Strategies for elastic full waveform inversion

by

Gian Matharu

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Geophysics

Department of Physics
University of Alberta

Abstract

The advent of modern supercomputers, in conjunction with larger, more comprehensive datasets, has led to a paradigm shift in seismic imaging. Full waveform inversion is routinely employed as a tool to estimate subsurface properties of the Earth with high resolution. The method fits simulated waveforms to observed data by iteratively updating estimates of subsurface properties. While recent advances have fostered seismic imaging success in areas with complex subsurface geology, a variety of challenges persist. Underdeveloped topics include the estimation of multiple physical parameters, uncertainty quantification, robust convergence, and the incorporation of more complex physics.

This thesis focuses on multi-parameter inversion in isotropic, elastic full waveform inversion. The transition from acoustic to elastic waveform inversion increases the computational cost, data complexity, and the ill-posed nature of the inverse problem. Estimating multiple independent subsurface parameters is challenging due to the limited, or overlapping, sensitivity of data to different parameters. In this thesis, I explore approaches to accelerate elastic full waveform inversion through simultaneous sources (Chapter 3) and second-order stochastic optimization (Chapter 4). Performance is assessed through controlled numerical experiments. Using the acoustic formulation, I present two forms of resolution/uncertainty analysis predicated on an approximation of the Hessian as a superposition of Kronecker products (Chapter 5). The final chapter compares applications of 2D acoustic and elastic full waveform inversion to a land dataset from the western Canadian basin (Chapter 6). I devise a workflow that includes data-preprocessing, initial model building and inversion.

Preface

A version of chapter 3 in this thesis has been published as a journal article: Matharu. G., and M. D. Sacchi, 2017, Source encoding in multiparameter full waveform inversion, Geophysical Journal International, Volume 214, Issue 2, Pages 792-810.

A version of chapter 4 of this thesis has been published as a journal article: Matharu. G., and M. D. Sacchi, 2019, A subsampled truncated-Newton method for multiparameter full-waveform inversion, Geophysics, 84, R333-R340.

Chapter 5 presents research done in collaboration with SAIG colleague Wenlei Gao. The work uses theory developed in a co-authored journal article: Gao. W., Matharu. G., and M. D. Sacchi, 2020, Fast least-squares reverse time migration via a superposition of Kronecker products, Geophysics, 85, S115-S134.

Chapter 6 contains a 2D seismic dataset that was provided to us by TGS (previously Arcis).

For the listed publications, I was the primary author while also developing and programming the modelling and inversion algorithms/examples. Dr. Sacchi served as the supervisory author, providing project guidance and manuscript editing. For the work in chapter 5, Wenlei Gao was the principal creator of the Kronecker-based Hessian factorization.

To my parents, who let me find my way.

Acknowledgements

I can only begin by expressing my gratitude for my supervisor, Professor Mauricio Sacchi. Throughout my PhD, he has given me the freedom to learn and develop as a researcher, allowing me to explore different avenues to satiate my curiosity. His expertise and uncanny intuition on complex problems are both invaluable and inspiring. I would like to thank my committee members, Prof. Bruce Sutherland, Prof. Richard Sydora, Prof. Xinwei Yu, and Prof. Mirko van der Baan for their support and suggestions over the years. I am deeply indebted to my many colleagues at the Signal Analysis and Imaging Group (SAIG), they have pushed me to grow not only academically, but personally. A special mention goes to my contemporaries Wenlei Gao and Fernanda Carozzi, who have undergone the trials and tribulations of this journey alongside me. I am grateful to NSERC and Alberta Innovates for providing scholarship funding during my studies. I am forever thankful to Dr. Jean Virieux, Dr. R.É. Plessix, Dr. Romain Brossier, Dr. Ludovic Métivier, and Dr. Andreas Fichtner for their expertise, courteousness, and willingness to entertain my questions at any time; their research continues to inspire me. Finally, I would be remiss if I did not mention the 5 pillars that form the foundations of my person: my parents, Gurmeet and Amrit; my sisters, Roopkamal and Gunwant; and my wife, Rajbir. This thesis is a product of your faith, patience, and unfaltering support.

Contents

Appendices

List of Tables

List of Figures

List of abbreviations and symbols

PDE	Partial differential equation
FWI	Full waveform inversion
SEFWI	Source-encoded FWI
SD	Steepest descent
CG	Linear conjugate gradients
NLCG	Non-linear conjugate gradients
TN	Truncated Newton method
STN	Subsampled truncated Newton method
SNR	Signal-to-noise ratio
WCSB	Western Canadian sedimentary basin

$\mathbf{d}$	Observed data
$\mathbf{u}$	Simulated data
$\mathbf{m}$	Model parameters
$J(\mathbf{m})$	Objective function
$\mathbf{m^0}$	Initial model
$\delta\mathbf{m}$	Model perturbation
N_s	Number of sources
N_r	Number of receivers
λ	Lamé parameter
μ	Shear modulus

ρ	Density
v_p or α	P-wave velocity
v_s or β	S-wave velocity
$\mathbf{g}$	FWI Gradient
$\mathbf{H}$	FWI Hessian
$\mathbf{L}(\mathbf{m})$	Wave equation operator
$\mathbf{L}^{\dagger}(\mathbf{m})$	Adjoint wave equation operator
$\mathcal{F}$	Fourier operator
$\mathbf{C_d}$	Prior data covariance
$\mathbf{C_m}$	Prior model covariance
$\tilde{\mathbf{C}}_\mathbf{m}$	Posterior covariance
$\rho(\mathbf{d})$	Data likelihood
$\rho(\mathbf{m})$	Prior model distribution
$\rho(\mathbf{m}\|\mathbf{d})$	Posterior distribution

CHAPTER 1

Introduction

Seismology is a scientific discipline that includes the study of the Earth's interior and seismic wave propagation. The canonical seismic experiment begins with a source that generates seismic waves that propagate into the Earth's subsurface. The downward propagating waves undergo refractions and reflections that are governed by subsurface material properties (e.g. rock densities, seismic velocities). These interactions regularly redirect waves towards the Earth's surface where they are recorded as ground motions across an array of receivers. Seismic sources are often categorized as "passive" or "active", with the former referring to naturally occurring sources e.g., earthquakes, whereas the latter is reserved for artificial man-made sources. Schematic representations of active and passive source seismic surveys are depicted in Figure 1.1.

Seismologists use time-series measurements of ground motion, known as seismograms, to make inferences about the structure of the Earth, geodynamics, source mechanisms, and a range of other topics. The variable scale of wave propagation leads to a natural division of seismology into the sub-disciplines of global and exploration seismology. Global seismology deals with wave propagation on a regional/continental/global scale (100-1000s km) relying on passive sources to generate data. Exploration (or reflection) seismology involves the use of active sources to survey the subsurface. It is the most prominently used technique for hydrocarbon exploration in the oil and gas industry.

Thus far, the description of the seismic experiment has been limited to data acquisition which marks one of its three distinct components. The remaining components are data processing and seismic imaging/inversion. The three components are often treated independently, but are inextricably linked as each component either caters to, or is reliant on another stage. This thesis focuses solely on seismic imaging and inversion in an exploration setting. Specifically, I address outstanding challenges in the seismic inversion technique known as full waveform

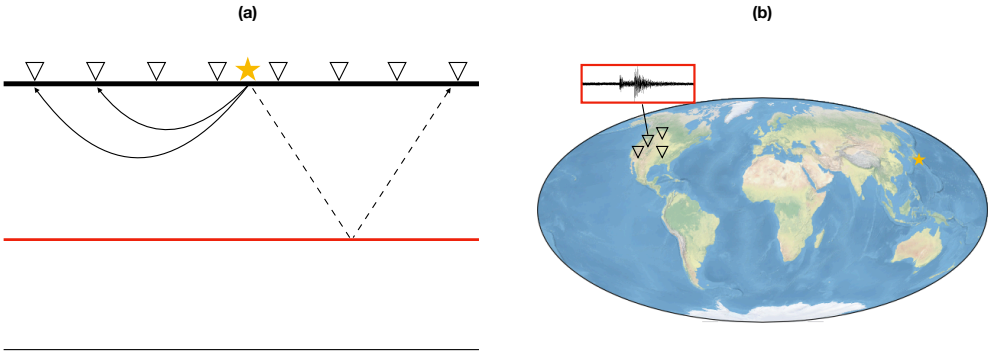

Figure 1.1: Schematic representation of a seismic survey for (a) exploration and (b) global seismology. Stars and inverted triangles represent seismic sources and receivers, respectively. Waves propagating through the Earth can undergo both refraction and reflection. Reflections commonly occur in the presence of strong material contrasts in the subsurface (illustrated with a red line in (a)).

inversion (FWI). FWI falls into a class of problems known as inverse problems. Inverse problems estimate model parameters given some observations and a mathematical model relating the model to the data. The remainder of the introduction provides a brief foray into forward and inverse problems before reviewing the current state of FWI research.

