Relative to μ_1, a penalty parameter $\lambda^2 \gg \mu_1$ can be considered large while $\lambda^2 \ll \mu_1$ is considered small. As a result, λ chosen much greater than this reference scale—i.e., when $\lambda^2 \gg \mu_1$, the minimizers for field variables $\overline{\mathbf{u}}(\mathbf{m})$ will converge to the solution of the wave equation $\mathbf{A}(\mathbf{m})^{-1}\mathbf{q}$ with a convergence rate of $\mathcal{O}(\lambda^{-2})$, and therefore our penalty misfit approaches the reduced misfit. A similar consideration applies when $\lambda^2 \ll \mu_1$. After extensive parameter testing, we found that choosing $\lambda^2 = 0.01\mu_1$ strikes the right balance so that the posterior PDF is less affected by local maxima compared to the reduced formulation while the data is still informative to the model parameters and the determinant term $\phi_1(\mathbf{m})$ remains negligible compared to $\phi_2(\mathbf{m})$. With this choice of λ, we can therefore neglect the $\phi_1(\mathbf{m})$ term, as it is small relative to $\phi_2(\mathbf{m})$, and consider an approximate posterior PDF $\overline{\rho}_{\text{post}}(\mathbf{m}|\mathbf{d})$ consisting only of the $\phi_2(\mathbf{m})$ term, i.e., we now have the approximate equality

$$\begin{aligned}\rho_{\text{post}}(\mathbf{m}|\mathbf{d}) &\approx \overline{\rho}_{\text{post}}(\mathbf{m}|\mathbf{d}) \\ &\propto \exp\Big(-\frac{\lambda^2}{2}\|\mathbf{A}(\mathbf{m})\overline{\mathbf{u}}(\mathbf{m}) - \mathbf{q}\|^2 - \frac{1}{2}\|\mathbf{P}\overline{\mathbf{u}}(\mathbf{m}) - \mathbf{d}\|^2_{\Gamma^{-1}_{\text{noise}}} \\ &\qquad - \frac{1}{2}\|\mathbf{m} - \tilde{\mathbf{m}}\|^2_{\Gamma^{-1}_{\text{prior}}}\Big).\end{aligned} \tag{6.25}$$

This approximation results in a posterior PDF that is much easier to evaluate. From here on out, we consider $\overline{\rho}_{\text{post}}(\mathbf{m}|\mathbf{d})$ as our PDF of interest in the subsequent sampling considerations. In practice, we may face situations that the inversion involves data corresponding to different frequencies. In such situations, both the noise levels and the PDEs can differ from frequency to frequency. As a result, the proposed selection criterion implies us to select different λ's for different frequencies.

6.2.3 Sampling method

Given this choice of λ, yielding the PDF in equation 6.25, the computational considerations of drawing samples from this approximate distribution are paramount in designing a tractable method. McMC-type methods are unfortunately computationally unfeasible for high-dimensional problems, owing to the relatively large number of the expensive evaluations of posterior distributions needed to converge adequately. For this reason, we follow an alternative approach—widely used in Bayesian inverse problems with PDE-constraints (Bui-Thanh et al., 2013; Zhu et al., 2016)—where we approximate the target posterior PDF by a Gaussian PDF. To construct this Gaussian PDF, we first find the MAP estimate $\mathbf{m}_*$ by solving

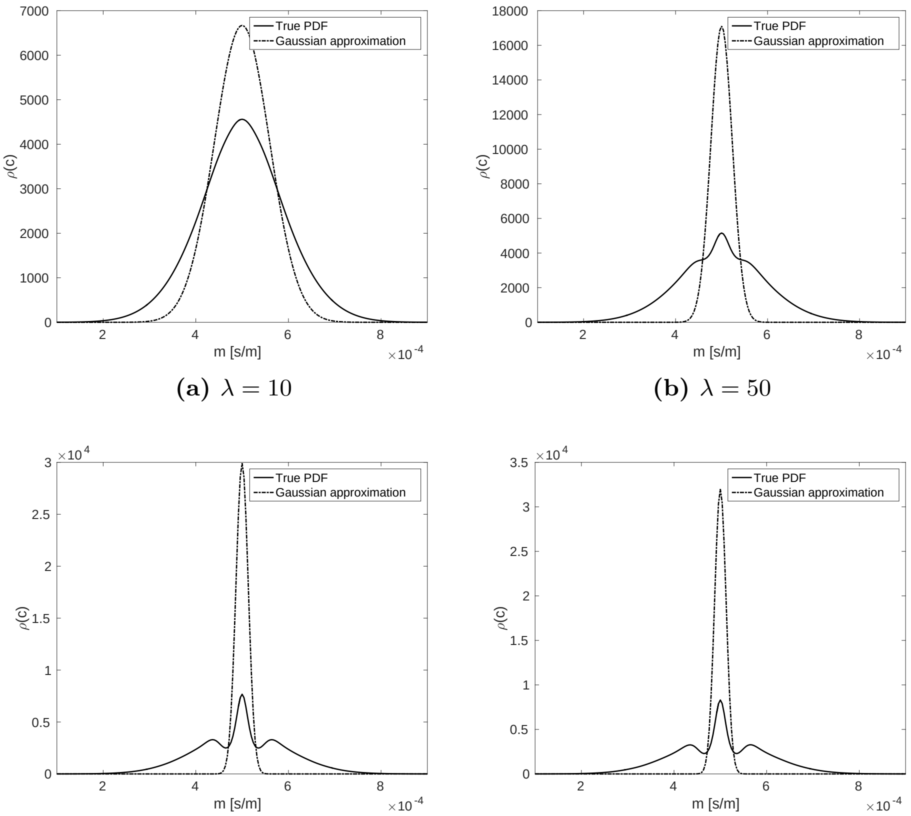
$$\mathbf{m}_* = \underset{\mathbf{m}}{\arg\min} - \log\big(\overline{\rho}_{\text{post}}\left(\mathbf{m}|\mathbf{d}\right)\big). \tag{6.26}$$

Next, we use the local second-order derivative information of the posterior PDF at the MAP point—i.e., the Hessian $\mathbf{H}_{\text{post}} = -\nabla^2_{\mathbf{m}} \log \overline{\rho}_{\text{post}}(\mathbf{m}|\mathbf{d})$—to construct the covariance matrix of the Gaussian PDF, which yields the Gaussian distribution $\mathcal{N}(\mathbf{m}_*, \mathbf{H}^{-1}_{\text{post}})$. Afterwards, we draw samples from the Gaussian distribution from which we compute statistical quantities. We incorporate this sampling strategy with the proposed posterior PDF with weak PDE-constraints and obtain the following Bayesian framework as shown in Algorithm 7:

Algorithm 7 Bayesian framework for inverse problems with weak PDE-constraints

```
1. Set $\Gamma_{\text{noise}}$, $\Gamma_{\text{prior}}$, prior mean model $\tilde{\mathbf{m}}$, and a value for the penalty
   parameter $\lambda$;
2. Formulate the posterior PDF $\overline{\rho}_{\text{post}}(\mathbf{m}|\mathbf{d})$ with equation 6.25;
3. Find the MAP estimate $\mathbf{m}_*$ by minimizing $-\log \overline{\rho}_{\text{post}}(\mathbf{m}|\mathbf{d})$;
4. Compute the Hessian matrix $\mathbf{H}_{\text{post}} = -\nabla^2_{\mathbf{m}}\log\big(\overline{\rho}_{\text{post}}(\mathbf{m}|\mathbf{d})\big)$ at the MAP
   estimate $\mathbf{m}_*$ and define the Gaussian PDF $\mathcal{N}(\mathbf{m}_*, \mathbf{H}^{-1}_{\text{post}})$;
5. Draw $n_{\text{smp}}$ samples from the Gaussian PDF $\mathcal{N}(\mathbf{m}_*, \mathbf{H}^{-1}_{\text{post}})$;
6. Compute statistical properties of interest from the $n_{\text{smp}}$ samples.
```

Compared to McMC type methods, the additional evaluations of the posterior PDF are not needed once we calculate the MAP estimate $\mathbf{m}_*$, which significantly reduces the computational cost. However, the accuracy of samples drawn from this surrogate PDF strongly depends on the accuracy of the Gaussian approximation in a neighborhood of $\mathbf{m}_*$, which is related



(a) $\lambda = 10$

(b) $\lambda = 50$

(c) $\lambda = 250$

(d) Reduced

Figure 6.3: The Gaussian approximations of the four posterior PDFs associated with the model in equation 6.21.

to our choice of λ. To illustrate this dependence, we continue the example shown in Figure 6.2 and compare the Gaussian approximation of the reduced formulation (i.e., $\lambda = \infty$) to the penalty formulation for different values of λ, plotted in Figure 6.3. In this case the largest eigenvalue is $\mu_1 = 10^4$ and the corresponding is $\lambda = 10$. Clearly, when selecting $\lambda = 10$, the Gaussian approximation is relatively close to the true PDF, whereas increasing λ decreases the accuracy of the Gaussian approximation.

Armed with an accurate Gaussian approximation to the unimodal PDF (for an appropriate choice of λ), we are now in a position to draw samples from the Gaussian distribution $\mathcal{N}(\mathbf{m}_*, \mathbf{H}_{\text{post}}^{-1})$. For small-scale problems, an explicit expression of the Hessian $\mathbf{H}_{\text{post}}$ is available. Hence, we can draw samples $\mathbf{m}_{\text{s}}$ from the distribution $\mathcal{N}(\mathbf{m}_{*}, \mathbf{H}_{\text{post}}^{-1})$ by utilizing the Cholesky factorization of the Hessian $\mathbf{H}_{\text{post}}$—i.e., $\mathbf{H}_{\text{post}} = \mathbf{R}_{\text{post}}^{\top}\mathbf{R}_{\text{post}}$—as follows (Rue, 2001):

$$\mathbf{m}_{\text{s}} = \mathbf{m}_{*} + \mathbf{R}_{\text{post}}^{-1}\mathbf{r}, \tag{6.27}$$

where the matrix $\mathbf{R}_{\text{post}}$ is an upper triangular matrix and the vector $\mathbf{r}$ is a random vector drawn from the n_{grid}-dimensional standard Gaussian distribution $\mathcal{N}(0, \mathcal{I}_{n_{\text{grid}} \times n_{\text{grid}}})$. Nevertheless, for large-scale problems, constructing and storing an explicit expression of the Hessian $\mathbf{H}_{\text{post}}$ is infeasible. Typically, to avoid the construction of the explicit Hessian matrix, the Hessian $\mathbf{H}_{\text{post}}$ is constructed as a matrix-free implicit operator, and we only have access to computing the matrix-vector product with an arbitrary vector. As a result, we need a matrix-free sampling method to draw samples. In the following section, we will develop the implementation details for a suitable sampling method for seismic wave-equation-constrained inverse problems.

6.3 Uncertainty quantification for seismic wave-equation constrained inverse problems

Wave-equation-based inversions, where the coefficients of the wave equation are the unknowns, are amongst the most computationally challenging inverse problems as they typically require wave-equation solves for multiple source experiments on a large domain where the wave travels many wavelengths. Additionally, these inverse problems, also known as full-waveform inversion (FWI) in the seismic community (Pratt, 1999), involve fitting oscillatory predicted and observed data, which can result in parasitic local minima.

Motivated by the penalty formulation with its weak PDE-constraints and the results presented above presumably, we will derive a time-harmonic Bayesian inversion framework that is capable of handling relatively large-scale problems where the wave propagates about 30 wavelengths, a moderate number for realistic problems. Given acoustic seismic data, collected at the surface from sources that fire along the same surface, our aim is to construct the statistical properties of the spatial distribution of the acoustic wave velocity. With these properties, we are able to conduct uncertainty quantification for the recovered velocity model. The subsections are organized roughly in according to the main steps outlined in Algorithm 7.

6.3.1 Steps 1 and 2: designing the posterior and prior PDF

To arrive at a Bayesian formulation for wave-equation-based inverse problems with weak constraints, we consider monochromatic seismic data collected at n_{rcv} receivers locations from n_{src} seismic source locations and sampled at n_{freq} frequencies. Hence, the observed data $\mathbf{d} \in \mathbb{C}^{n_{\text{data}}}$ with $n_{\text{data}} = n_{\text{rcv}} \times n_{\text{src}} \times n_{\text{freq}}$. As before, the synthetic data are obtained by applying, for each source, the projection operator $\mathbf{P} \in \mathbb{C}^{n_{\text{rcv}} \times n_{\text{grid}}}$. For the i^{th} source and j^{th} frequency, the time-harmonic wave equation corresponds to the following discretized Helmholtz system:

$$\mathbf{A}_j(\mathbf{m})\mathbf{u}_{i,j} = \mathbf{q}_{i,j} \quad \text{with} \quad \mathbf{A}_j(\mathbf{m}) = \Delta + \omega_j^2 \operatorname{diag}\left(\mathbf{m}^{-2}\right). \tag{6.28}$$

In this expression, the $\mathbf{q}_{i,j}$'s are the monochromatic sources, the symbol Δ refers to the discretized Laplacian, ω represents the angular frequency, and $\mathbf{m} \in \mathbb{R}^{n_{\text{grid}}}$ denotes the vector with the discretized velocities. With a slight abuse of notation, this vector appears as the elementwise reciprocal square on the diagonal. To discretize this problem, we use the Helmholtz discretization from Chen et al. (2013).

If we consider the data from all sources and frequencies simultaneously, the posterior PDF for the weak-constrained penalty formulation becomes

$$\begin{aligned}
&\overline{\rho}_{\text{post}}(\mathbf{m}|\mathbf{d}) \\
&\propto \exp\left(\sum_{i=1}^{n_{\text{src}}}\sum_{j=1}^{n_{\text{freq}}} -\frac{1}{2}\|\mathbf{P}\overline{\mathbf{u}}_{i,j}(\mathbf{m}) - \mathbf{d}_{i,j}\|^2_{\Gamma^{-1}_{\text{noise}}} - \frac{\lambda^2}{2}\|\mathbf{A}_j(\mathbf{m})\overline{\mathbf{u}}_{i,j}(\mathbf{m}) - \mathbf{q}_{i,j}\|^2\right) \\
&\quad \times \exp\left(-\frac{1}{2}\|\mathbf{m} - \tilde{\mathbf{m}}\|^2_{\Gamma^{-1}_{\text{prior}}}\right).
\end{aligned} \tag{6.29}$$

Aside from choosing a proper value for the penalty parameter λ, another crucial component of the posterior PDF in equation 6.29 is the choice of the prior PDF. From a computational perspective, a suitable prior should have a bounded variance and one should be able to draw samples with moderate cost (Bui-Thanh et al., 2013). More specifically, this results in having computationally feasible access to (matrix-free) actions of the square-root of the prior covariance operator or its inverse on random vectors. To meet this requirement, we utilize Gaussian smoothness priors (Matheron, 2012), which provide a flexible way to describe random fields and are commonly employed in Bayesian inference (Lieberman et al., 2010; Martin et al., 2012; Bardsley et al., 2015). Following Lieberman et al. (2010), we construct a

Gaussian smoothness prior $\rho_{\text{prior}}(\mathbf{m}) \propto \mathcal{N}(\tilde{\mathbf{m}}, \Gamma_{\text{prior}})$ with a reference mean model $\tilde{\mathbf{m}}$ and a covariance matrix Γ_{prior} given by

$$\Gamma_{\text{prior}}(k,l) = a \exp\left(\frac{-\|\mathbf{s}_k - \mathbf{s}_l\|^2}{2b^2}\right) + c\delta_{k,l}. \tag{6.30}$$

In this expression for the covariance, the vectors $\mathbf{s}_k = (z_k, x_k)$ and $\mathbf{s}_l = (z_l, x_l)$ denote the k^{th} and l^{th} spatial coordinates corresponding to the k^{th} and l^{th} elements in the vector $\mathbf{m}$, respectively. The parameters a, b, and c control the correlation strength, variance, and spatial correlation distance. The variance of the k^{th} element m_k is $\text{var}(m_k) = \Gamma_{\text{prior}}(k,k) = a+c$. The parameter c also ensures that the prior covariance matrix remains numerically well-conditioned (Martin et al., 2012). Clearly, when the distance between $\mathbf{s}_k$ and $\mathbf{s}_l$ is large—i.e., $\frac{\|\mathbf{s}_k - \mathbf{s}_l\|^2}{2b^2} \gg 1$, the cross-covariance $\Gamma_{\text{prior}}(k,l)$ vanishes quickly.

The covariance matrix Γ_{prior} is dense. Therefore, for large-scale problems, it is intractable to construct and store it. However, since $\Gamma_{\text{prior}}(k,l)$ tends to zero quickly when $|k-l|$ is sufficiently large, Γ_{prior} has a nearly sparse approximation. To construct this sparse approximation, we follow (Bardsley et al., 2015) and define the correlation length γ as the distance where the cross-covariance $\Gamma_{\text{prior}}(k,l)$ drops to 1% of the $\text{var}(m_k)$. The parameter b can be expressed in terms of the correlation length γ by the expression

$$b = \frac{\gamma}{\sqrt{2\log(100) - 2\log(1+c/a)}}. \tag{6.31}$$

With the correlation length γ, we force

$$\begin{cases} \Gamma_{\text{prior}}(k,l) = a \exp\left(\frac{-\|\mathbf{s}_k - \mathbf{s}_l\|^2}{2b^2}\right) + c\delta_{k,l}, & \text{if } \|\mathbf{s}_k - \mathbf{s}_l\| \le \gamma \\ \Gamma_{\text{prior}}(k,l) = 0, & \text{if } \|\mathbf{s}_k - \mathbf{s}_l\| > \gamma, \end{cases} \tag{6.32}$$

and obtain a sparse Γ_{prior}. Compared to (6.30), the construction of Γ_{prior} by (6.32) is much cheaper and requires less storage. Consequently, our prior is computationally feasible for large-scale problems.

6.3.2 Steps 3 and 4: Gaussian approximation

After setting up the posterior PDF, we need to construct its Gaussian approximation $\mathcal{N}(\mathbf{m}_*, \mathbf{H}_{\text{post}}^{-1})$, corresponding to steps 3 and 4 in Algorithm 7. In order to achieve this objective, first we compute the MAP estimate of the posterior PDF $\mathbf{m}_*$, which is equivalent to minimizing the negative logarithm $-\log \overline{\rho}_{\text{post}}(\mathbf{m}|\mathbf{d})$:

$$\begin{aligned}
\mathbf{m}_* &= \arg\min_{\mathbf{m}} -\log \overline{\rho}_{\text{post}}(\mathbf{m}|\mathbf{d}) \\
&= \arg\min_{\mathbf{m}} \sum_{i=1}^{n_{\text{src}}} \sum_{j=1}^{n_{\text{freq}}} \big(\frac{1}{2}\|\mathbf{P}\overline{\mathbf{u}}_{i,j}(\mathbf{m}) - \mathbf{d}_{i,j}\|^2_{\Gamma^{-1}_{\text{noise}}} + \frac{\lambda^2}{2}\|\mathbf{A}_j(\mathbf{m})\overline{\mathbf{u}}_{i,j}(\mathbf{m}) - \mathbf{q}_{i,j}\|^2\big) \\
&\qquad + \frac{1}{2}\|\mathbf{m} - \tilde{\mathbf{m}}\|^2_{\Gamma^{-1}_{\text{prior}}}.
\end{aligned} \tag{6.33}$$

Note that the objective function $-\log \overline{\rho}_{\text{post}}(\mathbf{m}|\mathbf{d})$ is analogous to the cost function of the deterministic optimization problem (van Leeuwen and Herrmann, 2015). Using similar techniques as in the aforementioned work, we can express the gradient $\mathbf{g}$ for this objective as

$$\mathbf{g} = \sum_{i=1}^{n_{\text{src}}} \sum_{j=1}^{n_{\text{freq}}} \lambda^2 \mathbf{G}_{i,j}^{\top}\big(\mathbf{A}_j(\mathbf{m})\overline{\mathbf{u}}_{i,j}(\mathbf{m}) - \mathbf{q}_{i,j}\big) + \Gamma^{-1}_{\text{prior}}(\mathbf{m} - \tilde{\mathbf{m}}), \tag{6.34}$$

where the sparse Jacobian matrix $\mathbf{G}_{i,j} = \big(\nabla_{\mathbf{m}}\mathbf{A}_j(\mathbf{m})\big)\overline{\mathbf{u}}_{i,j}(\mathbf{m})$. Following van Leeuwen and Herrmann (2013a), we use the limited-memory-Broyden–Fletcher–Goldfarb–Shanno method (l-BFGS, Nocedal and Wright, 2006) to solve the optimization problem in equation 6.33 to find the MAP estimate $\mathbf{m}_*$.

Once we have computed $\mathbf{m}_*$, we focus on approximating the posterior PDF $\overline{\rho}_{\text{post}}(\mathbf{m}|\mathbf{d})$ by a Gaussian distribution centered at $\mathbf{m}_*$. For simplicity, we omit the dependence of $\mathbf{A}_j(\mathbf{m})$ and $\overline{\mathbf{u}}_{i,j}(\mathbf{m})$ on $\mathbf{m}$. A necessary component in this process is computing the Hessian $\mathbf{H}_{\text{post}} = -\nabla^2_{\mathbf{m}} \log \overline{\rho}_{\text{post}}(\mathbf{m}|\mathbf{d})$, which is given by

$$\begin{aligned}
\mathbf{H}_{\text{post}} &= \mathbf{H}_{\text{like}} + \Gamma^{-1}_{\text{prior}} \\
&= \sum_{i}^{n_{\text{src}}} \sum_{j}^{n_{\text{freq}}} \lambda^2 \mathbf{G}_{i,j}^{\top}\mathbf{G}_{i,j} - \mathbf{S}_{i,j}^{\top}\left(\mathbf{P}^{\top}\Gamma^{-1}_{\text{noise}}\mathbf{P} + \lambda^2\mathbf{A}_j^{\top}\mathbf{A}_j\right)^{-1}\mathbf{S}_{i,j} + \Gamma^{-1}_{\text{prior}},
\end{aligned} \tag{6.35}$$

where

$$\mathbf{S}_{i,j} = \lambda^2(\nabla_{\mathbf{m}}\mathbf{A}_j^{\top})\,(\mathbf{A}_j\overline{\mathbf{u}}_{i,j} - \mathbf{q}_{i,j}) + \lambda^2\mathbf{A}_j^{\top}\mathbf{G}_{i,j}, \tag{6.36}$$

and the optimal wavefield $\overline{\mathbf{u}}_{i,j}$ is computed by equation 6.11.

The full Hessian $\mathbf{H}_{\text{post}}$ is a dense $n_{\text{grid}} \times n_{\text{grid}}$ matrix, which is prohibitive to construct explicitly when n_{grid} is large, even in the two-dimensional setting. On the other hand, it is also prohibitive if we formulate the Hessian $\mathbf{H}_{\text{post}}$ as a traditional implicit operator, which requires $2 \times n_{\text{src}} \times n_{\text{freq}}$ PDE solves to compute each matrix-vector product $\mathbf{H}_{\text{post}}\overline{\mathbf{m}}$ with any vector $\overline{\mathbf{m}}$, according to the expression in equation 6.35. Since the posterior covariance matrix is the inverse of the Hessian, we need to invert the square root of the Hessian operator in order to generate random samples. With the implicit Hessian operator, we would need to employ an iterative solver such as LSQR or CG (Golub and Van Loan, 2012), and the total number of PDE solves required is therefore proportional to $2 \times n_{\text{smp}} \times n_{\text{iter}} \times n_{\text{src}} \times n_{\text{freq}}$. As a result, this type of approach requires an extremely large computational cost when drawing sufficiently many samples. As a remedy, we exploit the structure of the Hessian matrix $\mathbf{H}_{\text{post}}$ to find an approximation that can be constructed and applied in a computationally efficient manner.

To exploit the structure of the Hessian matrix $\mathbf{H}_{\text{post}}$, we will focus on the Hessian matrix $\mathbf{H}_{\text{like}}$ of the likelihood term, as we already have discussed the matrix Γ_{prior} in the previous section. Based on equations 6.35 and 6.36, $\mathbf{H}_{\text{like}}$ consists of three components—the matrices $\mathbf{P}^\top\Gamma_{\text{noise}}^{-1}\mathbf{P} + \lambda^2\mathbf{A}_j^\top\mathbf{A}_j$, $(\nabla_{\mathbf{m}}\mathbf{A}_j^\top)(\mathbf{A}_j\overline{\mathbf{u}}_{i,j} - \mathbf{q}_{i,j})$, and $\mathbf{G}_{i,j}$. We first consider the Jacobian $(\nabla_{\mathbf{m}}\mathbf{A}_j^\top)(\mathbf{A}_j\overline{\mathbf{u}}_{i,j} - \mathbf{q}_{i,j})$ and specifically the PDE misfit $\mathbf{A}_j\overline{\mathbf{u}}_{i,j} - \mathbf{q}_{i,j}$. When the PDE misfit is approximately zero, this overall term is also expected to be small. As the MAP estimate $\mathbf{m}_*$ simultaneously minimizes the data misfit, PDE misfit, and model penalty term, the PDE misfit is expected to be small when model and observational errors are not too large. Thus, we obtain a good approximation $\tilde{\mathbf{H}}_{\text{like}}$ of the full Hessian $\mathbf{H}_{\text{like}}$, which corresponds to the Gauss-Newton Hessian derived by van Leeuwen and Herrmann (2015):

$$\tilde{\mathbf{H}}_{\text{like}} = \sum_{i}^{n_{\text{src}}}\sum_{j}^{n_{\text{freq}}} \lambda^2\mathbf{G}_{i,j}^\top\mathbf{G}_{i,j} - \lambda^4\mathbf{G}_{i,j}^\top\mathbf{A}_j(\mathbf{P}^\top\Gamma_{\text{noise}}^{-1}\mathbf{P} + \lambda^2\mathbf{A}_j^\top\mathbf{A}_j)^{-1}\mathbf{A}_j^\top\mathbf{G}_{i,j}. \tag{6.37}$$

Consequently, we can use the Gauss-Newton Hessian $\tilde{\mathbf{H}}_{\text{like}}$ to construct the Gaussian distribution that approximates the posterior PDF $\overline{\rho}_{\text{post}}(\mathbf{m}|\mathbf{d})$:

$$\overline{\rho}_{\text{post}}(\mathbf{m}|\mathbf{d}) \approx \rho_{\text{Gauss}}(\mathbf{m}) = \mathcal{N}(\mathbf{m}_*, \tilde{\mathbf{H}}_{\text{post}}^{-1}) = \mathcal{N}(\mathbf{m}_*, (\tilde{\mathbf{H}}_{\text{like}} + \Gamma_{\text{prior}}^{-1})^{-1}). \tag{6.38}$$

The Gauss-Newton Hessian $\tilde{\mathbf{H}}_{\text{like}}$ has a compact expression derived from the Sherman–Morrison–Woodbury formula (Golub and Van Loan, 2012):

$$\tilde{\mathbf{H}}_{\text{like}} = \sum_{i}^{n_{\text{src}}}\sum_{j}^{n_{\text{freq}}} \mathbf{G}_{i,j}^\top\mathbf{A}_j^{-\top}\mathbf{P}^\top(\Gamma_{\text{noise}} + \frac{1}{\lambda^2}\mathbf{P}\mathbf{A}_j^{-1}\mathbf{A}_j^{-\top}\mathbf{P}^\top)^{-1}\mathbf{P}\mathbf{A}_j^{-1}\mathbf{G}_{i,j}. \tag{6.39}$$

We shall see that this expression provides a factored formulation to implicitly construct the Gauss-Newton Hessian, which does not require any additional PDE solves to compute matrix-vector products. In order to construct the implicit Gauss-Newton Hessian $\tilde{\mathbf{H}}_{\text{like}}$, three matrices are necessary—$\mathbf{A}_j^{-\top}\mathbf{P}^\top \in \mathbb{C}^{n_{\text{grid}} \times n_{\text{rcv}}}$, $\mathbf{G}_{i,j} \in \mathbb{C}^{n_{\text{grid}} \times n_{\text{grid}}}$, and $\Gamma_{\text{noise}} + \frac{1}{\lambda^2}\mathbf{P}\mathbf{A}_j^{-1}\mathbf{A}_j^{-\top}\mathbf{P}^\top \in \mathbb{C}^{n_{\text{rcv}} \times n_{\text{rcv}}}$. For each frequency, constructing the matrix $\mathbf{A}_j^{-\top}\mathbf{P}^\top$ requires n_{rcv} PDE solves. As described in the previous section, the matrix $\mathbf{G}_{i,j}$ is sparse and driven by the corresponding wavefields $\overline{\mathbf{u}}_{i,j}$, whose computational cost approximately equals to one PDE solve for each source and each frequency. The computational complexity of inverting the matrix $\Gamma_{\text{noise}} + \frac{1}{\lambda^2}\mathbf{P}\mathbf{A}_j^{-1}\mathbf{A}_j^{-\top}\mathbf{P}^\top$ is $\mathcal{O}(n_{\text{rcv}}^3)$. Since $n_{\text{rcv}} \ll n_{\text{grid}}$, inverting this matrix is much cheaper than solving a PDE. Thus, to construct the Gauss-Newton Hessian $\tilde{\mathbf{H}}_{\text{like}}$, we only need to solve $n_{\text{freq}} \times (n_{\text{src}} + n_{\text{rcv}})$ PDEs. With the computed matrix $\mathbf{A}_j^{-\top}\mathbf{P}^\top$ and wavefield $\overline{\mathbf{u}}_{i,j}$, the action of the Gauss-Newton Hessian $\tilde{\mathbf{H}}_{\text{like}}$ on any vector $\overline{\mathbf{m}}$ can be performed with several efficient matrix-vector multiplications related to the matrices $\mathbf{A}_j^{-\top}\mathbf{P}^\top$, $\mathbf{G}_{i,j}$, and $\Gamma_{\text{noise}} + \frac{1}{\lambda^2}\mathbf{P}\mathbf{A}_j^{-1}\mathbf{A}_j^{-\top}\mathbf{P}^\top$. Compared to the conventional approach, this new factored formulation of the implicit operator does not require additional PDE solves to compute a matrix-vector multiplication once it is constructed. In general, we have that $n_{\text{rcv}} \ll n_{\text{smp}} \times n_{\text{iter}}$ and using this implicit Gauss-Newton Hessian operator to draw n_{smp} samples is significantly cheaper than the conventional approach. Another advantage of using this operator arises from the fact that the computations of the necessary matrices corresponding to different frequencies are independent from each other. As a result, we can compute and store these matrices in parallel for different frequencies, allowing us to speed up our computations in a distributed computing environment. Furthermore, the expression in equation 6.39 provides a natural decomposition of the Gauss-Newton Hessian $\tilde{\mathbf{H}}_{\text{like}}$ as follows:

$$
\begin{aligned}
&\tilde{\mathbf{H}}_{\text{like}} = \mathbf{R}_{\text{like}}^\top \mathbf{R}_{\text{like}}, \\
&\mathbf{R}_{\text{like}} = \begin{bmatrix}
(\Gamma_{\text{noise}} + \frac{1}{\lambda^2}\mathbf{P}\mathbf{A}_1^{-1}\mathbf{A}_1^{-\top}\mathbf{P}^\top)^{-\frac{1}{2}}\mathbf{P}\mathbf{A}_1^{-1}\mathbf{G}_{1,1} \\
\cdots \\
\\
(\Gamma_{\text{noise}} + \frac{1}{\lambda^2}\mathbf{P}\mathbf{A}_j^{-1}\mathbf{A}_j^{-\top}\mathbf{P}^\top)^{-\frac{1}{2}}\mathbf{P}\mathbf{A}_j^{-1}\mathbf{G}_{i,j} \\
\cdots \\
(\Gamma_{\text{noise}} + \frac{1}{\lambda^2}\mathbf{P}\mathbf{A}_{n_{\text{freq}}}^{-1}\mathbf{A}_{n_{\text{freq}}}^{-\top}\mathbf{P}^\top)^{-\frac{1}{2}}\mathbf{P}\mathbf{A}_{n_{\text{freq}}}^{-1}\mathbf{G}_{n_{\text{src}},n_{\text{freq}}}
\end{bmatrix}.
\end{aligned}
\tag{6.40}
$$

Similarly to the factored formulation of the implicit Gauss-Newton Hessian, we can construct the factor $\mathbf{R}_{\text{like}}$ as an implicit operator once we have computed the matrix $\mathbf{A}_j^{-\top}\mathbf{P}^\top$ and wavefield $\overline{\mathbf{u}}_{i,j}$. We will use this implicit operator $\mathbf{R}_{\text{like}}$ for the sampling method introduced in the next subsection.