1.1 Forward problems

The forward problem is the mathematical formulation of a system whereby observations $\mathbf{d}$ are predicted from model parameters $\mathbf{m}$ through a mapping by operator $\mathbf{G}$. The operator $\mathbf{G}$ is derived from the governing equations of a system. The quantities $\mathbf{d}$ and $\mathbf{m}$ are abstract and can represent continuous or discrete variables. In the case of a discrete linear system, $\mathbf{d}$ and $\mathbf{m}$ would be vectors and $\mathbf{G}$ a matrix such that

$$\mathbf{m} \mapsto \mathbf{d} = \mathbf{G}\mathbf{m}. \tag{1.1}$$

There are a wide array of problems that can be classed as linear systems, linear regression being one such example. For linear regression, $\mathbf{m}$ represents a coefficient vector, $\mathbf{G}$ a matrix containing the input data samples, and $\mathbf{d}$ the observations.

For non-linear systems, $\mathbf{G}$ represents a non-linear operator with the forward problem defined as

$$\mathbf{m} \mapsto \mathbf{d} = \mathbf{G}(\mathbf{m}). \tag{1.2}$$

In FWI, data are modelled by solving the seismic wave equation, which represented abstractly, reads as

$$\mathbf{L}(\mathbf{m})\mathbf{u} = \mathbf{f}, \tag{1.3}$$

where $\mathbf{u}$ is the seismic wavefield excited by an external source $\mathbf{f}$. The linear differential operator $\mathbf{L}(\mathbf{m})$ characterizes the seismic wave equation and can accommodate varying degrees of physical complexity. The operator $\mathbf{L}(\mathbf{m})$ may represent the acoustic or elastic forms of the wave equation in 1, 2, or 3 dimensions without a loss of generality. The seismic wavefield $\mathbf{u}$ is linear with respect to the external source. The non-linearity in FWI stems from the non-linear relationship between $\mathbf{u}$ and $\mathbf{m}$. The mapping from model-to-data space can be expressed as $\mathbf{u} = \mathcal{P}\mathbf{L}(\mathbf{m})^{-1}\mathbf{f}$, where $\mathcal{P}$ is a restriction/sampling operator that samples the wavefield at select spatial positions.

1.2 Inverse problems

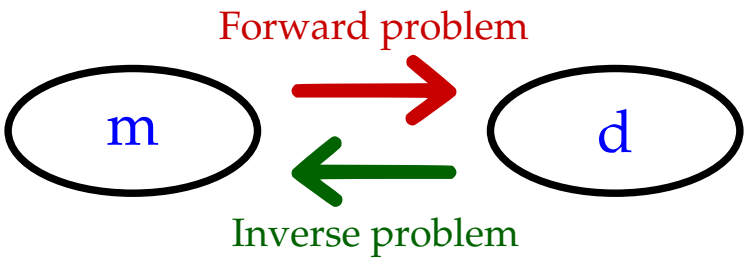

Figure 1.2: Illustration of the forward and inverse problem as mappings between data and model spaces.

Inverse problems estimate model parameters $\mathbf{m}$ given observations $\mathbf{d}$. Naively, the inverse problem might be considered as the inverse mapping of the forward problem, that is to say

$$\mathbf{d} \mapsto \mathbf{m} = \mathbf{G}^{-1}\mathbf{d}, \tag{1.4}$$

in the case of a linear problem. It is easy to see that Equation 1.4 is plagued with problems. If $\mathbf{G}$ is singular, its inverse does not exist. Assuming an inverse exists, in the presence of data errors $\mathbf{e}$,

$$\mathbf{m} = \mathbf{G}^{-1}(\mathbf{d} + \mathbf{e}), \tag{1.5}$$

in which case the inverse may be unstable if $\mathbf{G}$ is an ill-conditioned matrix. Furthermore, the inverse is only defined when $\mathbf{G}$ is a square matrix i.e. when the size of the $\mathbf{d}$ and $\mathbf{m}$ are the same, which is seldom true. Solutions to the inverse problem vary and depend on the properties of the system. For example, consider a linear system with $\mathbf{d} \in \mathbb{R}^n$, $\mathbf{m} \in \mathbb{R}^m$,

and $\mathbf{G} \in \mathbb{R}^{n \times m}$. Systems are defined as underdetermined ($n < m$, more unknowns than equations) or overdetermined ($n > m$, more equations than unknowns). Undetermined systems have an infinite number of solutions owing to the finite (non-zero) size of the system's null space; therefore, additional information is required to constrain candidate solutions. The minimum norm solution

$$\mathbf{m} = \mathbf{G}^T(\mathbf{G}\mathbf{G}^T)^{-1}\mathbf{d} \tag{1.6}$$

arises from minimizing $\mathbf{m}^T\mathbf{m}$ such that $\mathbf{G}\mathbf{m} = \mathbf{d}$. Overdetermined systems approximate solutions by minimizing the least-squares difference between the observations and the predictions,

$$\underset{\mathbf{m}}{\text{minimize}} \ \|\mathbf{G}\mathbf{m} - \mathbf{d}\|_2^2, \tag{1.7}$$

with the solution given by

$$\mathbf{m} = (\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T\mathbf{d}. \tag{1.8}$$

Both Equation 1.6 and Equation 1.8 impose requirements on the rank of $\mathbf{G}$. The linear problem has illustrated a number of difficulties that make the inverse problem non-trivial. The inverse problem for non-linear systems is no less problematic and is characteristically more challenging.

Non-linear inverse problems can be solved approximately using local optimization methods. For a discrete non-linear system of n equations, an objective function is defined as

$$\underset{\mathbf{m}}{\text{minimize}} \ J(\mathbf{m}) = \frac{1}{2}\|\mathbf{G}(\mathbf{m}) - \mathbf{d}\|_2^2, \tag{1.9}$$

where $\mathbf{d}$ and the predictions $\mathbf{G}(\mathbf{m})$ represent n-dimensional vectors. The objective function is at times referred to as the misfit functional or simply, the misfit. The principle behind local optimization is to make repeated small updates to an initial model $\mathbf{m}^0$ such that $J(\mathbf{m}) \rightarrow \min$. Written explicitly, the updates take the form

$$\mathbf{m}^{k+1} = \mathbf{m}^k + \nu^k \delta\mathbf{m}^k, \tag{1.10}$$

where k denotes the iteration number, ν^k is a scalar step length, and $\delta\mathbf{m}^k$ is a model perturbation/update. The model perturbation can be computed by minimizing the second-order Taylor expansion of the objective function around an initial model $\mathbf{m}^0$,

$$J(\mathbf{m}^0 + \delta\mathbf{m}) = J(\mathbf{m}^0) + \mathbf{g}^T\delta\mathbf{m} + \frac{1}{2}\delta\mathbf{m}^T\mathbf{H}\delta\mathbf{m} + \mathcal{O}(\delta\mathbf{m}^3) \tag{1.11}$$

where $\mathbf{g} = \nabla_m J(\mathbf{m}^0)$ and $\mathbf{H} = \nabla_m^2 J(\mathbf{m}^0)$ are the gradient and Hessian of $J(\mathbf{m}^0)$ (w.r.t. $\mathbf{m}$) written as a vector and matrix, respectively. After taking the derivative of Equation 1.11 with respect to $\mathbf{m}$ and setting it to zero, the minimizer of the local quadratic approximation

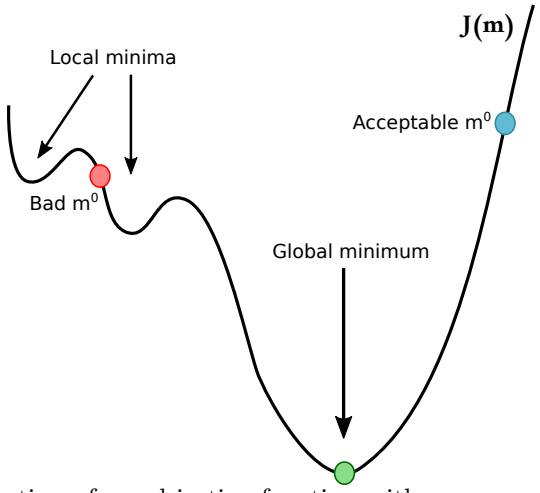

Figure 1.3: Illustration of an objective function with numerous local minima. A good starting point (initial model) is essential to ensuring proper convergence of gradient-based minimization algorithms. Poor initial models can lead to convergence to a local minimum that may not resemble the global minimum.

is given by the solution of the linear system

$$\mathbf{H}\delta\mathbf{m} = -\mathbf{g}. \tag{1.12}$$

The solution of the so called "Newton equation" (Equation 1.12) yields the Newton search direction. Explicitly computing and inverting the Hessian is computationally expensive and rarely done in practice. Economical alternatives for $\delta\mathbf{m}$ can be obtained from a range of gradient-based algorithms (Nocedal and Wright, 2006).