6.3.3 Steps 5 and 6: sampling the Gaussian PDFs

The covariance matrix $\tilde{\mathbf{H}}_{\text{post}}^{-1}$ is a dense $n_{\text{grid}} \times n_{\text{grid}}$ matrix, and the construction of its Cholesky factorization involves $\mathcal{O}(n_{\text{grid}}^3)$ operations. Both of these facts prohibit us from applying the Cholesky factorization method to sample the Gaussian distribution $\rho_{\text{Gauss}}(\mathbf{m})$ for large-scale problems. Here we propose to use the so-called optimization-driven Gaussian simulators (Papandreou and Yuille, 2010; Orieux et al., 2012) or the randomize-then-optimize (RTO) method (Solonen et al., 2014) to sample the Gaussian distribution $\rho_{\text{Gauss}}(\mathbf{m})$. This method belongs to the classical bootstrap method (Efron, 1981, 1992; Kitanidis, 1995), and it does not require the explicit formulation of the Hessian matrix as well as the expensive Cholesky factorization. To outline this method, we first use equation 6.38 to divide $\tilde{\mathbf{H}}_{\text{post}}$ into $\tilde{\mathbf{H}}_{\text{like}}$ and $\Gamma_{\text{prior}}^{-1}$. The Gauss-Newton Hessian $\tilde{\mathbf{H}}_{\text{like}}$ has the factorization in equation 6.40, and we can also compute the Cholesky factorization of the prior covariance matrix $\Gamma_{\text{prior}} = \mathbf{R}_{\text{prior}}^\top \mathbf{R}_{\text{prior}}$ with an upper-triangular matrix $\mathbf{R}_{\text{prior}}$. Substituting these two factorizations into equation 6.38, we can rewrite the Gaussian distribution $\rho_{\text{Gauss}}(\mathbf{m})$ as follows:

$$\rho_{\text{Gauss}}(\mathbf{m}) \propto \exp\left(-\frac{1}{2}\|\mathbf{R}_{\text{like}}\mathbf{m} - \mathbf{R}_{\text{like}}\mathbf{m}_*\|^2 - \frac{1}{2}\|\mathbf{R}_{\text{prior}}^{-\top}\mathbf{m} - \mathbf{R}_{\text{prior}}^{-\top}\mathbf{m}_*\|^2\right). \tag{6.41}$$

Papandreou and Yuille (2010) and Solonen et al. (2014) noted that independent realizations from the distribution in equation 6.41 can be computed by repeatedly solving the following linear data-fitting optimization problem:

$$\begin{aligned}\mathbf{m}_{\text{s}} = \arg\min_{\mathbf{m}} & \|\mathbf{R}_{\text{like}}\mathbf{m} - \mathbf{R}_{\text{like}}\mathbf{m}_* - \mathbf{r}_{\text{like}}\|^2 + \left\|\mathbf{R}_{\text{prior}}^{-\top}\mathbf{m} - \mathbf{R}_{\text{prior}}^{-\top}\mathbf{m}_* - \mathbf{r}_{\text{prior}}\right\|^2, \\ & \mathbf{r}_{\text{like}} \sim \mathcal{N}(\mathbf{0}, \mathcal{I}_{n_{\text{data}}\times n_{\text{data}}}), \\ & \mathbf{r}_{\text{prior}} \sim \mathcal{N}(\mathbf{0}, \mathcal{I}_{n_{\text{grid}}\times n_{\text{grid}}}).\end{aligned} \tag{6.42}$$

This optimization problem can be solved by iterative solvers such as LSQR and PCG, which do not require the explicit expression for the matrix $\mathbf{R}_{\text{like}}$ but merely an operator that can compute the matrix-vector product. As a result, we can use our implicit formulation of $\mathbf{R}_{\text{like}}$ in equation 6.40 to solve the optimization problem in equation 6.42 and draw samples from the Gaussian distribution $\rho_{\text{Gauss}}(\mathbf{m})$. The pseudo code of the RTO method to draw samples from the Gaussian distribution $\rho_{\text{Gauss}}(\mathbf{m})$ is shown in Algorithm 8, which realizes step 5 in Algorithm 7. Because the overall sampling strategy consists of the Gaussian approximation and the RTO method, we will refer to the proposed method as GARTO in the rest of the chapter.

Algorithm 8 Sample $\rho_{\text{Gauss}}(\mathbf{m})$ by the RTO method

```
1. Start with the MAP estimate $\mathbf{m}_*$, covariance matrices $\Gamma_{\text{noise}}$ and $\Gamma_{\text{prior}}$;
2. Formulate the operator $\mathbf{R}_{\text{like}}(\mathbf{m}_*)$ by equation 6.40, and compute the
   Cholesky factorization of $\Gamma_{\text{prior}} = \mathbf{R}_{\text{prior}}^{\top}\mathbf{R}_{\text{prior}}$;
3. for s = 1:$n_{\text{smp}}$
4.     Generate $\mathbf{r}_{\text{prior}} \sim \mathcal{N}(\mathbf{0}, \mathcal{I}_{n_{\text{grid}}\times n_{\text{grid}}})$ and $\mathbf{r}_{\text{like}} \sim \mathcal{N}(\mathbf{0}, \mathcal{I}_{n_{\text{data}}\times n_{\text{data}}})$;
5.     Solve $\mathbf{m}_{\text{s}} = \arg\min_{\mathbf{m}} \|\mathbf{R}_{\text{like}}\mathbf{m} - \mathbf{R}_{\text{like}}\mathbf{m}_* - \mathbf{r}_{\text{like}}\|^2$
                 $+ \left\|\mathbf{R}_{\text{prior}}^{-\top}\mathbf{m} - \mathbf{R}_{\text{prior}}^{-\top}\mathbf{m}_* - \mathbf{r}_{\text{prior}}\right\|^2$;
6. end
```

6.3.4 A benchmark method: the randomized maximum likelihood method

We have proposed a computationally efficient algorithm—GARTO—that can approximately sample the target distribution in equation 6.29 without additional PDE solves once the MAP estimate and the Gauss-Newton Hessian operator are computed. However, due to the loss of accuracy caused by the Gaussian approximation, it is important to investigate the accuracy of the GARTO method by comparing it to a benchmark method that can sample the target distribution in equation 6.29 regardless of the computational cost. The randomized maximum likelihood (RML) (Chen and Oliver, 2012) method is a viable candidate as a benchmark because it has the capability to draw samples that are good approximations of those from the target distribution for weakly nonlinear problems with Gaussian prior distribution and additive Gaussian distributed errors (Chen and Oliver, 2012; Bardsley et al., 2014, 2015). Indeed, as previously discussed, the target distribution with weak PDE-constraints that the GARTO method aims to sample is less prone to the nonlinearity with a carefully selected λ.

To draw a sample, the RML method performs a bootstraping-like method (Efron, 1981, 1992) that first samples the data and prior model and then computes the resulting MAP. More precisely, in order to draw a sample from the target distribution in equation 6.29, the RML method solves the following nonlinear optimization problem:

$$
\begin{aligned}
\mathbf{m}_s = \arg\min_{\mathbf{m}} \sum_{i=1}^{n_{\text{src}}} \sum_{j=1}^{n_{\text{freq}}} \Big(& \frac{1}{2} \| \Gamma_{\text{noise}}^{-\frac{1}{2}} \mathbf{P} \overline{\mathbf{u}}_{i,j}(\mathbf{m}) - \Gamma_{\text{noise}}^{-\frac{1}{2}} \mathbf{d}_{i,j} - \mathbf{r}_{i,j}^{(1)} \|^2 \\
& + \frac{1}{2} \| \lambda \mathbf{A}_j(\mathbf{m}) \overline{\mathbf{u}}_{i,j}(\mathbf{m}) - \lambda \mathbf{q}_{i,j} - \mathbf{r}_{i,j}^{(2)} \|^2 \Big) \\
& + \frac{1}{2} \| \Gamma_{\text{prior}}^{-\frac{1}{2}} \mathbf{m} - \Gamma_{\text{prior}}^{-\frac{1}{2}} \tilde{\mathbf{m}} - \mathbf{r}^{(3)} \|^2,
\end{aligned}
\tag{6.43}
$$

where the vectors $\mathbf{r}_{i,j}^{(1)}$, $\mathbf{r}_{i,j}^{(2)}$, and $\mathbf{r}^{(3)}$ are random realizations from the standard norm distribution $\mathcal{N}(0, \mathbf{I})$. We refer interested readers to Chen and Oliver (2012) for more details about the RML method. As a result of this approach, the computational cost of drawing one sample by the RML method is approximately equivalent to solving one FWI problem, which is significantly more expensive than the GARTO method. Therefore, we are only able to conduct an comparison with the RML method on a small-scale problem in the following section.

6.4 Numerical experiments

6.4.1 Influence of the penalty parameter λ

The feasibility of the proposed Bayesian framework relies on the accuracy of the approximations in equations 6.25 and 6.38, both of which depend on the selection of λ. To get a better understanding of the influence of this parameter on these approximations, we will work with a relatively simple 2D velocity model parameterized by a single varying parameter v_0—i.e., the velocity is given by $v(z) = v_0 + 0.75z\,\text{m/s}$ and z increases with vertical depth. We simulate data with a single source and single frequency for $v_0 = 2000\,\text{m/s}$ using a grid spacing of 25 m. The frequency of the data is 5 Hz, which is suitable for avoiding numerical dispersion on this particular grid spacing. We place our source at (z, x) coordinates $(50\,\text{m}, 50\,\text{m})$ and record the data at 200 receivers located at the depth of 50 m with a sampling interval of 25 m. We do not simulate the free surface in this example. After the simulation, we add 10% Gaussian noise to the observed data. Because the prior distribution is independent of λ, we only investigate its influence on the negative log-likelihood of the associated PDFs in this experiment. We abbreviate the negative log-likelihood with "NLL".

Figure 6.4 shows the NLL for the reduced approach (cf. equation 6.8) as

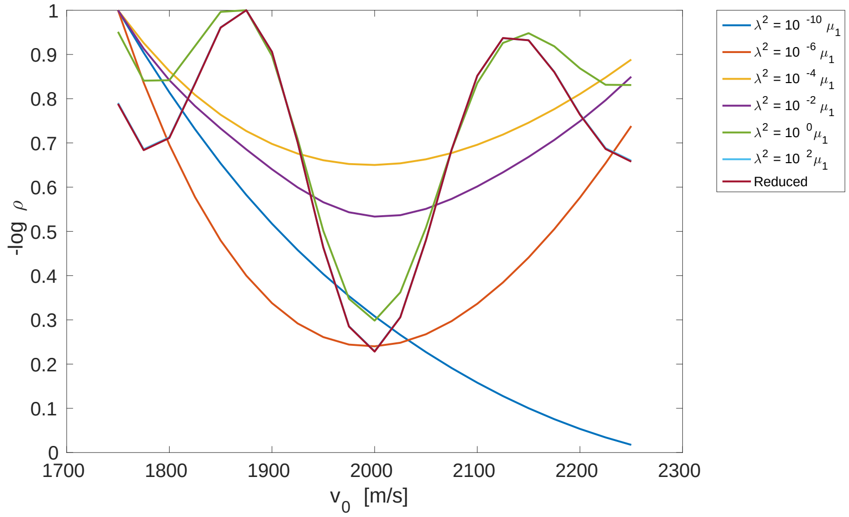


Figure 6.4: The comparison of the negative log-likelihood functions.

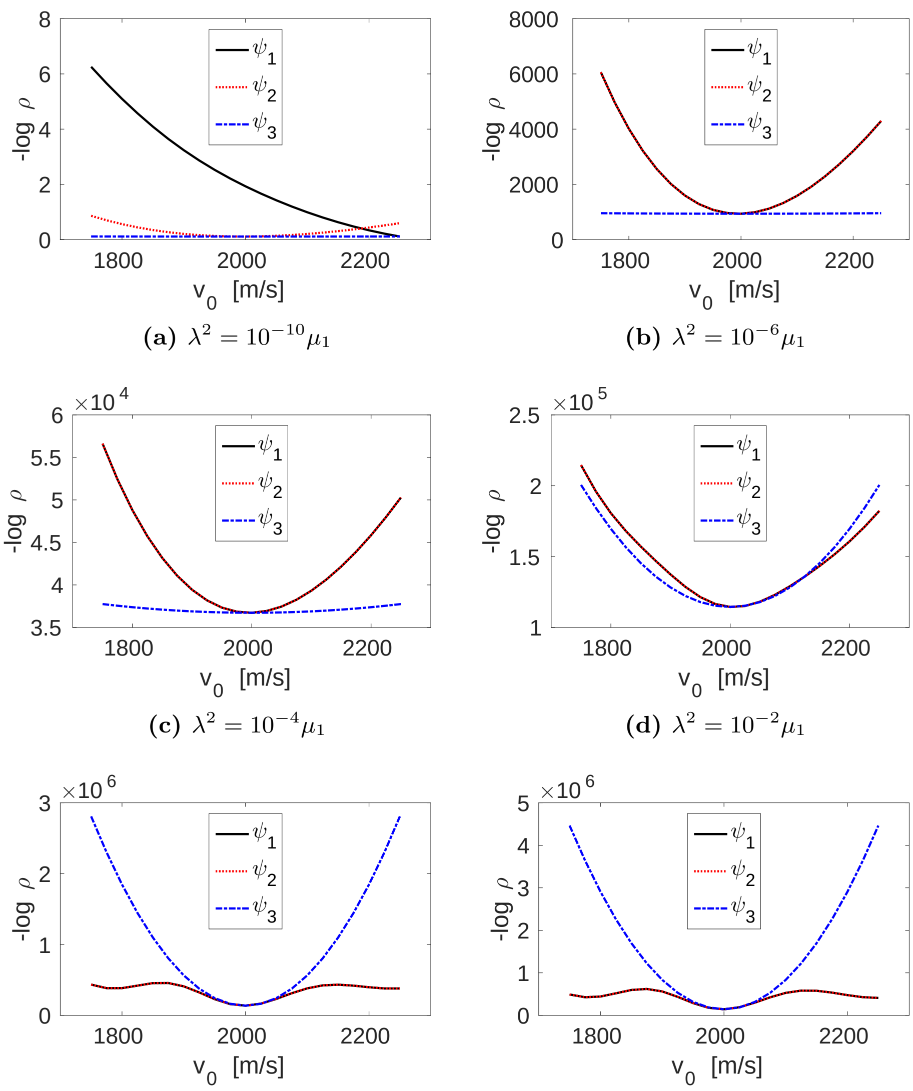
well as for the penalty approach (cf. equation 6.16) for various values of λ as a function of v_0. As discussed previously, we select values of λ^2 proportional to the largest eigenvalue μ_1 of the matrix $\mathbf{A}(\mathbf{m})^{-\top}\mathbf{P}^{\top}\Gamma_{\text{noise}}^{-1}\mathbf{P}\mathbf{A}(\mathbf{m})^{-1}$, i.e., $\lambda^2 = 10^{-10}\mu_1$, $10^{-6}\mu_1$, $10^{-4}\mu_1$, $10^{-2}\mu_1$, $10^{0}\mu_1$, and $10^{2}\mu_1$. From this figure, we observe that, when λ is large, i.e., $\lambda^2 = 10^2\mu_1$ and $10^0\mu_1$, the NLL exhibits several local minima, and, as λ increases, it converges to the reduced approach formulation, as expected. We also note that for small λ, i.e., $\lambda^2 = 10^{-10}\mu_1$, the resulting NLL is monotonically decreasing and does not have a global minimum, which is due to the fact that the determinant term in equation 6.22 dominates the NLL. Additionally, in between these two extreme values for λ (i.e., when $\lambda^2 = 10^{-6}\mu_1$, $10^{-4}\mu_1$, and $10^{-2}\mu_1$), the resulting NLLs are unimodal with a global minimum, which slightly differs from the one of the reduced formulation due to the fact that the wavefields computed by the penalty formulation tend to simultaneously match the data and PDEs instead of the PDEs alone. This observation implies that with a carefully selected λ, the posterior distribution with weak PDE-constraints potentially contains less local maxima.

To investigate the influence of the parameter λ on the accuracy of the approximations in equations 6.25 and 6.38, we plot in Figure 6.5 the NLL corresponding to the true (cf. equation 6.16) $\rho_{\text{post}}(\mathbf{m}|\mathbf{d})$, its approximation neglecting the determinant term $\overline{\rho}_{\text{post}}(\mathbf{m}|\mathbf{d})$ (cf. equation 6.25), and the

Gaussian approximation $\rho_{\text{Gauss}}(\mathbf{m})$ (cf. equation 6.38) for various values of λ. For simplicity, we refer to these three different functions as ψ_1, ψ_2, and ψ_3, respectively. From Figure 6.5a, we observe that when $\lambda^2 = 10^{-10}\mu_1$, ψ_2 fails to adequately approximate ψ_1—i.e., the approximation in equation 6.25 fails—because the determinant term in equation 6.16 dominates the negative logarithm function. As λ increases, the determinant term becomes negligible, and ψ_2 becomes a reasonable approximation of ψ_1, as shown in Figures 6.5b to 6.5f. However, among the five various selections of λ, only when $\lambda^2 = 10^{-2}\mu_1$ does ψ_3 adequately approximate ψ_2. This occurs because when λ is relatively large, ψ_2 contains a number of nonoptimal local minima, resulting in ψ_2 being poorly modeled by its Gaussian approximation. Additionally, when $\lambda^2 < 10^{-2}\mu_1$, the term $\mathbf{A}_j\overline{\mathbf{u}}_{i,j} - \mathbf{q}_{i,j}$ in equation 6.36 is not negligible, resulting in the Gauss-Newton Hessian being a poor approximation of the full Hessian. These results imply that the proposed criterion—i.e., $\lambda^2 = 10^{-2}\mu_1$—provides a reasonable choice for the parameter λ, which can simultaneously satisfy both the approximations in equations 6.25 and 6.38. As a result, the corresponding posterior distribution is less prone to the local maxima and can be appropriately approximated by the Gaussian distribution in equation 6.38, which ensures the feasibility of the proposed framework.

6.4.2 A 2D layered model

In this section, we develop an example to compare the accuracy of the statistical quantities produced by the GARTO method relative to those produced by the RML method. Considering the large computational cost required by the RML method, we use a small three-layer model as our baseline model, as shown in Figure 6.6a, whose size is 1500 m × 3000 m. We discretize the velocity model with a grid spacing of 50 m, yielding 1800 unknown parameters. At the surface, we place sixty sources and receivers with a horizontal sampling interval of 50 m. To control the computational cost and ensure the feasibility of the RML method, we use a Ricker wavelet centered at 6 Hz to simulate the observed data with frequencies of 5, 6, and 7 Hz. We use an absorbing boundary condition at the surface, so that no surface-related multiples are included. After the simulation, we add 15% Gaussian noise to the observed data resulting in a covariance matrix of $\Gamma_{\text{noise}} = 175^2\mathbf{I}$. To set up the prior distribution, we first construct the monotonically increasing model shown in Figure 6.6b as the prior mean model, which corresponds to the well-known observation that the velocity of the Earth, in general, increases with depth. Following the strategy used by Bardsley et al. (2015)



(a) $\lambda^2 = 10^{-10}\mu_1$

(b) $\lambda^2 = 10^{-6}\mu_1$

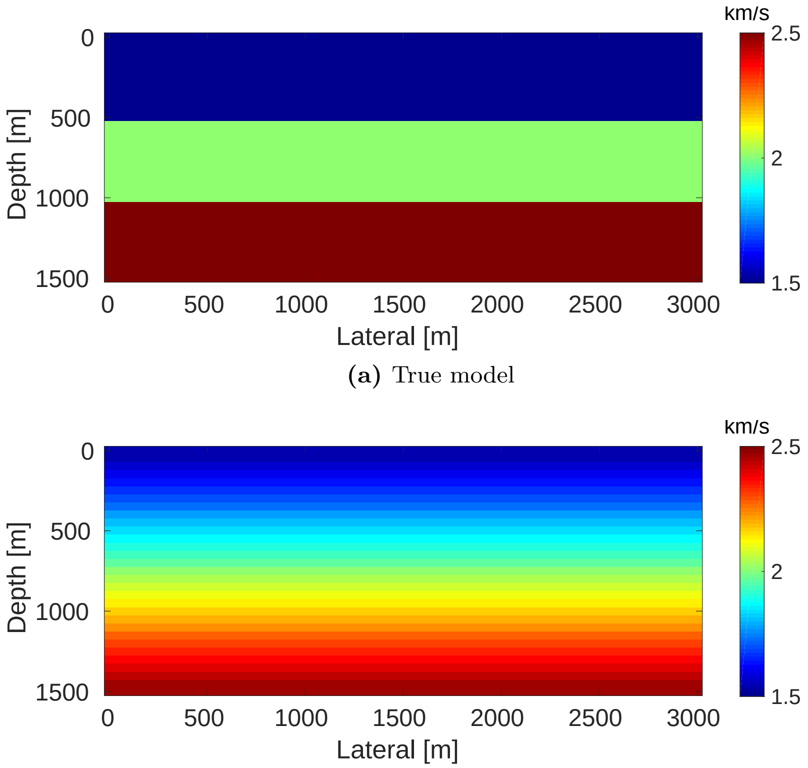
(c) $\lambda^2 = 10^{-4}\mu_1$

(d) $\lambda^2 = 10^{-2}\mu_1$

(e) $\lambda^2 = 10^{0}\mu_1$

(f) $\lambda^2 = 10^{2}\mu_1$

Figure 6.5: The comparison of the negative log-likelihood functions of the true distribution (ψ_1), the approximate distribution (ψ_2), and the Gaussian approximation distribution (ψ_3) with the values of $\lambda^2 = 10^{-10}\mu_1$, $10^{-6}\mu_1$, $10^{-4}\mu_1$, $10^{-2}\mu_1$, $10^{0}\mu_1$, and $10^{2}\mu_1$.



(a) True model

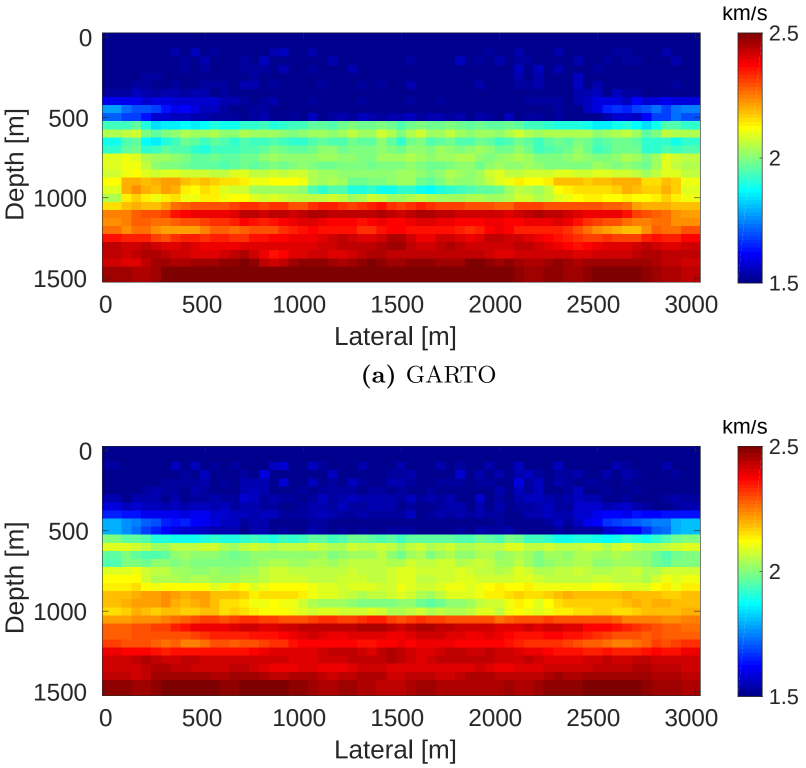
(b) Prior mean model

Figure 6.6: The true model (a) and the prior mean model (b).

that ensures the prior covariance matrix is well-conditioned, we construct the prior covariance matrix by selecting $a = 0.1\,\mathrm{km}^2/\mathrm{s}^2$, $b = 0.65$, and $c = 0.01\,\mathrm{km}^2/\mathrm{s}^2$, resulting in a prior standard deviation of $0.33\,\mathrm{km/s}$ that meets the maximal difference between the true and prior mean models. We select the penalty parameter λ for each frequency according to the proposed selection criterion, resulting in $\lambda = 13,\ 12,$ and 11, respectively. To compute reliable statistical quantities and while also ensuring that the RML method terminates in a reasonable amount of time, we use both the GARTO and the RML methods to generate $10,000$ samples from the posterior distribution. Based on our experience, generating $10,000$ samples is sufficient for both methods to produce stable results in this case.

Given that the computational overhead introduced by these methods is negligible compared to the number of PDEs that need to be solved, we use the number of PDEs as our performance metric. To generate $10,000$ samples, the RML method needs to solve $10,000$ nonlinear optimization problems. To solve each nonlinear optimization problem, following van Leeuwen et al. (2014), we stop the optimization when the relative misfit difference between two iterations drops below 10^{-3} resulting in 100 l-BFGS iterations. During each l-BFGS iteration, we have to solve $2 \times 3 \times 60$ PDEs to compute the objective and gradient. As a result, the RML method requires $360 \times 100 \times 10000 = 360$ million PDE solves to draw the $10,000$ samples. Contrary to the RML method, the GARTO method requires significantly fewer PDE solves. The total number of PDE solves required by the GARTO method is $36,360$, which includes $36,000$ PDE solves to find the MAP estimate and another 360 PDE solves to construct the Gauss-Newton operator. With the MAP estimate and the Gauss-Newton operator in hand, the GARTO method uses the RTO approach to sample the $10,000$ samples without involving any additional PDE solves as we explained above. Therefore, neglecting the costs associated with solving the least-squares systems, compared to the RML method, the GARTO method requires only $\frac{1}{10000}$th the computational budget to generate the same number of samples.

In addition to the significant computational speedups introduced by the GARTO method, the GARTO method also generates samples that yield similar statistical quantities as those produced by the RML method. For instance, the posterior mean models obtained by the two methods are shown in Figure 6.7. From Figures 6.7a and 6.7b, we observe that aside from some slight differences in the second and third layers, the two results are roughly identical, with an average pointwise relative difference of 1.5%. Both results provide acceptable reconstructions of the original velocity model, despite the fact that the data are noisy. We also use the two posterior mean models to compute the predicted data and compare them with the observed data in Figures 6.8a and 6.8b, which faithfully match the observed data, aside from the noise. The pointwise standard deviations computed from both methods, shown in Figure 6.9, result in estimates that are visually quite similar throughout the entire model. The average relative difference between the standard deviations produced by both methods is 6%, an acceptable error level, and results from the Gaussian approximation in GARTO. Figure 6.10 depicts the 95% confidence intervals obtained with the two methods at three vertical cross-sections through the model. The confidence intervals obtained by the two methods are virtually identical, as are the pointwise marginal distributions shown in Figure 6.11. All these results illustrate that, compared to the RML method, the proposed GARTO method can produce accurate estimates of statistical quantities, while requiring a fraction of the computational costs.



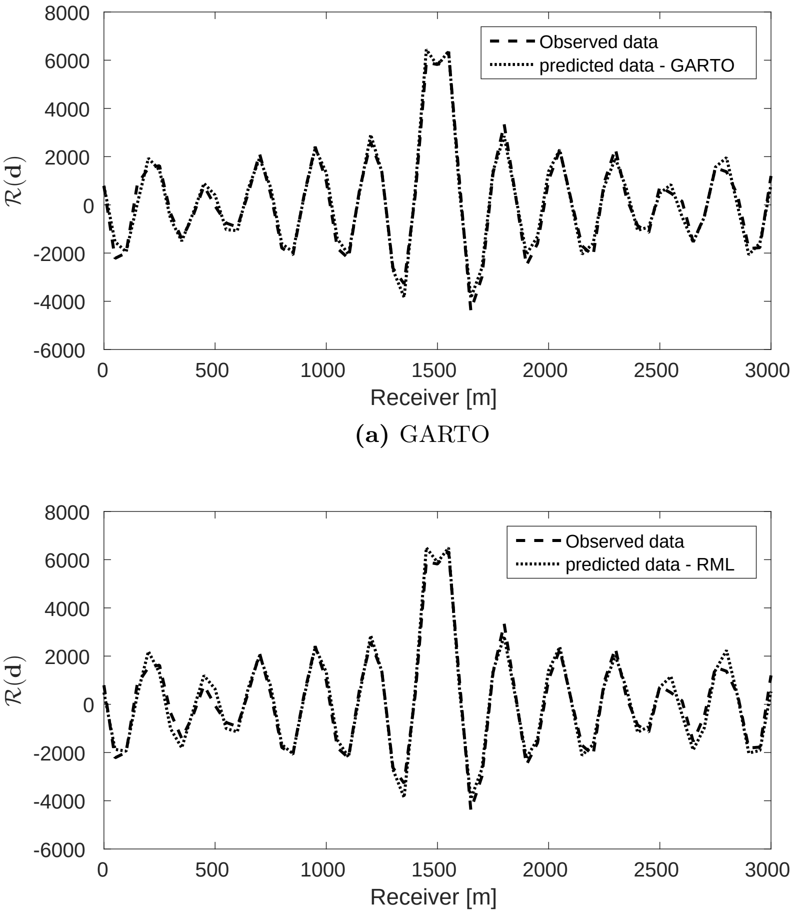
(a) GARTO

(b) RML

Figure 6.7: The posterior mean models obtained by the GARTO method (a) and the RML method (b).