The challenges of non-linear inversion originate from the non-uniqueness and non-linearity of the system. The objective function in Equation 1.9 is non-convex due to the non-linear dependence of the predictions on $\mathbf{m}$. The non-convexity manifests as numerous local minima in the objective function (illustrated in Figure 1.3). Convergence towards the global minimum is therefore contingent on the starting point. The optimization may converge to a local minimum if $\mathbf{m}^0$ is far from the true model. Even if an acceptable minimum is found, it may not be unique (e.g., consider an objective function with an extended, flat global minimum). An ensemble of models may fit the data equally well. In the context of seismic imaging, some models may not be resolvable if, for example, they contain fine-scale structure that is below the resolution limit of the band-limited data. Furthermore, if seismic waves do not propagate through a particular region of the Earth, the data will exhibit no sensitivity to changes in that region. Limitations in the data and its acquisition introduce a null space into the problem that results in non-unique solutions. Until this point, $\mathbf{G}(\mathbf{m})$ has been

assumed to be an accurate mapping from the model to the data. Mathematical models of physical systems often involve simplifications or approximations to the true physics. Errors in the forward modelling restrict the class of solutions an inversion is able to recover.

Global optimization methods that probabilistically, or stochastically, search the model space could help navigate the numerous local minima in FWI. However, for typical 2D/3D problems the model space is simply too large to sample efficiently. This problem is compounded by the prohibitive cost of evaluating the objective function in FWI as it requires numerically solving PDEs. These two factors have precluded the use of global optimization methods in FWI for realistic problems. Despite issues with convergence, local optimization methods remain the workhorse for practical applications of FWI.

1.3 Full waveform inversion: A review

Full waveform inversion is a seismic inversion technique that estimates subsurface parameters (e.g. density, P- and S-wave velocities) by fitting simulated data to real-world observations. A critical component of FWI is that the simulated data represent numerical solutions to the seismic wave equation. By modelling the physics of the wave equation, complex wave phenomena can be modelled and utilized for inversion. Unlike other techniques, FWI advocates the fitting of complete waveforms/seismograms, where possible, as opposed to some compressed feature/s of the data (e.g., traveltimes of seismic phases). These two points mark a significant departure from traditional imaging techniques such as travel-time tomography. By accounting for a range of wave modes and the band-limited nature of the data, FWI offers significantly better resolving power than ray-based tomography. Despite this, the name "full waveform inversion" is a misnomer in some regards, as FWI seldom fits entire waveforms. In practice, FWI only attempts to fit wave modes that are accurately modelled by the physics model, or modes that provide desirable updates to the subsurface model.

At its conception in the 1980s full waveform inversion was a mathematical construct that lacked a definitive proof of concept. Lailly (1983) and Tarantola (1984b) present acoustic, time-domain formulations of FWI. Elastic extensions are later derived by Tarantola (1986) and Mora (1987). Subsequent 2D numerical studies characterize important properties and challenges of FWI. Gauthier et al. (1986) and Mora (1988) perform 2D acoustic and elastic inversion, respectively. Both studies demonstrate the scale separation achieved when using transmitted and reflected waves for inversion. Specifically, transmitted waves capture long wavelength components of the velocity model, whereas the inversion of reflection data yields high wavenumber models that resemble seismic images. Early attempts at applying FWI to real marine (Pica et al., 1990) and land (Crase et al., 1992) data resulted in limited success.

Initial studies were hindered by the relative immaturity of the algorithm, the non-linearity, and the prohibitive cost of the inverse problem.

Fundamental discoveries significantly advanced FWI in the 1990-2000s. Frequency-domain FWI formulates the FWI forward problem as a discrete linear system (Pratt and Worthington, 1990; Pratt et al., 1996; Pratt, 1999). Wave equation solutions are obtained following an LU decomposition of the forward modelling operator. The LU decomposition can be reused for different sources thereby reducing the computational cost relative to 2D time-domain implementations. The importance of wide-aperture surveys in recovering intermediate-to-long wavelength structure was established by Mora (1988); Pratt and Worthington (1990); Pratt et al. (1996). Multi-scale inversion schemes were introduced to mitigate the non-linearities of FWI (Bunks et al., 1995; Pratt et al., 1996); such schemes have become standard practice in modern FWI applications. Multi-scale schemes estimate large scale structure prior to incorporating fine scale details into the model. Typically, this involves fitting low-to-high frequencies in the data. Sirgue and Pratt (2004) propose a frequency progression scheme that exploits redundancies in the wavenumber domain. Pratt et al. (1998) investigates the role of the objective function's Hessian in FWI. Pratt (1999) validated FWI for a physical model, demonstrating a definitive improvement in the FWI inverted model over that recovered from traveltime tomography. Shortly thereafter, multiple applications of FWI on marine streamer data emerged (Shipp and Singh, 2002; Ravaut et al., 2004; Brenders and Pratt, 2007; Sears et al., 2008).

Simultaneously, FWI has been adapted in global tomography. The transition was stimulated by the discovery of "banana-doughnut" kernels in the study of finite-frequency sensitivity kernels (Marquering et al., 1999; Zhao et al., 2000; Dahlen et al., 2000). Finite-frequency kernels indicate that seismic traveltimes are sensitive to model perturbations in a frequency-dependent volume around, but not on, the geometric ray path. Tromp et al. (2005) unifies the idea of finite-frequency kernels with the principles of FWI derived by Tarantola (1984b). Full waveform tomography has been successfully applied for both regional and global tomography studies (e.g., Chen et al. (2007); Tape et al. (2009); Fichtner et al. (2009); French et al. (2013); Zhu et al. (2015)).

Over the past decade, interest in FWI has grown dramatically with increased industry adoption and general accessibility. The fundamental challenges currently plaguing FWI can be summarized as: computational cost, robust convergence, uncertainty quantification, utilizing reflections, multi-parameter inversion, and extensions to more complex physics. Numerous studies have explored reducing the computational burden of FWI through more advanced optimization schemes. Efforts include incorporating Hessian information to accelerate convergence (Brossier et al., 2009; Epanomeritakis et al., 2008; Métivier et al., 2013; Anagaw and Sacchi, 2014; Pan et al., 2016), the use of simultaneous sources (Romero et al.,

2000; Capdeville et al., 2005; Vigh and Starr, 2008; Krebs et al., 2009; Schuster et al., 2011; Anagaw and Sacchi, 2014; Castellanos et al., 2015), and stochastic optimization (Haber et al., 2012; van Leeuwen and Herrmann, 2013a).

Reflection data produce FWI gradients that contain high-wavenumber migration-like components and low-wavenumber, tomographic components. The migration-type terms dominate the update and are the reason why reflection data are often not included in FWI. Mora (1989) estimates low- and high-wavenumber components of a velocity model by utilizing both the migration and tomographic terms in the FWI gradient. Chavent et al. (1994) use a migration-based traveltime approach to invert for background slowness models from a poor initial guess. Xu et al. (2012) use the Born approximation to isolate the transmission, or tomographic, component of reflection ray paths to extract long-wavelength updates from reflection data. The long-wavelength information from reflection data can be used to provide deeper updates than conventional diving wave (i.e. waves that are redirected to the Earth's surface through refraction) FWI. Zhou et al. (2015) devise a joint inversion scheme that estimates background velocities and an impedance model in an alternating manner. The algorithm achieves scale separation by explicitly separating short-spread reflections and wide-angle transmitted waves in the data. The joint inversion uses both diving waves and reflections to better constrain shallow structure. Reflection waveform inversion requires an initial reflector/impedance model to generate reflections in the simulated data. The inherent ambiguity between velocity and reflector depth limits the vertical resolution of reflection waveform inversion (Gomes and Chazalnoel, 2017).

Robust convergence in conventional FWI is dictated by the quality of the initial model and data properties. Two classes of algorithm have been proposed to improve the tolerance of FWI to deficiencies in either of these factors; the classes are modified objective functions and extended inversion methods. Modified objective functions seek to improve the convexity of the optimization landscape i.e. to reduce the number of local minima. Objectives based on traveltimes/correlation (Luo and Schuster, 1991; Van Leeuwen and Mulder, 2010) and envelopes (Bozdag et al., 2011) benefit from simplifying the representation of the data. Approaches that maximize the delta-ness of matching filters that match synthetic and observed data have observed success in improving convergence (e.g., Luo and Sava (2011); Warner and Guasch (2016); Sun and Alkhalifah (2018)). Similarly, objectives based on optimal-transport distances are more robust against poor starting models (Métivier et al., 2016; Yang et al., 2018b; Yang and Engquist, 2018; Métivier et al., 2018). Extended inversion methods expand the model space in a non-physical manner (Symes, 2008; van Leeuwen and Herrmann, 2013b; Biondi and Almomin, 2014; Warner and Guasch, 2016). Extended models make it trivial to fit the data exactly; however, not all extended models are physically realizable. The optimization uses annihilators that gradually push the extended model to a physically realizable subset of the extended model space. While extended inversion methods

have demonstrated promising ability to navigate the non-convexity of FWI, they typically come with an increased computational cost.