6.4.3 BG Model

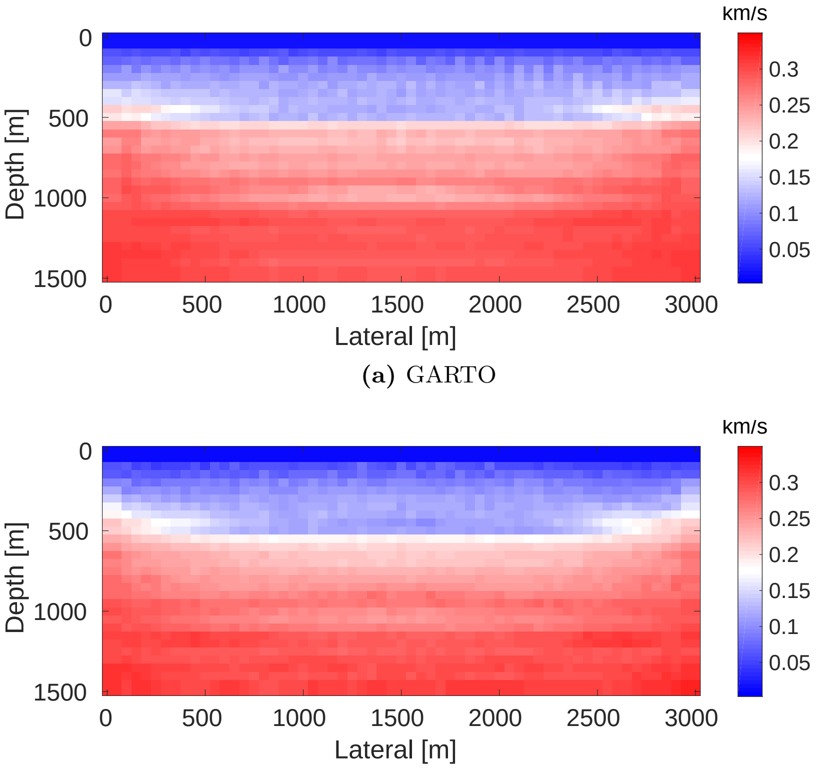
To demonstrate the effectiveness of our method when applied to a more realistic problem, we conduct an experiment on a 2D slice from the 3D BG Compass model shown in Figure 6.12a. This is a synthetic velocity model created by BG Group, which contains a large degree of variability and has been widely used to evaluate performances of different FWI methods (Li et al., 2012; van Leeuwen and Herrmann, 2013a). The model is 2000 m deep and 4500 m wide, with a grid size of 10 m resulting in 92,455 unknown parameters. Following van Leeuwen and Herrmann (2013a), we use a Ricker wavelet with a central frequency of 15 Hz to simulate 91 sources at the depth of 50 m with a horizontal sampling interval of 50 m. As before, we do not model the free-surface, so that the data do not contain free-surface related multiples. We place 451 equally spaced receivers at the same depth



(a) GARTO

(b) RML

Figure 6.8: The comparison between the observed data and the predicted data with the posterior mean models obtained by the GARTO method (a) and the RML method (b).

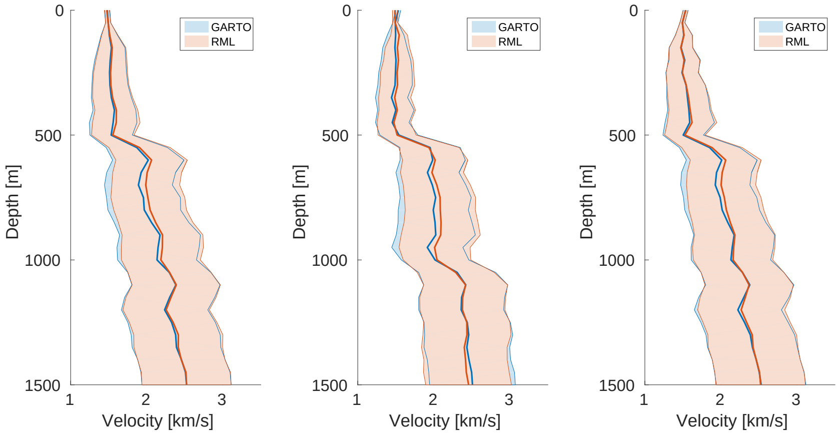


(a) GARTO

(b) RML

Figure 6.9: The posterior standard deviations obtained by the GARTO method (a) and the RML method (b).

as the sources to record the data, which contain 30 equally spaced frequency components ranging from 2 Hz to 31 Hz. This results in 1,231,230 observed data values. To mimic a real-world noise level, we corrupt the synthetic observations with 15% random Gaussian noise, leading to $\Gamma_{\text{noise}} = 46^2\mathbf{I}$. To construct the prior distribution, we first set its mean model (Figure 6.12b) by smoothing the true model followed by horizontal averaging. Second, we construct the covariance matrix of the prior distribution utilizing the fact that we have the true 3D model, which contains 1800 2D slices. We regard these 2D slices as 1800 2D samples, from which we compute the pointwise standard deviation. After horizontal averaging, we obtain the prior standard deviation shown in Figure 6.12c. With the prior standard deviation, we select $a = 0.02\,\text{km}^2/\text{s}^2$ and $b = 19.5$ to construct a well-conditioned covariance matrix with a correlation length of 60 m (Bardsley et al., 2015).

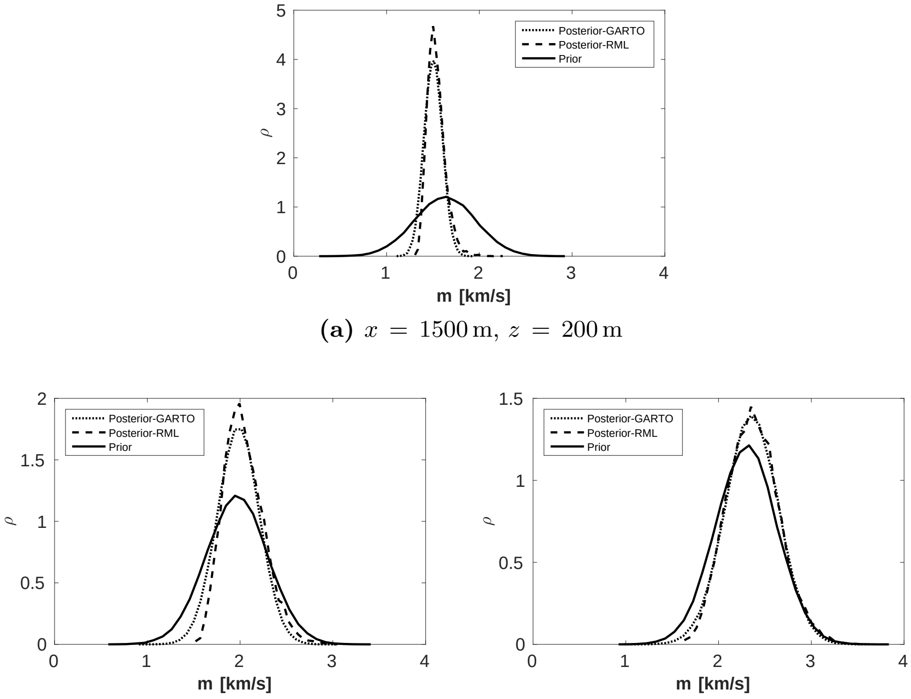


(a) $x = 500\,\text{m}$ **(b)** $x = 1500\,\text{m}$ **(c)** $x = 2500\,\text{m}$

Figure 6.10: The mean (line) and 95% confidence interval (background patch) comparisons of the GARTO method (blue) and the RML method (red) at $x = 500\,\text{m}$, $1500\,\text{m}$, and $2500\,\text{m}$. The similarity between these two results implies that the confidence intervals obtained with the GARTO method is a good approximation of the ones obtain with the RML method.

The parameter c in equation 6.30 is calculated according to the standard deviation and the parameters a and b. Finally, we compute the penalty parameter λ for each frequency (listed in Table 6.1) according to the criterion introduced earlier in order to obtain a posterior distribution that is less prone to local maxima. Considering both the computational resources and the accuracy of the inverted statistical quantities, we will use the GARTO method to draw 2000 samples according to Bardsley et al. (2015). Compared to the previous example, which had a much simpler model, this model contains a significantly larger number of unknown parameters and data points and is a better proxy for real-world problems.

During the inversion, we use 200 l-BFGS iterations to compute the MAP estimate plotted in Figure 6.13a with the same stopping criterion as in the previous example. Compared to the true model, we observe that most of the observable velocity structures are reconstructed in the MAP estimate, aside from some small measurement noise related imprints near the boundary of the model. We also observe that the shallow parts ($z \leq 1400\,\text{m}$) of the BG model, where the turning waves exist (for a maximal offset of 4500 m



(a) $x\,=\,1500\,\mathrm{m}$, $z\,=\,200\,\mathrm{m}$

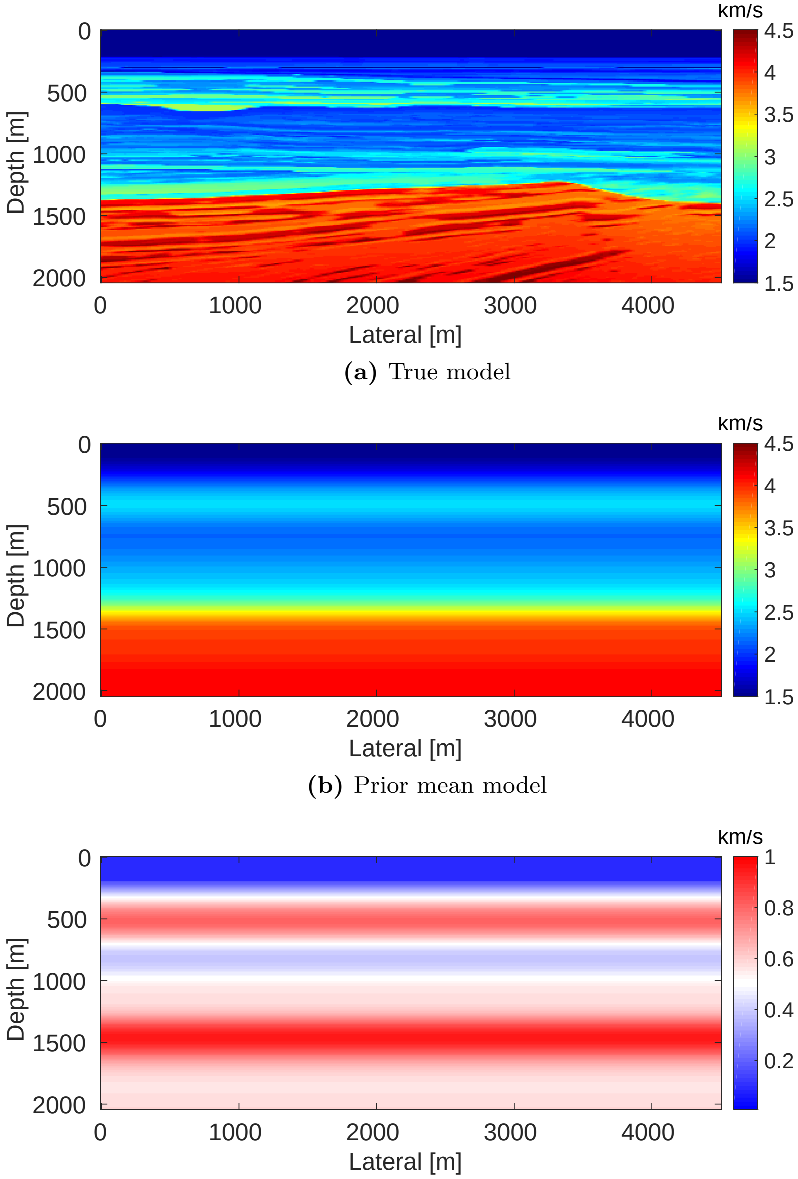
(b) $x\,=\,1500\,\mathrm{m}$, $z\,=\,700\,\mathrm{m}$

(c) $x\,=\,1500\,\mathrm{m}$, $z\,=\,1200\,\mathrm{m}$

Figure 6.11: The comparison of the prior marginal distribution (solid line) and the posterior marginal distributions obtained by the GARTO method (dotted line) and the RML method (dashed line) at the locations of $(x,\,z)\;=\;$(1500m, 200m), (1500m, 700m), and (1500m, 1200m).

Frequency	2	3	4	5	6	7	8	9	10	11
λ	37.8	29.3	24.4	20.5	17.6	15.2	13.5	12.1	10.9	10.1
Frequency	12	13	14	15	16	17	18	19	20	21
λ	9.3	8.8	8.3	7.9	7.5	7.1	6.9	6.5	6.4	6.1
Frequency	22	23	24	25	26	27	28	29	30	31
λ	5.9	5.7	5.5	5.4	5.2	5.1	4.9	4.8	4.7	4.6

Table 6.1: The selection of λ for each frequency



(a) True model

(b) Prior mean model

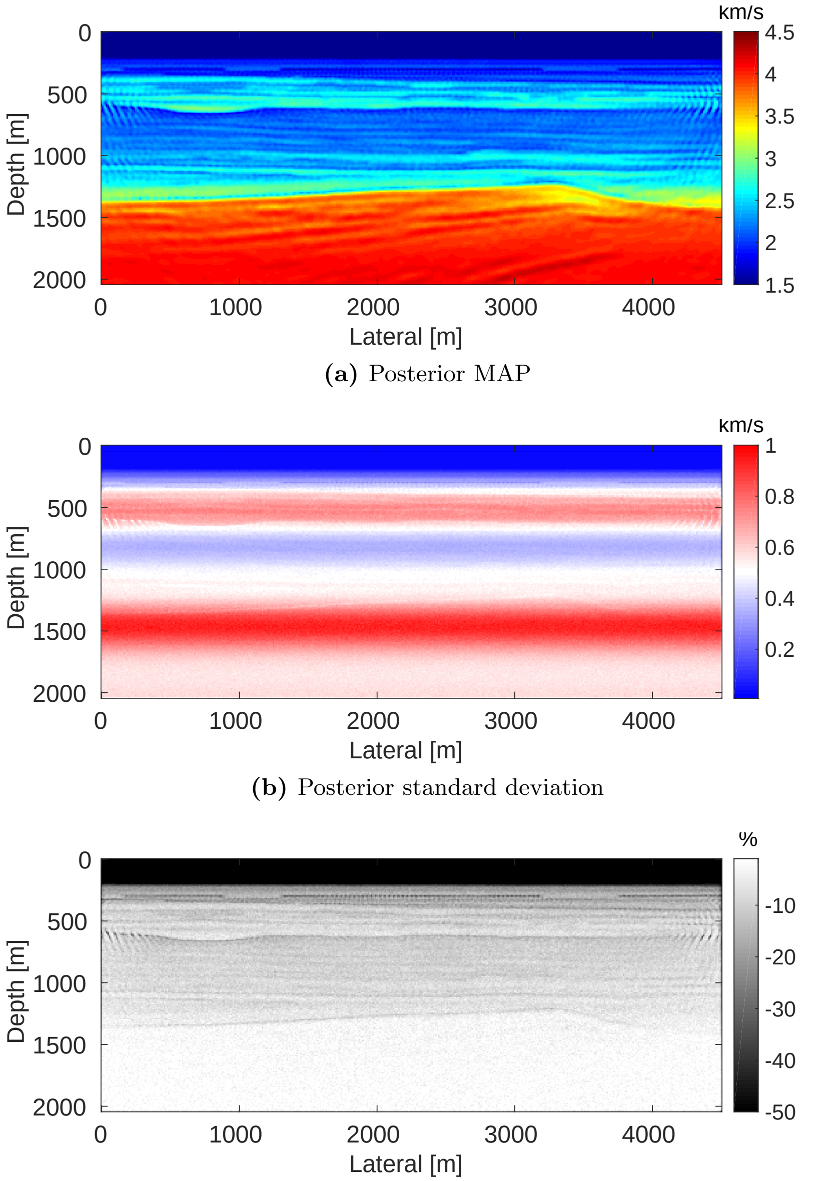
(c) Prior standard deviation

Figure 6.12: (a) The true model, (b) the prior mean model, and (c) the prior standard deviation.