Multi-parameter inversion is closely linked to the inclusion of more complex physics in FWI. In scenarios where the acoustic approximation no longer suitably approximates the data, additional physics such as attenuation, anisotropy, and elasticity should be considered. With each extension, the number of independent parameters required to parametrize the subsurface increases. For example, an acoustic medium can be characterized by P-wave velocity and density whereas an isotropic, elastic medium also requires S-wave velocity. Anisotropy can introduce up to 21 independent parameters in the most general description. Perturbations to different parameters can produce similar responses in the data. This ambiguity leads trade-offs between independent parameters. The difficulty inverting multiple parameters lies in the problem becoming more ill-posed. The dimensions of the model space increases while the number of data constraints generally remains fixed. In spite of the challenges, studies have achieved success in multi-parameter variants of FWI (Sears et al., 2008; Brossier et al., 2009; Plessix and Cao, 2011; Warner et al., 2013; Operto et al., 2013; Gholami et al., 2013; Prieux et al., 2013b; Alkhalifah and Plessix, 2014; Pan et al., 2016).

A neglected branch of FWI research is that of uncertainty quantification. Practitioners rely on heuristic quality control measures rather than rigorous uncertainty analysis. In recent years, some effort has been made in this department. In global tomography, spike tests that probe resolution/Hessian matrices have emerged as more reliable than classical checkerboard tests (Fichtner and Trampert, 2011b; Trampert et al., 2013; Rawlinson and Spakman, 2016; Fichtner and Leeuwen, 2015). While the framework for linearized Bayesian inversion has been around for some time (Tarantola, 2005), it has not been computationally feasible until recently (Bui-Thanh et al., 2013; Zhu et al., 2015). Fang et al. (2018) generalize Bayesian inversion to a penalty formulation of FWI known as wavefield reconstruction inversion. Ely et al. (2018) use a Metropolis-Hastings algorithm and a fast local solver to sample the posterior distribution. Thurin et al. (2019) use an ensemble data assimilation technique based on ensemble Kalman filters to quantify uncertainty in frequency-domain FWI.

1.4 Contributions of this thesis

The main contributions of this thesis are as follows:

- We explore the application of source encoding to multi-parameter FWI. Properties of the multi-parameter Hessian, with and without source encoding, are explored to establish the implications of encoding on parameter trade-off. We demonstrate potential limitations of encoding-based approaches for data-driven inversion schemes.

- We introduce a stochastic second-order optimization scheme for multi-parameter FWI. The inclusion of Hessian information into the optimization allows for improved per-iteration convergence rates and parameter decoupling. Relative to conventional second-order optimization schemes, the proposed method achieves comparable inversion results and convergence rates while reducing the per-iteration computational cost of FWI. The algorithm can utilize non-uniform sub-sampling of the data to further reduce cost.

- We investigate two computationally feasible forms of resolution analysis in FWI: a subsurface probing method and linearized Bayesian inversion. The analyses benefit from a compact representation of the Hessian in the form of a superposition of Kronecker products. The Hessian approximation permits fast Hessian-vector products that improve the efficiency and capability of both approaches. In addition to extracting uncertainty information associated with the inversion, we explore the unique properties of the Kronecker-based factorization of the Hessian.

- We develop an inversion workflow for elastic FWI applied to a land dataset. The workflow consists of data processing, initial model building and a tailored inversion strategy. We also perform acoustic FWI and evaluate the merits of each. We investigate the inclusion of reflection data through the use of a modified free-surface boundary condition.

1.5 Thesis overview

Chapter 2 formally introduces full waveform inversion as a PDE-constrained optimization. I define the forward problem as the isotropic, elastic wave equation and present the inverse problem as a gradient-based optimization. The adjoint-state method is used to derive expressions for the FWI gradient. I discuss challenges that arise when considering multi-parameter forms of FWI. I discuss features of a generic FWI algorithm and provide details of the software developed for this thesis.

Chapter 3 explores the application of source-encoding methods to multi-parameter FWI. The algorithm uses simultaneous sources to reduce the number of PDE solves required per FWI iteration, reducing the overall computational cost. This chapter examines the implications of crosstalk noise in the source-encoded, multi-parameter Hessian on parameter trade-off and inversion stability. We demonstrate that, when using source encoding, crosstalk noise does not cause deteriorated inversion results or divergence away from a reasonable solution. Parameter trade-off has similar characteristics to conventional FWI provided the inversion follows a similar workflow; we present examples of simultaneous and the sequential

inversion of multiple parameters to demonstrate this. We identify a limitation of source-encoded FWI for certain data-driven FWI schemes. A version of this chapter is published as a journal article (Matharu. G., and M. D. Sacchi, 2017, Source encoding in multi-parameter full waveform inversion, Geophysical Journal International, Volume 214, Issue 2, Pages 792-810).

Chapter 4 introduces a stochastic truncated-Newton method for multi-parameter inversion. The method uses a random subset of sources to compute approximate FWI gradients and Hessian-vector products. Model updates are estimated by iteratively solving an approximate Newton equation. The approximate Newton step reduces the number of PDE solves required per FWI iteration, while retaining advantageous properties of Truncated Newton methods (e.g., higher per-iteration convergence rates, improved resolution). We propose a non-uniform sampling scheme that preferentially selects sources based on a criteria that minimizes the error in the Hessian-vector product approximation. We evaluate the performance of the algorithm through synthetic inversions and comparisons with quasi-Newton and truncated Newton methods. A version of this chapter is published as a journal article (Matharu. G., and M. D. Sacchi, 2019, A subsampled truncated-Newton method for multi-parameter full-waveform inversion, Geophysics, 84, R333-R340).

Chapter 5 presents two techniques to assess resolution and uncertainty in FWI. A subsurface probing approach assumes convergence to a model in the vicinity of the global minimum. Horizontal and vertical resolution lengths are extracted by applying the Hessian to spike perturbations placed throughout the subsurface. A secondary analysis formulates a linearized Bayesian inversion that assumes Gaussian priors. The inverted model is presented as a conditional probability distribution. Confidence intervals are obtained using the diagonal of the posterior covariance. The proposed analyses are facilitated by approximating the Hessian as a superposition of Kronecker products. Important components of this chapter were developed in a preceding co-authored journal article (Gao. W., Matharu. G., and M. D. Sacchi, 2020, Fast least-squares reverse time migration via a superposition of Kronecker products, Geophysics, 85, S115-S134).

In **Chapter 6**, we perform a case study of FWI applied to a 2D land dataset from the western Canadian sedimentary basin. We conduct data-preprocessing and obtain initial velocity models using traveltime tomography combined with prior information from well constraints. We tailor an inversion workflow to ensure robust convergence given the properties of the data. Separate workflows are deployed for acoustic and elastic forms of FWI. We update near-surface P-wave velocity structure by focusing the inversion on fitting diving waves; the elastic inversion also estimates S-wave velocity despite limited sensitivity to the parameter. In the elastic inversion, we adopt a modified free-surface boundary condition and remove surface waves from the data to allow fitting of reflection events. We discuss the merits and

limitations of acoustic and elastic FWI. A version of this chapter is being prepared for a manuscript submission to the journal *Geophysics*.

CHAPTER 2

Multi-parameter full waveform inversion

2.1 Mathematical formulation: A PDE constrained optimization problem

Full waveform inversion estimates a set of subsurface material parameters that minimize the difference between observed and synthetic data; the latter are numerical solutions to partial differential equations (PDEs). Mathematically, FWI can be formulated as a PDE-constrained optimization problem of the form

$$\begin{aligned} \underset{\mathbf{m}}{\text{minimize}} \quad & J(\mathbf{m}), \\ \text{subject to} \quad & \mathbf{L}(\mathbf{m})\mathbf{u} - \mathbf{f} = 0, \end{aligned} \tag{2.1}$$

where $\mathbf{L}(\mathbf{m})$ is the wave-equation operator (Plessix, 2006). The misfit functional $J(\mathbf{m})$ — also known as the cost or objective function— quantifies the difference between observed and synthetic data $\mathbf{u}$ by comparing an observable quantity. The PDE constraint requires that the synthetic data be solutions to the seismic wave equation. The optimization is performed over an estimated model $\mathbf{m}$; $\mathbf{f}$ is an external source. The classical choice of objective function in FWI is the least-squares waveform misfit functional

$$J(\mathbf{m}) = \frac{1}{2} \sum_{s=1}^{N_s} \sum_{r=1}^{N_r} \int_T |\mathbf{u}_s(\mathbf{x}_r, t; \mathbf{m}) - \mathbf{d}_s(\mathbf{x}_r, t)|^2 \, \mathrm{d}t. \tag{2.2}$$

The simulated multi-component data $\mathbf{u}_s(\mathbf{x}_r, t; \mathbf{m})$ are recorded at the r-th receiver and generated by the s-th source $\mathbf{f}_s(\mathbf{x}, t)$ for model $\mathbf{m}$. An analogous definition holds for the observed data $\mathbf{d}_s(\mathbf{x}_r, t)$. The model parameters $\mathbf{m}(\mathbf{x}) = [m_1(\mathbf{x}), m_2(\mathbf{x}), \ldots, m_{N_p}(\mathbf{x})]^T$, represent N_p

independent physical properties of the Earth's subsurface; T denotes the transpose. The number of sources and receivers are denoted by N_s and N_r, respectively. For the sake of brevity, we henceforth omit the spatial and temporal dependencies of variables after they are first introduced, provided that no ambiguities arise from the omission. The least-squares waveform misfit (Equation 2.2) is a non-linear functional owing to the squared term and the non-linear dependence of $\mathbf{u}$ on $\mathbf{m}$ (Virieux and Operto, 2009).

Non-linear functionals can be minimized via iterative gradient-based minimization algorithms (Plessix, 2006). Lailly (1983) and Tarantola (1984b) demonstrated this by formulating FWI as a linearized inverse problem. Solutions to Equation 2.1 can be estimated by iteratively updating the model parameters via

$$\mathbf{m}^{k+1} = \mathbf{m}^k + \nu^k \delta\mathbf{m}^k, \tag{2.3}$$

where k denotes the iteration number, ν^k is a scalar step length, and the model perturbation/update is $\delta\mathbf{m}^k(\mathbf{x}) = [\delta m_1(\mathbf{x}), \delta m_2(\mathbf{x}), \ldots, \delta m_{N_p}(\mathbf{x})]^T$. A suitable ν^k can be estimated using various line-search algorithms (Nocedal and Wright, 2006). In optimization literature, the model update $\delta\mathbf{m}^k$ is referred to as the search or descent direction and can be derived from the gradient of the objective function with respect to the model parameters. Thus, the key components of FWI are computing $\mathbf{u}(\mathbf{m})$ (forward problem) and the gradient of the objective function with respect to the model parameters $\nabla_m J(\mathbf{m})$ (for the inverse problem).

Gradient-based optimization schemes are inherently local in nature i.e., they assume that the global minimum can be reached via iterative gradient-based updates. For non-linear optimization problems, the objective function typically contains numerous local minima. This property stipulates that to reach the global minimum, the starting point must already be within the same basin of attraction. Search based optimization methods such as Markov-Chain Monte Carlo methods or simulated annealing are not computationally feasible for FWI. In principle, these methods attempt to find the global minimum by evaluating the objective function at various points, chosen stochastically or probabilistically, in the model space. In FWI, the dimensions of the model space are too large to explore, particularly given the computational cost of evaluating the objective function once (proportional to N_s PDE solves). For these reasons, gradient-based optimization continues to be the preferred approach for FWI.

2.1.1 The forward problem

The time-domain formulation of the elastic wave equation is defined as

$$\rho(\mathbf{x})\ddot{\mathbf{u}}(\mathbf{x},t) - \nabla\cdot\boldsymbol{\sigma}(\mathbf{x},t) = \mathbf{f}(\mathbf{x},t), \tag{2.4}$$

inside the Earth Ω with surface $\partial\Omega$ (Aki and Richards, 2002). Time is denoted by $t \in [0, T]$ and $\mathbf{x} \in \Omega \subset \mathbb{R}^d$ denote spatial coordinates with dimensions $d = 1, 2, 3$; the derivations in this section take $d = 3$. The particle displacement $\mathbf{u}(\mathbf{x}, t)$ is excited by an external force $\mathbf{f}(\mathbf{x}, t)$. The density $\rho(\mathbf{x})$ is a material property and $\boldsymbol{\sigma}(\mathbf{x}, t)$ is the stress tensor. Single and double dots above a variable indicate first and second time derivatives, respectively. The spatial gradient operator is denoted by ∇. Equation 2.4 is subject to the free-surface boundary condition which states that the traction (normal component of the stress) is zero at the boundary $\partial\Omega$

$$\boldsymbol{\sigma} \cdot \hat{\mathbf{n}}|_{\mathbf{x}\in\partial\Omega} = 0, \tag{2.5}$$

where $\hat{\mathbf{n}}$ is the unit normal vector. The initial conditions require the particle displacement and velocity be zero at $t = 0$,

$$\mathbf{u}(\mathbf{x}, t = 0) = \dot{\mathbf{u}}(\mathbf{x}, t = 0) = 0. \tag{2.6}$$

The linear strain tensor $\boldsymbol{\epsilon}(\mathbf{x}, t) = \frac{1}{2}(\nabla\mathbf{u} + \nabla\mathbf{u}^T)$ is related to the stress tensor $\boldsymbol{\sigma}(\mathbf{x}, t)$ via the constitutive relation

$$\boldsymbol{\sigma}(\mathbf{x}, t) = \mathbf{C}(\mathbf{x}) : \boldsymbol{\epsilon}(\mathbf{x}, t). \tag{2.7}$$

The elastic tensor $\mathbf{C} = C_{ijkl}$ is a fourth order tensor that characterizes the subsurface material properties of the Earth. The : is defined as a contraction over repeated indices such that Equation 2.7 in index form is

$$\sigma_{ij} = \sum_{k=1}^{d}\sum_{l=1}^{d} C_{ijkl}\epsilon_{kl}. \tag{2.8}$$

The physical parameters comprised in $\mathbf{m}$ are problem dependent. For the general elastic wave equation (Equation 2.4), $\mathbf{m} = [\rho, \mathbf{C}]^T$. While $\mathbf{C}$ contains 81 elements, the symmetries of $\boldsymbol{\sigma}$ and $\boldsymbol{\epsilon}$ along with thermodynamic considerations require that

$$C_{ijkl} = C_{jikl} = C_{ijlk} = C_{klij}, \tag{2.9}$$

thereby limiting the number of independent elastic constants to 21 (Aki and Richards, 2002). Imposing additional symmetries on the medium, by making structural assumptions, further reduces the number of independent parameters required to characterize the subsurface. For an isotropic medium, the elastic tensor may be represented succinctly in terms of only two independent parameters

$$C_{ijkl} = \lambda\delta_{ij}\delta_{kl} + \mu(\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}), \tag{2.10}$$

where $\lambda(\mathbf{x})$ and $\mu(\mathbf{x})$ are known as the Lamé parameters; the latter is also known as the shear modulus. Substituting Equation 2.10 into Equation 2.4, leads to an expression for the elastic wave equation in heterogeneous isotropic media,

$$\mathbf{L}(\rho, \lambda, \mu)\mathbf{u} = \rho\ddot{\mathbf{u}} - \nabla \cdot [\lambda(\nabla \cdot \mathbf{u})\mathbf{I} + \mu(\nabla\mathbf{u} + \nabla\mathbf{u}^T)] = \mathbf{f}, \tag{2.11}$$

where $\mathbf{I}$ is the identity. A potential model parametrization for an elastic isotropic medium is in terms of density and the Lamé parameters, $\mathbf{m} = [\rho, \lambda, \mu]^T$. Analytic solutions to eqs. 2.4-2.7 do not exist for general heterogeneous media; therefore, solutions must be computed using numerical methods for partial differential equations. Details about my implementation are provided in section 2.2.1.

2.1.2 The inverse problem

In FWI, the inverse problem equates to performing a gradient-based optimization to iteratively update $\mathbf{m}$. This section describes how $\nabla_m J(\mathbf{m})$ is computed via Lagrange multipliers and the adjoint-state method.