(Virieux and Operto, 2009)) are better recovered relative to the deep parts ($z \geq 1400\,\mathrm{m}$), where the only received energy arises from reflected waves. This implies that the portion of the data corresponding to the turning waves is more informative to the velocity reconstruction than that of the reflected waves, which is a well-known observation in seismic inversions.

After obtaining the MAP estimate, we construct the Gauss-Newton Hessian operator and apply the RTO method to generate 2000 samples. This allows us to compute the posterior standard deviation (Figure 6.13b) and compare it with the prior standard deviation (Figure 6.12c). To have a better understanding of the information that the data introduce, we also compute the relative difference $\frac{\mathrm{STD}_{\mathrm{post}}(m_k)-\mathrm{STD}_{\mathrm{prior}}(m_k)}{|\mathrm{STD}_{\mathrm{prior}}(m_k)|}$ between the posterior and prior standard deviations (Figure 6.13c). In the shallow parts of the model ($z \leq 1400\,\mathrm{m}$), the posterior standard deviation decreases between $10\% - 50\%$ compared to the prior standard deviation, while in the deep parts ($z \geq 1400\,\mathrm{m}$), the reduction in standard deviation is smaller than 3%. This observation is consistent with the notion that, owing to the amplitude decay of propagating waves, the data place more constraints on the velocity variations in the shallow parts of the model compared to the deep parts. Additionally, considering the areas where the turning waves and the reflected waves exist, this observation also implies that the portion of the data corresponding to the turning waves can reduce more uncertainties in the recovered velocity compared to the reflected waves. To further evaluate this inversion result, we compare the prior model, the MAP estimate of the posterior, and the true velocity at three different cross sections in Figure 6.14 (i.e., $x = 1000\,\mathrm{m}$, $2500\,\mathrm{m}$, and $4000\,\mathrm{m}$), in which we also plot the prior standard deviation (red patch) and posterior standard deviation (blue patch). In the shallow region of the model, the MAP estimates closely match the true model, while they diverge in the deeper region in a more pronounced manner. This is again consistent with the notion that the data are more informative in the shallow area of the model compared to the deeper areas.

We also consider the pointwise posterior marginal distribution generated by the posterior and prior distributions to further understand the results of the GARTO method. Figure 6.15 compares these distributions at four different locations, $(x, z) = (2240\,\mathrm{m}, 40\,\mathrm{m})$, $(2240\,\mathrm{m}, 640\,\mathrm{m})$, $(2240\,\mathrm{m}, 1240\,\mathrm{m})$, and $(2240\,\mathrm{m}, 1840\,\mathrm{m})$. The posterior distribution is more concentrated than the prior distribution in the shallow regions of the model, while in the deep parts, the two distributions are almost identical. Therefore, the recovered velocity in the shallow parts is more reliable than in the deep parts.

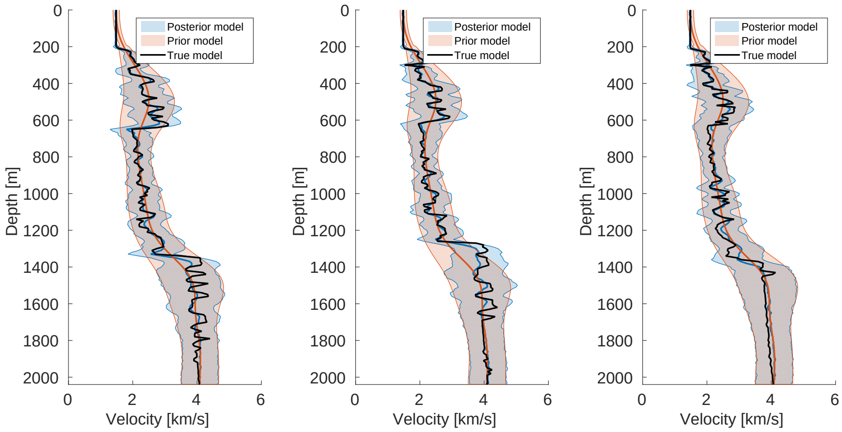


(a) Posterior MAP

(b) Posterior standard deviation

(c) Relative difference between the posterior and prior standard deviations

Figure 6.13: (a) The posterior MAP estimate of the model, (b) the posterior standard deviation, and (c) the relative difference between the posterior and the prior standard deviations, i.e., $\frac{\mathrm{STD}_{\mathrm{post}}(m_k)-\mathrm{STD}_{\mathrm{prior}}(m_k)}{|\mathrm{STD}_{\mathrm{prior}}(m_k)|}$.



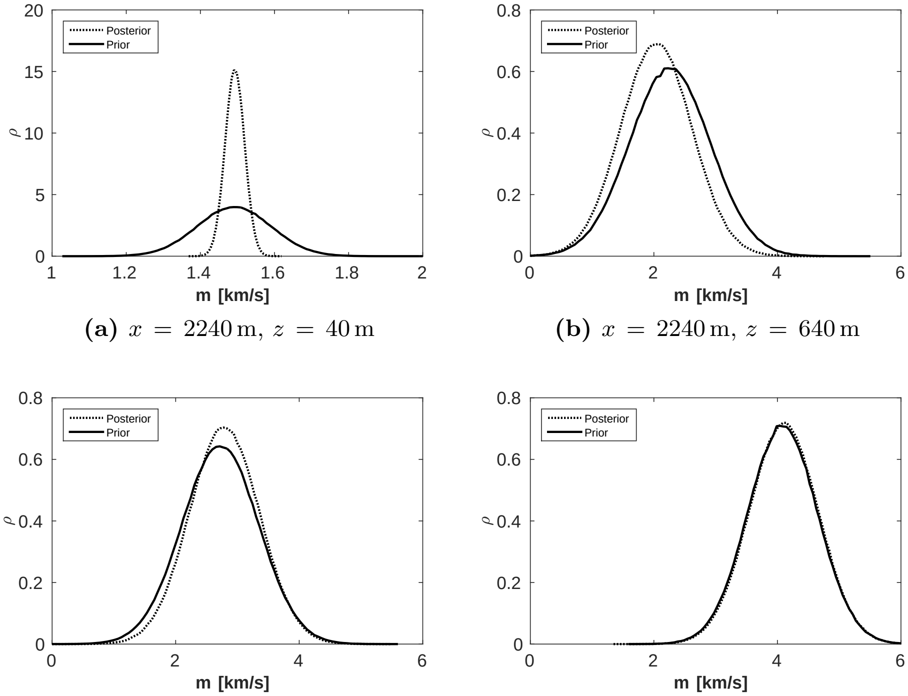
(a) $x = 1000\,\mathrm{m}$ **(b)** $x = 2500\,\mathrm{m}$ **(c)** $x = 4000\,\mathrm{m}$

Figure 6.14: The mean and standard deviation comparisons of the posterior (blue) and the prior (red) distributions at $x = 1000\,\mathrm{m}$, $2500\,\mathrm{m}$, and $4000\,\mathrm{m}$.

To verify our statistical results, we also utilize the RML method as a baseline approach. For this example, drawing a single sample with the RML method using 200 l-BFGS iterations requires at least 1.09 million PDE solves, which is computationally daunting. As a result, we only generate 10 samples via the RML method and compare them to the 95% confidence intervals (i.e., the blue patch) obtained by the GARTO method in Figure 6.16. From these figures, it is clear that the majority of the 10 samples lie in the blue patch. Moreover, the variation of the ten samples also matches the 95% confidence interval. In this case, we conclude that the estimated confidence intervals are likely accurate approximations of the true confidence intervals.

6.5 Discussion

When the underlying model is given by PDE-constraints consisting of multiple experiments, one is forced to make various approximations and tradeoffs in order to develop a computationally tractable procedure. There are a large number of discretized parameters, corresponding to a discretized 2D or 3D function, and one must necessarily avoid having to explicitly construct dense covariance matrices of the size of the model grid squared, whose construction requires a large number of PDE solves. Moreover, each evaluation of



(a) $x\,=\,2240\,\mathrm{m}$, $z\,=\,40\,\mathrm{m}$

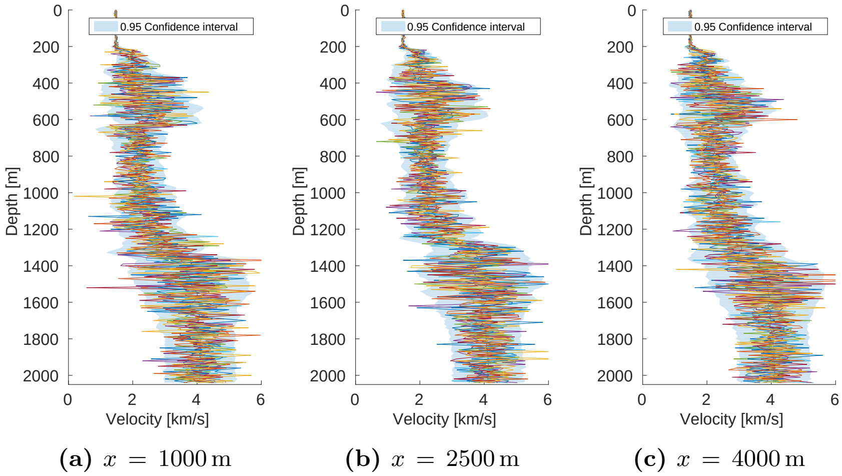
(b) $x\,=\,2240\,\mathrm{m}$, $z\,=\,640\,\mathrm{m}$

(c) $x\,=\,2240\,\mathrm{m}$, $z\,=\,1240\,\mathrm{m}$

(d) $x\,=\,2240\,\mathrm{m}$, $z\,=\,1840\,\mathrm{m}$

Figure 6.15: The comparison of the prior (solid line) and the posterior (dotted line) marginal distributions at the locations of $(x,\,z)\,=\,$(2240m, 40m), (2240m, 640m), (2240m, 1240m), and (2240m, 1840m).

the posterior distribution involves the solution of multiple PDEs, a computationally expensive affair. Methods that require tens or hundreds of thousands of such posterior evaluations, such as McMC-type methods, do not scale to realistically sized problems. The original PDE-constrained formulation of Bayesian inference, while theoretically convenient, results in a posterior that cannot be appropriately approximated by a Gaussian distribution, whereas the relaxation of the exact PDE-constraints results in a posterior that is much more amenable to Gaussian approximation. Ideally, one would like to parameterize the distribution over the joint model/field variables $(\mathbf{m}, \mathbf{u})$ and estimate the variances accordingly. Hidden in this notation, however, is the fact that in real-world problems, we have hundreds or potentially thousands of source experiments, each of which corresponds



(a) $x = 1000\,\mathrm{m}$ **(b)** $x = 2500\,\mathrm{m}$ **(c)** $x = 4000\,\mathrm{m}$

Figure 6.16: The 95% confidence intervals and the 10 realizations via the RML method at $x = 1000\,\mathrm{m}$, $2500\,\mathrm{m}$, and $4000\,\mathrm{m}$.

to a different **u**. Storing all of these fields and updating them explicitly becomes prohibitive memory-wise, even on a large cluster. As a result, our approach aims to keep the benefits of relaxing the PDE-constraints while still resulting in a computationally feasible algorithm.

The initial results illustrate that with the specific selection of λ, i.e., $\lambda^2 = 0.01\mu_1$, the relaxed formulation of the posterior PDF is less prone to local maxima, which enables us to analyze it via an arguably accurate Gaussian approximation. Once we can manipulate the covariance matrix of the Gaussian approximation through the implicit PDE-free formulation, we have access to a variety of statistical quantities, including the pointwise standard deviation, the confidence intervals, and the pointwise marginal distributions, in a tractable manner. We can use these quantities to assess the uncertainties of our inversion results and to identify the areas with high/low reliability in the model. This information is important and useful for the subsequent processing and interpretation. A straightforward example is that it allows us to assess the reliability of the visible image features obtained by the subsequent procedure of imaging, as in Ely et al. (2017).

While the initial results are promising, some aspects of the proposed framework warrant further investigation. Although numerical examples illustrate the feasibility of the proposed method for the case with the selection of $\lambda^2 = 0.01\mu_1$, it does not guarantee the feasibility of the proposed approach for PDFs arising from other choices of λ. For other selections of λ, the posterior PDFs can significantly differ from a Gaussian PDF, which makes approximately sampling them challenging. Potentially other sampling schemes can be explored for these distributions.

While we have shown the feasibility of the proposed sampling method for 2D problems, the application of the proposed sampling method to 3D problems is still challenging from the perspective of memory usage. To satisfy the $\mathcal{O}\big(n_{\text{grid}} \times (n_{\text{rcv}} + n_{\text{src}}) \times n_{\text{freq}}\big)$ storage requirement for formulating the implicit Gauss-Newton Hessian operator, a large high-performance cluster with enough computational workers and memory is needed to store all of the necessary matrices in parallel.

6.6 Conclusions

We have described a new Bayesian framework for partial-differential-equation-constrained inverse problems. In our new framework, we introduce the field variables as auxiliary variables and relax the partial-differential-equation-constraint. By integrating out the field parameters, we avoid having to store and update the high number of field parameters, and exploit the fact that the new distribution is Gaussian in the field variables for every fixed model. We propose a method for selecting an appropriate penalty parameter such that the new posterior distribution can be approximated by a Gaussian distribution, which is more accurate than the conventional formulation.

We apply the new formulation to the seismic inverse problems and derive each component of the general framework. For this specific application, we use a partial-differential-equation-free Gauss-Newton Hessian to formulate the Gaussian approximation of the posterior distribution. We also illustrate that with this Gauss-Newton Hessian, the Gaussian approximation can be sampled without the requirement of the explicit formulation of the Gauss-Newton Hessian and its Cholesky factorization.

Our proposed method compares favorably to the existing randomized maximum likelihood method for generating different statistics on a simple layered model and the more challenging BG Compass model. Compared to the randomized maximum likelihood method, our method produces results that are quite visually similar, while requiring significantly fewer partial-differential-equation solves, which are the computational bottleneck in these problems. We expect that these methods will scale reasonably well to 3D models, where traditional methods have only begun to scratch the surface of the problem.

While in this chapter we utilize the Gaussian smoothness prior distribution, indeed, the proposed sampling method can also handle other choices of prior distributions, as long as they have sparse covariance matrices. In the future, we can investigate the incorporation of the proposed method with other kinds of prior distributions.

Chapter 7

Conclusion

Since oil and gas industries rely on images of the subsurface structure to evaluate the location, size, and profitability of oil and gas reservoirs, as well as a variety of exploration and economic risks, research into producing high-resolution images of complexly structured areas is exceptionally significant. In turn, this thesis has contributed source estimation techniques for seismic inversion and imaging problems, which can be applied to realistic cases, allowing for us to construct both background velocity models and high-resolution details without prior knowledge of the source functions. In addition, we have also contributed a computationally efficient framework to analyze uncertainties in the recovered velocity model, which can be used to assess risks in the subsequent procedures. In order to do this, we have addressed the topics of (1) source estimation for time-domain sparsity promoting least-squares reverse-time migration, (2) wavefield-reconstruction inversion with source estimation, and (3) uncertainty quantification for weak wave-equation constrained inverse problems.

7.1 Source estimation for time-domain sparsity promoting least-squares reverse-time migration

In the first part of the thesis (chapter 2), we proposed an on-the-fly source estimation technique for the time-domain least-squares reverse-time migration. The linearized forward operator is linear with respect to the source function. Based on this fact, we proposed to project out the source function by solving a regularized linear data-fitting optimization problem during each iteration of LS-RTM. The resulting source function can minimize the difference between the predicted and observed data given the current model update. Through iteratively updating the source function in this way along with the update of the model parameters, the time-domain least-squares reverse-time migration is able to reach an image close to the one using the accurate source function. As a result, we mitigate the dependence of the time-domain least-squares reverse-time migration on the accurate source function.