Let $\mathbb{U}$ be the space of all displacement fields $\tilde{\mathbf{u}}(\mathbf{x}, t)$ and $\mathbb{F}$ be the space of all potential source fields $\tilde{\mathbf{f}}(\mathbf{x}, t)$. For any $\tilde{\mathbf{u}} \in \mathbb{U}$ and $\tilde{\mathbf{f}} \in \mathbb{F}$, the inner products in either space are defined as

$$\left\langle \tilde{\mathbf{u}}, \tilde{\mathbf{f}} \right\rangle_{\mathbb{U}} = \left\langle \tilde{\mathbf{f}}, \tilde{\mathbf{u}} \right\rangle_{\mathbb{F}} = \int_\Omega \int_T \tilde{\mathbf{u}}(\mathbf{x}, t)\tilde{\mathbf{f}}(\mathbf{x}, t) \, \mathrm{d}t \, \mathrm{d}^3\mathbf{x}. \tag{2.12}$$

In this section, we make a distinction between any wavefield $\tilde{\mathbf{u}} \in \mathbb{U}$ and solutions of eqs. 2.4-2.7. Specifically, we refer to solutions of the wave equation as $\mathbf{u}$ or $\mathbf{u}_s$.

To solve the constrained optimization problem in Equation 2.1, we invoke the method of Lagrange multipliers (e.g. Plessix (2006)). Consider a vector of Lagrange multipliers $\tilde{\mathbf{q}} = [\tilde{\mathbf{q}}_1, \tilde{\mathbf{q}}_2, \ldots, \tilde{\mathbf{q}}_{N_s}]^T$ and a corresponding vector of wavefields $\tilde{\mathbf{u}} = [\tilde{\mathbf{u}}_1, \tilde{\mathbf{u}}_2, \ldots, \tilde{\mathbf{u}}_{N_s}]^T$. The Lagrangian functional $\mathcal{L}(\tilde{\mathbf{u}}, \tilde{\mathbf{q}}, \mathbf{m})$ for a system of N_s equations is

$$\begin{aligned} \mathcal{L}(\tilde{\mathbf{u}}, \tilde{\mathbf{q}}, \mathbf{m}) &= \hat{J}(\tilde{\mathbf{u}}, \mathbf{m}) - \sum_{s=1}^{N_s} \langle \tilde{\mathbf{q}}_s, \mathbf{L}\tilde{\mathbf{u}}_s(\mathbf{x}, t) - \mathbf{f}_s(\mathbf{x}, t) \rangle_{\mathbb{F}}, \\ &= \frac{1}{2} \sum_{s=1}^{N_s} \sum_{r=1}^{N_r} \int_T |\tilde{\mathbf{u}}_s(\mathbf{x}_r, t) - \mathbf{d}_s(\mathbf{x}_r, t)|^2 \, \mathrm{d}t - \sum_{s=1}^{N_s} \langle \tilde{\mathbf{q}}_s, \mathbf{L}\tilde{\mathbf{u}}_s(\mathbf{x}, t) - \mathbf{f}_s(\mathbf{x}, t) \rangle_{\mathbb{F}}, \end{aligned} \tag{2.13}$$

where $\hat{J}(\tilde{\mathbf{u}}, \mathbf{m})$ is the misfit functional for arbitrary wavefields $\tilde{\mathbf{u}}_s(\mathbf{x}_r, t)$; we assume the waveform misfit functional introduced in Equation 2.2 for this development. We seek to find $\tilde{\mathbf{u}}, \tilde{\mathbf{q}}$, and $\mathbf{m}$ for which $\mathcal{L}(\tilde{\mathbf{u}}, \tilde{\mathbf{q}}, \mathbf{m})$ satisfies the first-order optimality conditions (Kuhn

and Tucker, 1951)

$$\frac{\partial \mathcal{L}(\tilde{\mathbf{u}}, \tilde{\mathbf{q}}, \mathbf{m})}{\partial \tilde{\mathbf{u}}_s} = 0, \quad s = \{1, \dots, N_s\} \tag{2.14}$$

$$\frac{\partial \mathcal{L}(\tilde{\mathbf{u}}, \tilde{\mathbf{q}}, \mathbf{m})}{\partial \tilde{\mathbf{q}}_s} = 0, \quad s = \{1, \dots, N_s\} \tag{2.15}$$

$$\frac{\partial \mathcal{L}(\tilde{\mathbf{u}}, \tilde{\mathbf{q}}, \mathbf{m})}{\partial \mathbf{m}} = 0. \tag{2.16}$$

Examining these derivatives individually provides insight into the relevance behind each stationary point. For example, satisfying

$$\frac{\partial \mathcal{L}(\tilde{\mathbf{u}}, \tilde{\mathbf{q}}, \mathbf{m})}{\partial \tilde{\mathbf{q}}_s} = \mathbf{L}\tilde{\mathbf{u}}_s(\mathbf{x}, t) - \mathbf{f}_s(\mathbf{x}, t) = 0. \tag{2.17}$$

is equivalent to finding a solution to the wave equation, To satisfy the first condition, we require that $\tilde{\mathbf{u}}_s = \mathbf{u}_s, \forall s \in \{1, \dots, N_s\}$. The derivative of Equation 2.20 with respect to $\mathbf{m}$ gives

$$\frac{\partial \mathcal{L}(\tilde{\mathbf{u}}, \tilde{\mathbf{q}}, \mathbf{m})}{\partial \mathbf{m}} = -\sum_{s=1}^{N_s} \left\langle \tilde{\mathbf{q}}_s, \frac{\partial \mathbf{L}}{\partial \mathbf{m}} \tilde{\mathbf{u}}_s \right\rangle_{\mathbb{F}}. \tag{2.18}$$

Taking the derivative of Equation 2.20 with respect to $\tilde{\mathbf{u}}_s$ gives

$$\frac{\partial \mathcal{L}(\tilde{\mathbf{u}}, \tilde{\mathbf{q}}, \mathbf{m})}{\partial \tilde{\mathbf{u}}_s} = \frac{\partial \hat{J}(\tilde{\mathbf{u}}, \mathbf{m})}{\partial \tilde{\mathbf{u}}_s} - \mathbf{L}^\dagger \tilde{\mathbf{q}}_s = 0, \quad s = \{1, \dots, N_s\}, \tag{2.19}$$

where $\hat{J}(\tilde{\mathbf{u}}, \mathbf{m})$ is,

$$\hat{J}(\tilde{\mathbf{u}}, \mathbf{m}) = \frac{1}{2} \sum_{s=1}^{N_s} \sum_{r=1}^{N_r} \int_T |\tilde{\mathbf{u}}_s(\mathbf{x}_r, t) - \mathbf{d}_s(\mathbf{x}_r, t)|^2 \, \mathrm{d}t. \tag{2.20}$$

Also we have used the definition of an adjoint operator

$$\langle \tilde{\mathbf{q}}_s, \mathbf{L}\tilde{\mathbf{u}}_s \rangle_{\mathbb{F}} = \left\langle \mathbf{L}^\dagger \tilde{\mathbf{q}}_s, \tilde{\mathbf{u}}_s \right\rangle_{\mathbb{U}}. \tag{2.21}$$

Rearranging Equation 2.19 yields a more recognisable form

$$\mathbf{L}^\dagger \tilde{\mathbf{q}}_s(\mathbf{x}, t) = \sum_{r=1}^{N_r} \tilde{\mathbf{u}}_s(\mathbf{x}_r, t) - \mathbf{d}_s(\mathbf{x}_r, t). \tag{2.22}$$

Indeed, Equation 2.22, referred to as the adjoint-state equation, resembles the wave equation in Equation 2.4. The operator $\mathbf{L}^\dagger$ is the adjoint wave-equation operator. It can be shown that solving Equation 2.22 is equivalent to solving Equation 2.4 in reverse time ($t = T \to 0$) with time reversed residuals, located at the receiver positions, acting as sources. For the

elastic wave equation in non-dissipative media, $\mathbf{L}$ is self-adjoint ($\mathbf{L} = \mathbf{L}^\dagger$) (Fichtner, 2010). The derivation of the adjoint operator is well established and is therefore omitted; interested readers may refer to Tarantola (1984b); Mora (1987); Tromp et al. (2005); Plessix (2006); Fichtner et al. (2006) for a complete description. The solutions of the adjoint-state equation are the adjoint wavefields $\tilde{\mathbf{q}}_s(\mathbf{x}, t)$. The term on the right hand side of Equation 2.22 is the adjoint source $\tilde{\mathbf{f}}_s^\dagger(\mathbf{x}, t)$.