We embedded the proposed source estimation technique in the robust time-domain sparsity promoting least-squares reverse-time migration. This approach combines the stochastic optimization technique and a curvelet-based sparsity-promoting inversion algorithm. As a result, it can produce high-resolution and artifact-free images at the cost of approximately one reverse time migration using all the sources. We illustrated that the sparsity-promoting inverse problem can be solved by an easy-to-implement approach—linearized Bregman method, which allows us to straightforwardly embed the proposed source estimation technique in each iteration.

Finally, we provided numerical examples illustrating that the proposed algorithm yields results that are comparable to those produced with the true source functions. As a result, we have contributed an approach that provides high-resolution subsurface images but without the need for obtaining correct source functions, and is, therefore, more applicable to real-world problems.

7.2 Source estimation for wavefield-reconstruction inversion

In the second part of the thesis (chapters 3, 4, and 5), we proposed a computationally efficient on-the-fly source estimation technique in the context of wavefield-reconstruction inversion, which enhances the realistic application of wavefield-reconstruction inversion. To achieve this task, we derived an optimization problem with respect to the velocity model, wavefields, and source functions. We illustrated that this optimization problem can be solved by the variable projection method, which simultaneously projects out the wavefields and source functions during each iteration. After the projection, we obtain a reduced objective function that only depends on the velocity. We illustrated that minimizing this reduced objective function can produce an inverted velocity model that is close to the one using the correct source functions.

During the procedure of source estimation, the projection of the wavefields and source functions involves inverting a source-dependent linear system, which requires an independent Cholesky factorization for each source. To reduce the computational cost, we applied the block matrix inversion formula and arrived at a new inversion scheme that only requires inverting a source-independent matrix. Therefore, all sources share the same Cholesky factorization and the computational cost is significantly reduced.

We demonstrated the effectiveness of the proposed source estimation technique by a series of numerical experiments conducted with the BG compass model dataset. From these experiments, we observed that the proposed approach is stable with respect to the modeling errors and measurement noise. Additionally, we demonstrated the feasibility of the proposed approach to practical problems by applying it to the Chevron 2014 benchmark blind test dataset. Although the data contain elastic events and our modeling kernel is based on the acoustic wave equation, the proposed approach is still able to recover both the source functions and velocity model. In sum, the proposed source estimation technique can enable us to conduct wavefield-reconstruction inversion without prior information about the source functions.

7.3 Uncertainty quantification for weak wave-equation constrained inverse problems

In the last portion of the thesis (chapter 6), we presented an uncertainty quantification framework based on weak wave-equation constraints. The conventional approach strictly satisfies the wave-equation constraint by solving the wave equation. Different from the conventional approach, we relaxed the wave-equation constraints by allowing for Gaussian deviations in the wave equations. We discovered that with a carefully selected variance choice for the wave-equation misfit, the relaxing strategy results in a posterior distribution containing fewer modes in comparison with the conventional statistical FWI framework. This allows us to appropriately approximate the new posterior distribution by a Gaussian distribution with a higher accuracy than the conventional one.

We also proposed an implicit and wave-equation-free formulation to construct the covariance matrix of the Gaussian distribution. This implicit formulation enables us to compute the product of the covariance matrix and any vectors with several efficient matrix-vector multiplications instead of solving many expensive wave equations. Based on this fact, We embedded the implicit formulation in the randomize-then-optimize method. The resulting approach can sample the Gaussian distribution without the requirement of the explicit formulation of the covariance matrix and its Cholesky factorization.

We favorably compared the proposed approach to the existing randomized maximum likelihood method for generating different statistical quantities on a simple layered model and the more challenging BG Compass model. Compared to the randomized maximum likelihood method, the proposed approach produces results that are quite visually similar, while requiring significantly fewer wave-equation solves, which are the main computational bottleneck in these problems.

7.4 Current limitations

Some limitations of the work presented in this thesis are as follows:

1. The current implementation of the projection procedure in wavefield-reconstruction inversion is based on the Cholesky factorization method, which is feasible for small or intermediate-scale 2D problems but unfeasible for large-scale 3D problems.

2. The Gaussian approximation for the new posterior distribution is only valid with the specific selection of the variance in the wave-equation misfit. For other selections of the variance, the posterior distribution can significantly differ from a Gaussian distribution, which makes approximately sampling them challenging.

3. The source estimation procedure in the proposed time-domain sparsity-promoting LS-RTM requires solving a deconvolution problem. Due to the large volume of data in 3D problems, this deconvolution problem can be computationally expensive to solve for 3D problems.

7.5 Future research directions

Some ideas for future work are as follows:

1. Develop a computationally fast and memory efficient implementation for 3D wavefield-reconstruction inversion with source estimation. In 3D cases, it is computationally unfeasible to simultaneously project out the wavefields and source functions by solving the data-augmented system with direct solvers. Therefore, an efficient iterative solver with an appropriate preconditioner is necessary.

2. Develop a computationally efficient second-order method for wavefield-reconstruction inversion. We can also apply the implicit Gauss-Newton Hessian proposed in the chapter of uncertainty quantification to the deterministic optimization problem of wavefield-reconstruction inversion. Because the Hessian-vector product only involves several simple matrix-vector multiplications, we can use iterative solves including PCG and LSQR to invert the Gauss-Newton Hessian without solving additional wave equations. As a result, we can derive a computationally efficient second-order method by using the implicit Gauss-Newton Hessian.

3. Incorporate non-Gaussian prior distribution with the proposed uncertainty quantification framework with weak wave-equation constraints. Gaussian prior models are used in this thesis since they provide mathematically simple and computationally efficient formulations of important inverse problems. Unfortunately, the Gaussian prior fails to capture a range of important properties including sparsity and natural constraints such as positivity. This fact motivates us to incorporate non-Gaussian priors with the current framework.

4. The strategy of relaxing the wave-equation provides us with the access to separate the noise into the receiver part (measurement noise) and source part (background noise). Indeed, the conditional distribution $\rho(\mathbf{u}|\mathbf{m})$ can be understood as the distribution associated with the background noise. We can easily replace the penalty parameter by a weighting matrix to describe the statistical properties of the background noise. Therefore, we can use the proposed formulation to study the influence of the measurement and background noise on the uncertainties of the inverted velocity model.

5. Develop a 3D implementation for the sparsity-promoting least-squares reverse-time migration with source estimation. Recent computational improvements allow us to simulate wavefields in a 3D manner and undertake the challenge of executing least-squares reverse-time migration. Therefore, an extension of the current work to 3D problems would be beneficial and feasible.

Bibliography

Akcelik, V., 2002, Multiscale Newton-Krylov methods for inverse acoustic wave propagation: PhD thesis, Carnegie Mellon University. → pages 52

Amestoy, P., R. Brossier, A. Buttari, J.-Y. L'Excellent, T. Mary, L. Métivier, A. Miniussi, and S. Operto, 2016, Fast 3D frequency-domain full-waveform inversion with a parallel block low-rank multifrontal direct solver: Application to OBC data from the North Sea: Geophysics, **81(6)**, R363–R383. → pages 74

Aoki, N., and G. T. Schuster, 2009, Fast least-squares migration with a deblurring filter: Geophysics, **74**, WCA83–WCA93. → pages 2, 7

Aravkin, A. Y., and T. van Leeuwen, 2012, Estimating nuisance parameters in inverse problems: Inverse Problems, **28**, 115016. → pages 46

Askan, A., V. Akcelik, J. Bielak, and O. Ghattas, 2007, Full waveform inversion for seismic velocity and anelastic losses in heterogeneous structures: Bulletin of the Seismological Society of America, **97**, 1990–2008. → pages 52

Aster, R. C., B. Borchers, and C. H. Thurber, 2011, Parameter estimation and inverse problems: Academic Press, **90**. → pages 94

Bardsley, J. M., A. Seppänen, A. Solonen, H. Haario, and J. Kaipio, 2015, Randomize-then-optimize for sampling and uncertainty quantification in electrical impedance tomography: SIAM/ASA Journal on Uncertainty Quantification, **3**, 1136–1158. → pages 94, 98, 109, 110, 115, 118, 124, 125

Bardsley, J. M., A. Solonen, H. Haario, and M. Laine, 2014, Randomize-then-optimize: A method for sampling from posterior distributions in nonlinear inverse problems: SIAM Journal on Scientific Computing, **36**, A1895–A1910. → pages 95, 115

Baumstein, A., 2013, POCS-based geophysical constraints in multi-parameter full wavefield inversion: Presented at the 75th EAGE Conference and Exhibition 2013, EAGE. → pages 81, 88

Bauschke, H., and J. Borwein, 1994, Dykstras alternating projection algorithm for two sets: Journal of Approximation Theory, **79**, 418 – 443. → pages 83

Baysal, E., D. D. Kosloff, and J. W. Sherwood, 1983, Reverse time migration: Geophysics, **48**, 1514–1524. → pages 2, 7

Bernstein, D. S., 2005, Matrix mathematics: Theory, facts, and formulas with application to linear systems theory: Princeton University Press Princeton. → pages 54

Bertsekas, D. P., and J. N. Tsitsiklis, 1995, Neuro-dynamic programming: an overview: Decision and Control, 1995., Proceedings of the 34th IEEE Conference on, IEEE, 560–564. → pages 11

Biegler, L. T., T. F. Coleman, A. Conn, and F. N. Santosa, 2012, Large-scale optimization with applications: Part I: Optimization in inverse problems and design: Springer Science & Business Media, **92**. → pages 93

Biondi, B., 2001, Kirchhoff imaging beyond aliasing: Geophysics, **66**, 654–666. → pages 7

Brenders, A. J., and R. G. Pratt, 2007, Full waveform tomography for lithospheric imaging: results from a blind test in a realistic crustal model: Geophysical Journal International, **168**, 133–151. → pages 81, 83, 90

Brossier, R., S. Operto, and J. Virieux, 2009, Seismic imaging of complex onshore structures by 2D elastic frequency-domain full-waveform inversion: Geophysics, **74(6)**, WCC105–WCC118. → pages 52

Bui-Thanh, T., O. Ghattas, J. Martin, and G. Stadler, 2013, A computational framework for infinite-dimensional Bayesian inverse problems part I: The linearized case, with application to global seismic inversion: SIAM Journal on Scientific Computing, **35**, A2494–A2523. → pages 16, 94, 98, 106, 109

Bunks, C., F. M. Saleck, S. Zaleski, and G. Chavent, 1995, Multiscale seismic waveform inversion: Geophysics, **60**, 1457–1473. → pages 12, 37

Cai, J.-F., S. Osher, and Z. Shen, 2009, Convergence of the linearized Bregman iteration for L1-norm minimization: Mathematics of Computation, **78**, 2127–2136. → pages 24

Caldwell, J., and C. Walker, 2011, An overview of marine seismic operations: Technical report 448: Technical report. → pages xi, 1, 3

Candes, E., L. Demanet, D. Donoho, and L. Ying, 2006, Fast discrete curvelet transforms: Multiscale Modeling & Simulation, **5**, 861–899. → pages 22

Candès, E. J., et al., 2006, Compressive sampling: Proceedings of the international congress of mathematicians, Madrid, Spain, 1433–1452. →

pages 11

Candès, E. J., and M. B. Wakin, 2008, An introduction to compressive sampling: IEEE signal processing magazine, **25**, 21–30. → pages 11

Chen, S. S., D. L. Donoho, and M. A. Saunders, 2001, Atomic decomposition by basis pursuit: SIAM review, **43**, 129–159. → pages 22, 23

Chen, Y., and D. S. Oliver, 2012, Ensemble randomized maximum likelihood method as an iterative ensemble smoother: Mathematical Geosciences, **44**, 1–26. → pages 95, 115, 116

Chen, Z., D. Cheng, W. Feng, and T. Wu, 2013, An optimal 9-point finite difference scheme for the Helmholtz equation with PML: International Journal of Numerical Analysis and Modeling, **10**, 389–410. → pages 57, 109

Cohen, J. K., and N. Bleistein, 1979, Velocity inversion procedure for acoustic waves: Geophysics, **44**, 1077–1087. → pages 2

Cordua, K. S., T. M. Hansen, and K. Mosegaard, 2012, Monte Carlo full-waveform inversion of crosshole GPR data using multiple-point geostatistical a priori information: Geophysics, **77**, H19–H31. → pages 16

Donoho, D. L., 2006, Compressed sensing: IEEE Transactions on information theory, **52**, 1289–1306. → pages 11, 22

Duijndam, A., 1988, Bayesian estimation in seismic inversion. part ii: Uncertainty analysis1: Geophysical Prospecting, **36**, 899–918. → pages 14

Dummit, D. S., and R. M. Foote, 2004, Abstract algebra: Wiley Hoboken, **3**. → pages 100

Eckstein, J., and W. Yao, 2012, Augmented Lagrangian and alternating direction methods for convex optimization: A tutorial and some illustrative computational results: RUTCOR Research Reports, **32**. → pages 85

Efron, B., 1981, Nonparametric estimates of standard error: the jackknife, the bootstrap and other methods: Biometrika, **68**, 589–599. → pages 95, 114, 115

———, 1992, Bootstrap methods: another look at the jackknife, *in* Breakthroughs in statistics: Springer, 569–593. → pages 95, 114, 115

Ely, G., A. Malcolm, and O. V. Poliannikov, 2017, Assessing uncertainties in velocity models and images with a fast nonlinear uncertainty quantification method: Geophysics, **83(2)**, R63–R75. → pages 14, 16, 94, 132

Epanomeritakis, I., V. Akçelik, O. Ghattas, and J. Bielak, 2008, A

Newton-CG method for large-scale three-dimensional elastic full-waveform seismic inversion: Inverse Problems, **24**, 034015. → pages 93

Esser, E., L. Guasch, T. van Leeuwen, A. Y. Aravkin, and F. J. Herrmann, 2018, Total-variation regularization strategies in full-waveform inversion: SIAM Journal on Imaging Sciences, **11**, 376–406. → pages 78, 101

Etgen, J., S. H. Gray, and Y. Zhang, 2009, An overview of depth imaging in exploration geophysics: Geophysics, **74**, WCA5–WCA17. → pages 1, 2, 7

Fang, Z., C. Lee, C. Da Silva, F. J. Herrmann, and R. Kuske, 2015, Uncertainty quantification for wavefield reconstruction inversion: Presented at the 77th EAGE Conference and Exhibition 2015, EAGE. → pages 95

Fang, Z., C. Y. Lee, C. Da Silva, F. J. Herrmann, and T. Van Leeuwen, 2016, Uncertainty quantification for wavefield reconstruction inversion using a PDE-free semidefinite Hessian and randomize-then-optimize method: 86th Annual International Meeting, SEG, Expanded Abstracts, 1390–1394. → pages 95

Fisher, M., M. Leutbecher, and G. Kelly, 2005, On the equivalence between Kalman smoothing and weak-constraint four-dimensional variational data assimilation: Quarterly Journal of the Royal Meteorological Society, **131**, 3235–3246. → pages 95

French, W. S., 1975, Computer migration of oblique seismic reflection profiles: Geophysics, **40**, 961–980. → pages 7

Golub, G., and V. Pereyra, 2003, Separable nonlinear least squares: the variable projection method and its applications: Inverse Problems, **19**, R1. → pages 14, 40, 46, 99

Golub, G. H., and C. F. Van Loan, 2012, Matrix computations: Johns Hopkins University Press, **3**. → pages 112

Haario, H., M. Laine, A. Mira, and E. Saksman, 2006, Dram: efficient adaptive MCMC: Statistics and Computing, **16**, 339–354. → pages 94

Haber, E., U. M. Ascher, and D. Oldenburg, 2000, On optimization techniques for solving nonlinear inverse problems: Inverse problems, **16**, 1263. → pages 93

Haber, E., M. Chung, and F. Herrmann, 2012, An effective method for parameter estimation with PDE constraints with multiple right-hand sides: SIAM Journal on Optimization, **22**, 739–757. → pages 11

Herrmann, F. J., C. R. Brown, Y. A. Erlangga, and P. P. Moghaddam, 2009, Curvelet-based migration preconditioning and scaling: Geophysics, **74**, A41–A46. → pages 11

Herrmann, F. J., and X. Li, 2012, Efficient least-squares imaging with sparsity promotion and compressive sensing: Geophysical prospecting, **60**, 696–712. → pages 2, 7, 11, 22, 24

Herrmann, F. J., P. P. Moghaddam, and C. Stolk, 2008, Sparsity- and continuity-promoting seismic image recovery with curvelet frames: Applied and Computational Harmonic Analysis, **24**, 150–173. → pages 11, 22, 29

Herrmann, F. J., N. Tu, and E. Esser, 2015, Fast ”online” migration with Compressive Sensing: Presented at the 77th EAGE Conference and Exhibition 2015, EAGE. → pages 22, 24

Hinze, M., R. Pinnau, M. Ulbrich, and S. Ulbrich, 2008, Optimization with PDE constraints: Springer Science & Business Media, **23**. → pages 12, 94

Huang, G., R. Nammour, and W. Symes, 2017, Full-waveform inversion via source-receiver extension: Geophysics, **82(3)**, R153–R171. → pages 12

Jaiswal, P., C. A. Zelt, A. W. Bally, and R. Dasgupta, 2008, 2-D traveltime and waveform inversion for improved seismic imaging: Naga Thrust and Fold Belt, India: Geophysical Journal International, **173**, 642–658. → pages 2

Kaipio, J., and E. Somersalo, 2006, Statistical and computational inverse problems: Springer Science & Business Media. → pages 14, 16, 97, 98

Kennett, B. L. N., M. S. Sambridge, and P. R. Williamson, 1988, Subspace methods for large inverse problems with multiple parameter classes: Geophysical Journal International, **94**, 237–247. → pages 50

Kitanidis, P. K., 1995, Recent advances in geostatistical inference on hydrogeological variables: Reviews of Geophysics, **33**, 1103–1109. → pages 114

Lailly, P., et al., 1983, The seismic inverse problem as a sequence of before stack migrations: Presented at the Conference on Inverse Scattering. → pages 2, 11, 37

Li, M., J. Rickett, and A. Abubakar, 2013, Application of the variable projection scheme for frequency-domain full-waveform inversion: Geophysics, **78(6)**, R249–R257. → pages 14, 46

Li, Q.-Z., 2017, High-resolution seismic exploration: Society of Exploration Geophysicists. → pages 2

Li, X., A. Y. Aravkin, T. van Leeuwen, and F. J. Herrmann, 2012, Fast randomized full-waveform inversion with compressive sensing: Geophysics, **77(3)**, A13–A17. → pages 2, 24, 57, 122

Lieberman, C., K. Willcox, and O. Ghattas, 2010, Parameter and state model reduction for large-scale statistical inverse problems: SIAM

Journal on Scientific Computing, **32**, 2523–2542. → pages 109

Lim, S., 2017, Bayesian inverse problems and seismic inversion: PhD thesis, University of Oxford. → pages 95

Liu, J., A. Abubakar, T. Habashy, D. Alumbaugh, E. Nichols, and G. Gao, 2008, Nonlinear inversion approaches for cross-well electromagnetic data collected in cased-wells: 78th Annual International Meeting, SEG, Expanded Abstracts, 304–308. → pages 45, 46

Lorenz, D. A., F. Schopfer, and S. Wenger, 2014, The linearized bregman method via split feasibility problems: Analysis and generalizations: SIAM Journal on Imaging Sciences, **7**, 1237–1262. → pages 22, 24

Martin, J., L. C. Wilcox, C. Burstedde, and O. Ghattas, 2012, A stochastic Newton MCMC method for large-scale statistical inverse problems with application to seismic inversion: SIAM Journal on Scientific Computing, **34**, A1460–A1487. → pages 16, 94, 109, 110

Matheron, G., 2012, Estimating and choosing: an essay on probability in practice: Springer Science & Business Media. → pages 16, 97, 109

McMechan, G. A., 1983, Migration by extrapolation of time-dependent boundary values: Geophysical Prospecting, **31**, 413–420. → pages 2

Nemeth, T., C. Wu, and G. T. Schuster, 1999, Least-squares migration of incomplete reflection data: Geophysics, **64**, 208–221. → pages 11

Nemirovski, A., A. Juditsky, G. Lan, and A. Shapiro, 2009, Robust stochastic approximation approach to stochastic programming: SIAM Journal on optimization, **19**, 1574–1609. → pages 11, 24

Nocedal, J., and S. J. Wright, 2000, Numerical optimization: Springer. → pages 90

———, 2006, Numerical optimization: Springer-Verlag New York. → pages 40, 47, 52, 111

Orieux, F., O. Féron, and J.-F. Giovannelli, 2012, Sampling high-dimensional Gaussian distributions for general linear inverse problems: Signal Processing Letters, IEEE, **19**, 251–254. → pages 114

Osypov, K., N. Ivanova, Y. Yang, A. Fournier, D. Nichols, R. Bachrach, C. E. Yarman, and D. Nikolenko, 2013a, Uncertainty as a Rosetta Stone Uniting Geoscience Knowledge and E&P Business Value: Presented at the IPTC 2013: International Petroleum Technology Conference. → pages 7

Osypov, K., D. Nichols, M. Woodward, O. Zdraveva, C. Yarman, et al., 2008, Uncertainty and resolution analysis for anisotropic tomography using iterative eigendecomposition: Presented at the 2008 SEG Annual Meeting, Society of Exploration Geophysicists. → pages 14

Osypov, K., Y. Yang, A. Fournier, N. Ivanova, R. Bachrach, C. E. Yarman,

Y. You, D. Nichols, and M. Woodward, 2013b, Model-uncertainty quantification in seismic tomography: method and applications: Geophysical Prospecting, **61**, 1114–1134. → pages 2, 14, 16, 94, 98

Paffenholz, J., J. Stefani, B. McLain, and K. Bishop, 2002, SIGSBEE_2A synthetic subsalt dataset-image quality as function of migration algorithm and velocity model error: Presented at the 64th EAGE Conference and Exhibition 2002, EAGE. → pages 23

Papandreou, G., and A. L. Yuille, 2010, Gaussian sampling by local perturbations: Advances in Neural Information Processing Systems, 1858–1866. → pages 114

Peters, B., C. Greif, and F. J. Herrmann, 2015, An algorithm for solving least-squares problems with a Helmholtz block and multiple right-hand-sides: Presented at the International Conference On Preconditioning Techniques For Scientific And Industrial Applications. → pages 75

Peters, B., and F. J. Herrmann, 2016, Constraints versus penalties for edge-preserving full-waveform inversion: The Leading Edge, **36**, 94–100. → pages 57

Peters, B., F. J. Herrmann, and T. van Leeuwen, 2014, Wave-equation based inversion with the penalty method: adjoint-state versus wavefield-reconstruction inversion: Presented at the 76th EAGE Conference and Exhibition 2014, EAGE. → pages 12

Plessix, R.-E., 2006, A review of the adjoint-state method for computing the gradient of a functional with geophysical applications: Geophysical Journal International, **167**, 495–503. → pages 11, 45, 94

Pratt, R. G., 1999, Seismic waveform inversion in the frequency domain, Part 1: Theory and verification in a physical scale model: Geophysics, **64**, 888–901. → pages 2, 11, 12, 14, 37, 45, 46, 52, 57, 93, 108

Qin, B., T. Allemand, G. Lambaré, et al., 2015, Full waveform inversion using preserved amplitude reverse time migration: Presented at the 2015 SEG Annual Meeting, Society of Exploration Geophysicists. → pages 80

Rawlinson, N., A. M. Reading, and B. L. Kennett, 2006, Lithospheric structure of tasmania from a novel form of teleseismic tomography: Journal of Geophysical Research: Solid Earth, **111**. → pages 50

Roberts, G. O., J. S. Rosenthal, et al., 2001, Optimal scaling for various Metropolis-Hastings algorithms: Statistical Science, **16**, 351–367. → pages 16, 94

Rosasco, L., E. De Vito, A. Caponnetto, M. Piana, and A. Verri, 2004, Are loss functions all the same?: Neural Computation, **16**, 1063–1076. → pages 26

Rue, H., 2001, Fast sampling of Gaussian Markov random fields: Journal of the Royal Statistical Society: Series B (Statistical Methodology), **63**, 325–338. → pages 108

Sambridge, M., 1990, Non-linear arrival time inversion: constraining velocity anomalies by seeking smooth models in 3-d: Geophysical Journal International, **102**, 653–677. → pages 50

Sambridge, M., K. Gallagher, A. Jackson, and P. Rickwood, 2006, Trans-dimensional inverse problems, model comparison and the evidence: Geophysical Journal International, **167**, 528–542. → pages 14, 99

Scales, J. A., and L. Tenorio, 2001, Prior information and uncertainty in inverse problems: Geophysics, **66**, 389–397. → pages 14

Schmidt, M., D. Kim, and S. Sra, 2012, 11, *in* Projected Newton-type Methods in Machine Learning: MIT Press, **35**, 305–327. → pages 84

Schneider, W. A., 1978, Integral formulation for migration in two and three dimensions: Geophysics, **43**, 49–76. → pages 7

Sen, M., and I. Roy, 2003, Computation of differential seismograms and iteration adaptive regularization in prestack waveform inversion: Geophysics, **68**, 2026–2039. → pages 90

Shapiro, A., D. Dentcheva, and A. Ruszczyński, 2009, Lectures on stochastic programming: modeling and theory: SIAM. → pages 11, 24

Sheriff, R. E., and L. P. Geldart, 1995, Exploration seismology: Cambridge university press. → pages 1

Sirgue, L., O. Barkved, J. Dellinger, J. Etgen, U. Albertin, and J. Kommedal, 2010, Thematic set: Full waveform inversion: The next leap forward in imaging at valhall: First Break, **28**, 65–70. → pages 11, 37

Sirgue, L., and R. G. Pratt, 2004, Efficient waveform inversion and imaging: A strategy for selecting temporal frequencies: Geophysics, **69**, 231–248. → pages 12, 37

Solonen, A., A. Bibov, J. M. Bardsley, and H. Haario, 2014, Optimization-based sampling in ensemble Kalman filtering: International Journal for Uncertainty Quantification, **4**. → pages 114

Stuart, G., W. Yang, S. Minkoff, and F. Pereira, 2016, A two-stage Markov chain Monte Carlo method for velocity estimation and uncertainty quantification: 86th Annual International Meeting, SEG, Expanded Abstracts, 3682–3687. → pages 94

Sun, D., K. Jiao, D. Vigh, and R. Coates, 2014, Source wavelet estimation in full waveform inversion: Presented at the 76th EAGE Conference and Exhibition 2014, EAGE. → pages 12

Symes, W. W., D. Sun, and M. Enriquez, 2011, From modelling to inversion: designing a well-adapted simulator: Geophysical Prospecting,

59, 814–833. → pages 64

Tarantola, A., 1987, Inverse problem theory: Methods for data fitting and parameter estimation. → pages 94

———, 2005, Inverse problem theory and methods for model parameter estimation: siam, **89**. → pages 94

Tarantola, A., and B. Valette, 1982a, Generalized nonlinear inverse problems solved using the least squares criterion: Reviews of Geophysics, **20**, 219–232. → pages 2, 11, 37, 93

———, 1982b, Inverse problems = quest for information: Journal of Geophysics, **50**, 159–170. → pages 14, 94

Tu, N., A. Aravkin, T. van Leeuwen, T. Lin, and F. J. Herrmann, 2016, Source estimation with surface-related multiples—fast ambiguity-resolved seismic imaging: Geophysical Journal International, **205**, 1492–1511. → pages 10

Tu, N., A. Y. Aravkin, T. van Leeuwen, and F. J. Herrmann, 2013, Fast least-squares migration with multiples and source estimation: Presented at the 75th EAGE Conference and Exhibition 2013, EAGE. → pages 22, 24

Tu, N., and F. J. Herrmann, 2015a, Fast imaging with surface-related multiples by sparse inversion: Geophysical Journal International, **201**, 304–317. → pages 25

———, 2015b, Fast least-squares imaging with surface-related multiples: Application to a north sea data set: The Leading Edge, **34**, 788–794. → pages 2, 7, 11

van Leeuwen, T., A. Y. Aravkin, and F. J. Herrmann, 2011, Seismic waveform inversion by stochastic optimization: International Journal of Geophysics, **2011**. → pages 25

———, 2014, Comment on: "Application of the variable projection scheme for frequency-domain full-waveform inversion" (M. Li, J. Rickett, and A. Abubakar, Geophysics, 78, no. 6, R249–R257): Geophysics, **79(3)**, X11–X17. → pages 46, 121

van Leeuwen, T., and F. J. Herrmann, 2013a, Fast waveform inversion without source-encoding: Geophysical Prospecting, **61**, 10–19. → pages 2, 52, 57, 111, 122

———, 2013b, Mitigating local minima in full-waveform inversion by expanding the search space: Geophysical Journal International, **195**, 661–667. → pages 12, 37, 38, 39, 40, 41, 42, 85, 94, 98, 99

———, 2015, A penalty method for PDE-constrained optimization in inverse problems: Inverse Problems, **32**, 015007. → pages 12, 37, 38, 39, 40, 41, 52, 56, 58, 94, 95, 98, 99, 105, 111, 112

Vigh, D., K. Jiao, W. Huang, N. Moldoveanu, and J. Kapoor, 2013, Long-offset-aided full-waveform inversion: Presented at the 75th EAGE Conference and Exhibition 2013, EAGE. → pages 12

Vigh, D., K. Jiao, D. Watts, and D. Sun, 2014, Elastic full-waveform inversion application using multicomponent measurements of seismic data collection: Geophysics, **79**, R63–R77. → pages 11, 37

Virieux, J., and S. Operto, 2009, An overview of full-waveform inversion in exploration geophysics: Geophysics, **74(6)**, WCC1–WCC26. → pages 1, 2, 11, 14, 37, 39, 45, 70, 87, 93, 94, 128

Warner, M., and L. Guasch, 2014, Adaptive Waveform Inversion-FWI Without Cycle Skipping-Theory: Presented at the 76th EAGE Conference and Exhibition 2014, EAGE. → pages 12

Warner, M., T. Nangoo, N. Shah, A. Umpleby, J. Morgan, et al., 2013a, Full-waveform inversion of cycle-skipped seismic data by frequency down-shifting: 83th Annual International Meeting, SEG, Expanded Abstracts, 903–907. → pages 12, 39

Warner, M., A. Ratcliffe, T. Nangoo, J. Morgan, A. Umpleby, N. Shah, V. Vinje, I. Štekl, L. Guasch, C. Win, et al., 2013b, Anisotropic 3D full-waveform inversion: Geophysics, **78**, R59–R80. → pages 2, 11, 37

Whitmore, N., 1983, Iterative depth migration by backward time propagation, *in* SEG Technical Program Expanded Abstracts 1983: Society of Exploration Geophysicists, 382–385. → pages 2

Yilmaz, Ö., 2001, Seismic data analysis: Society of Exploration Geophysicists Tulsa, **1**. → pages 7

Ying, L., L. Demanet, and E. Candes, 2005, 3D discrete curvelet transform: Wavelets XI, International Society for Optics and Photonics, 591413. → pages 22

Zhou, B., and S. A. Greenhalgh, 2003, Crosshole seismic inversion with normalized full-waveform amplitude data: Geophysics, **68**, 1320–1330. → pages 12, 46

Zhu, H., S. Li, S. Fomel, G. Stadler, and O. Ghattas, 2016, A Bayesian approach to estimate uncertainty for full-waveform inversion using a priori information from depth migration: Geophysics, **81(5)**, R307–R323. → pages 16, 94, 106