Finally, to obtain $\nabla_m J(\mathbf{m})$ we recognize that

$$\mathcal{L}(\mathbf{u}, \tilde{\mathbf{q}}, \mathbf{m}) = J(\mathbf{m}). \tag{2.23}$$

Equation 2.23 holds since $\mathbf{u}$ are solutions to the state equations specified in Equation 2.17. Using this in Equation 2.18 gives the desired gradient after a couple of steps. First, take

$$\frac{\partial \mathcal{L}(\mathbf{u}, \tilde{\mathbf{q}}, \mathbf{m})}{\partial \mathbf{m}} = \frac{\partial J(\mathbf{m})}{\partial \mathbf{m}} + \frac{\partial \mathcal{L}}{\partial \mathbf{u}} \frac{\partial \mathbf{u}}{\partial \mathbf{m}}. \tag{2.24}$$

The adjoint wavefield $\mathbf{q}_s$ is a solution to Equation 2.22 having substituted $\mathbf{u}_s$ on the right hand side. By substituting $\mathbf{q}_s$ into Equation 2.24, the second term on the right hand side vanishes by definition of the adjoint state. The gradient simplifies to

$$\frac{\partial \mathcal{L}(\mathbf{u}, \mathbf{q}, \mathbf{m})}{\partial \mathbf{m}} = \frac{\partial J(\mathbf{m})}{\partial \mathbf{m}} = -\sum_{s=1}^{N_s} \left\langle \mathbf{q}_s, \frac{\partial \mathbf{L}}{\partial \mathbf{m}} \mathbf{u}_s \right\rangle_{\mathbb{F}}. \tag{2.25}$$

It is more meaningful to examine the gradient in dimensions of the model. The volumetric densities of the derivative are

$$\frac{\partial J}{\partial \mathbf{m}}(\mathbf{x}) = -\sum_{s=1}^{N_s} \int_T \mathbf{q}_s \cdot \frac{\partial \mathbf{L}}{\partial \mathbf{m}} \mathbf{u}_s \, \mathrm{d}t. \tag{2.26}$$

Equation 2.26 states that the gradient kernels are the zero-lag correlation between the forward and adjoint-wavefields at each point in space. The FWI gradient bears resemblance to the classic imaging condition in reflection seismology (Claerbout, 1971). The computation of $\nabla_m J(\mathbf{m})$ can be summarized as follows: 1) Compute $\mathbf{u}_s$ as solutions to the wave equation (eqs. 2.4-2.7). 2) Compute the adjoint wavefield $\mathbf{q}_s$ as solutions to the adjoint-state equation (2.22). The adjoint sources are the time-reversed waveform residuals located at the receiver positions. 3) Compute $\nabla_m J(\mathbf{m})$ from the interaction between forward and adjoint wavefields as indicated by Equation 2.26. The cost of each gradient computation is $2N_s$ PDE solves, N_s to compute $\mathbf{u}_s$ and N_s to compute $\mathbf{q}_s$ for $s \in \{1, \ldots, N_s\}$.

For elastic, isotropic FWI the gradients parametrized in terms of density and the Lamé

parameters are expressed as

$$\frac{\partial J}{\partial \rho}(\mathbf{x}) = -\int_T \mathbf{u}^\dagger \cdot \ddot{\mathbf{u}} \, \mathrm{d}t, \tag{2.27}$$

$$\frac{\partial J}{\partial \lambda}(\mathbf{x}) = \int_T (\nabla \mathbf{u}^\dagger) \cdot (\nabla \mathbf{u}) \, \mathrm{d}t, \tag{2.28}$$

$$\frac{\partial J}{\partial \mu}(\mathbf{x}) = \int_T 2(\nabla \mathbf{u}^\dagger) : \boldsymbol{\epsilon} \, \mathrm{d}t. \tag{2.29}$$

A more common parametrization uses seismic velocities, $\mathbf{m} = [\rho, v_p, v_s]$, where v_p and v_s are the P-wave and S-wave velocities, respectively. The Lamé parameters are related to v_p and v_s through $v_p = \sqrt{(\lambda + 2\mu)/\rho}$ and $v_s = \sqrt{\mu/\rho}$. The gradients for the new parametrization can be obtained via the chain rule; the new gradients are linear combinations of the existing ones,

$$\frac{\partial J}{\partial \rho'} = \frac{\partial J}{\partial \rho} + (v_p^2 - 2v_s^2)\frac{\partial J}{\partial \lambda} + v_s^2 \frac{\partial J}{\partial \mu}, \tag{2.30}$$

$$\frac{\partial J}{\partial v_p} = 2\rho v_p \frac{\partial J}{\partial \lambda}, \tag{2.31}$$

$$\frac{\partial J}{\partial v_s} = 2\rho v_s \frac{\partial J}{\partial \mu} - 4\rho v_s \frac{\partial J}{\partial \lambda}. \tag{2.32}$$

We reiterate that the spatial dependencies of variables have been omitted. The gradients and model parameters are functions of space. Field variables are functions of space and time.

For the waveform misfit functional, the gradient of Equation 2.2 can also be expressed as

$$\frac{\partial J}{\partial \mathbf{m}}(\mathbf{x}) = \sum_{s=1}^{N_s} \sum_{r=1}^{N_r} \int_T \frac{\partial \mathbf{u}_s(\mathbf{x}_r, t)}{\partial \mathbf{m}(\mathbf{x})} \left[\mathbf{u}_s(\mathbf{x}_r, t) - \mathbf{d}_s(\mathbf{x}_r, t)\right] \, \mathrm{d}t, \tag{2.33}$$

where the derivative of $\mathbf{u}_s$ with respect to $\mathbf{m}$ is the Fréchet derivative operator. Computing the Fréchet derivative operator requires as many PDE solves as model parameters in the discrete setting (Pratt et al., 1998). The adjoint-state method avoids explicitly computing the Fréchet derivatives by computing an equivalent product. The Hessian can be computed by differentiating 2.1.2 with respect to $\mathbf{m}(\mathbf{y})$,

$$\begin{aligned}\mathbf{H}(\mathbf{x}, \mathbf{y}) = \frac{\partial^2 J}{\partial \mathbf{m}(\mathbf{x}) \partial \mathbf{m}(\mathbf{y})} = \sum_{s=1}^{N_s} \sum_{r=1}^{N_r} \int_T & \frac{\partial \mathbf{u}_s(\mathbf{x}_r, t)}{\partial \mathbf{m}(\mathbf{x})} \frac{\partial \mathbf{u}_s(\mathbf{x}_r, t)}{\partial \mathbf{m}(\mathbf{y})} \\ & + \frac{\partial^2 \mathbf{u}_s(\mathbf{x}_r, t)}{\partial \mathbf{m}(\mathbf{x}) \partial \mathbf{m}(\mathbf{y})} \left[\mathbf{u}_s(\mathbf{x}_r, t) - \mathbf{d}_s(\mathbf{x}_r, t)\right] \, \mathrm{d}t\end{aligned} \tag{2.34}$$

The first term on the right-hand side is the Gauss-Newton, or approximation, Hessian and

accounts for first-order scattering effects (Pratt et al., 1998). The second term is related to second-order scattering effects. The second term is often neglected due difficulties in evaluating the term. Furthermore, close to the global minimum, the data residuals should be small making the term less of lesser importance. For the majority of this thesis, we focus on the Gauss-Newton form of the Hessian; however, where possible, the notation is general and the concepts are applicable to both approximate and full forms of the Hessian. The Hessian is seldom evaluated in FWI due to the computational cost associated with computing the Fréchet derivatives. Hessian-vector products can be computed relatively efficiently using second-order adjoint-state methods (Fichtner and Trampert, 2011a). Hessian-vector products require at least 2 additional PDE solves not including those required to compute the gradient. Hessian-vector products have a range of utilities that will be exploited throughout this thesis.

2.1.3 The algorithm

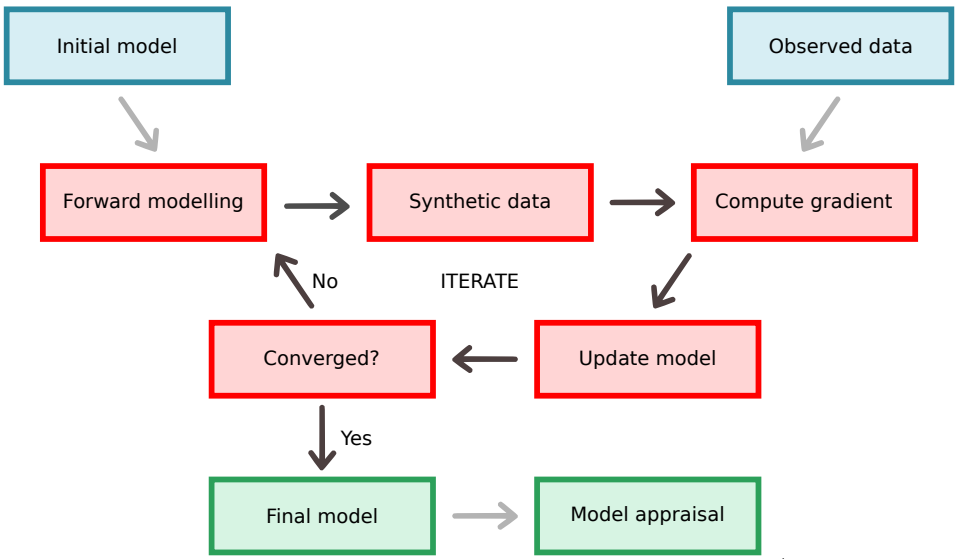

Figure 2.1: Simplified FWI workflow. Grey arrows mark input/output for the algorithm.

With the forward problem established and a recipe to compute $\nabla_m J$, an algorithm for FWI can be described. A simplified workflow is displayed in Figure 2.1 while a more complete time-domain FWI algorithm is presented in algorithm 1.

From algorithm 1, the computational cost of a single FWI iteration can be deduced. Each gradient computation (lines 2-8), requires $2N_s$ PDE solves per iteration. Each iteration of the line-search selects a trial step length ν^t, computes a trial model update $\mathbf{m}^t = \mathbf{m}^k +$

Algorithm 1 Conventional FWI using an L_2 waveform difference objective function

1: **for** $k = 1, \text{Max. iterations}$ **do**
2: **for** $s = 1, N_s$ **do** ▷ Loop over sources
3: Compute $\mathbf{u}_s(\mathbf{x}, t; \mathbf{m}^k)$ ▷ 1 PDE solve
4: Process $\mathbf{d}_s(\mathbf{x}_r, t)$ and $\mathbf{u}_s(\mathbf{x}_r, t; \mathbf{m}^k)$
5: Compute $\mathbf{f}_s^\dagger(\mathbf{x}, t) = \sum_{r=1}^{N_r}[\mathbf{u}_s(\mathbf{x}_r, t; \mathbf{m}^k) - \mathbf{d}_s(\mathbf{x}_r, t)]\delta(\mathbf{x} - \mathbf{x}_r)$
6: Compute $\mathbf{q}_s(\mathbf{x}, t)$ ▷ 1 PDE solve
7: Compute $(\nabla_m J)_s = -\int_T \mathbf{q}_s \cdot \frac{\partial \mathbf{L}}{\partial \mathbf{m}} \mathbf{u}_s \, \mathrm{d}t$
8: **end for**
9: $\nabla_m J = \sum_{s=1}^{n_s} (\nabla_m J)_s$
10: Compute $\delta \mathbf{m}^k$ ▷ Gradient-based techniques
11: Compute ν^k ▷ Line search. Requires PDE solves
12: $\mathbf{m}^{k+1} = \mathbf{m}^k + \nu^k \delta \mathbf{m}^k$
13: **if** $||\mathbf{m}^{k+1} - \mathbf{m}^k||/||\mathbf{m}^k|| < \epsilon$ **then** ▷ Check for convergence
14: **return** $\mathbf{m}^{k+1}$
15: **end if**
16: **end for**

$\nu^t \delta \mathbf{m}^k$, and then evaluates the objective function $J(\mathbf{m}^t)$ incurring a cost of N_s additional PDE solves. In my implementation, the line search is based on either a bracketing or backtracking search that satisfies the Armijo stopping criteria (Nocedal and Wright, 2006). A line search that runs for N_l iterations requires $N_l N_s$ PDE solves. Ultimately, the total cost of an FWI iteration is $(2 + N_l)N_s$ PDE solves. Since PDE solves for each source are independent, algorithm 1 can be readily parallelized with an embarrassingly parallel scheme i.e. by distributing the same (or very similar) tasks to independent CPU processes.

Line 7 of Algorithm 1 performs a zero-lag correlation between the forward wavefield (at time t) and the adjoint wavefield propagating in reverse time (time $T-t$). The correlation requires simultaneous access to both wavefields at various points in time. The simplest approach to address this is to pre-compute the forward wavefield and store all time steps in memory. In step 7, the forward wavefield is then read from memory/disk to compute the gradient. This approach can incur large memory costs and potentially introduce I/O bottlenecks for large models. An alternative, yet similar, approach stores the wavefield at select time intervals and recomputes intermediate time steps (Symes, 2007). Optimized checkpointing algorithms can reduce the storage requirements with a slight increase in computational cost. We adopt a wavefield reconstruction approach that recomputes the forward wavefield on-the-fly during adjoint wavefield computations. The reconstruction method stores the forward wavefield at the final time step along with a small number of interior boundary elements for all time steps (only for absorbing boundaries). The number of interior elements stored is equal to half the length of the finite-difference stencil used (e.g., 2 elements for a fourth-order stencil). By storing the boundary elements, we are able to re-inject energy that is dissipated by the absorbing boundaries. Wavefield reconstruction has the smallest memory overhead,

but effectively adds an additional PDE solve per gradient computation. In the presence of attenuation, carefully designed checkpointing algorithms can account for energy dissipation that occurs within the modelling domain (Yang et al., 2016).

2.1.4 Practical considerations

Multi-scale strategies

There are numerous adaptations to the generic FWI algorithm that can improve convergence properties. Adaptations are often necessary when deficiencies in the data prohibit proper convergence of the inversion. The most common alteration is the inclusion of multi-scale strategies (e.g., Bunks et al. (1995)). Multi-scale methods perform inversion in a hierarchical manner, fitting simple/more reliable features of the data before progressing to more challenging ones. A standard multi-scale approach is that of frequency continuation where low frequencies in the data are fit before transitioning to higher frequencies. In these schemes, the inverted model from the end of each stage is used as the initial model for the subsequent stage. Additional multi-scale strategies include time windowing to isolate various seismic events, and relaxation of scale lengths for gradient smoothers.

Choice of objective function

The L2 waveform difference is seldom used in practice due to its susceptibility to cycle skipping. Cycle skipping occurs when distinct seismic phases are shifted by more than half a wavelength relative to one another (Virieux and Operto, 2009). In such cases, the waveform difference will subtract unintended portions of the data and synthetics from each other. In addition, the waveform difference promotes fitting amplitudes, which in real data FWI is generally not prioritized. Inadequate physics models can lead to inaccurate dynamics during wave propagation making amplitude fitting unreliable. For real data applications, emphasis is placed on fitting the kinematics observed in the data. For this, alternative objective functions (e.g., traveltime/correlation-based, envelope, optimal-transport) are more suitable. Some objective functions offer more robust convergence owing to fewer local minima and/or broader basins of attraction in the optimization landscape. Changing the objective function merely alters the adjoint source in step 5 of algorithm 1 (see (2.19)).

Optimization algorithm

The search direction $\delta\mathbf{m}^k$ can be computed from a range of gradient-based optimization algorithms. Common choices include steepest descent, non-linear conjugate gradients, quasi-Newton methods (e.g., L-BFGS), and truncated Newton methods (Nocedal and Wright,

2006). The simplest choice is the steepest descent step $\delta\mathbf{m}^k = -\nabla_m J^k$. More advanced methods can improve convergence rates of the algorithm, although this comes at a significant cost with Newton-type methods. Truncated Newton methods are explored in more detail in Chapter 4. Gradient preconditioning can help provide more reliable updates; common choices include scale and/or structure dependent smoothers and weighting functions.

Source estimation

Thus far we have neglected discussion of the seismic source $\mathbf{f}$ that is used to generate the simulated data. In real data studies, the seismic source is not typically known and has to be estimated. An initial estimate of the source can, at times, be estimated through data processing techniques such as water bottom stacking for marine data or from vibroseis sweeps for vibroseis land data. The treatment of the source in FWI varies between studies. Pratt et al. (1998) proposes a technique that estimates the source through the solution of a linear inverse problem. The method finds a filter $s(t)$ that minimizes the expression

$$\sum_{r=1}^{N_r} \int_T \left[\mathbf{d}(\mathbf{x}_r, t) - s(t) * \mathbf{u}(\mathbf{x}_r, t)\right]^2 \, \mathrm{d}t, \tag{2.35}$$

where $*$ denotes convolution. The initial simulated data are best modelled with a source that possesses a similar frequency spectrum to the data. An Ormsby wavelet is a useful candidate as the amplitude spectrum can be controlled more readily. The solution to Equation 2.35 can be expressed conveniently in the frequency-domain as

$$s(\omega) = \frac{\sum_{r=1}^{N_r} \bar{\mathbf{u}}(\mathbf{x}_r, \omega)\mathbf{d}(\mathbf{x}_r, \omega)}{\sum_{r=1}^{N_r} \bar{\mathbf{u}}(\mathbf{x}_r, \omega)\mathbf{u}(\mathbf{x}_r, \omega) + \epsilon}, \tag{2.36}$$

where $\mathbf{d}(\mathbf{x}_r, \omega) = \mathcal{F}\{\mathbf{d}(\mathbf{x}_r, t)\}, \mathbf{u}(\mathbf{x}_r, \omega) = \mathcal{F}\{\mathbf{u}(\mathbf{x}_r, t)\}, s(\omega) = \mathcal{F}\{s(t)\}$, $\mathcal{F}$ is the Fourier transform operator, and ϵ is a small value to avoid division by zero. The bar above variables denotes complex conjugation. The Fourier transform is defined as

$$\mathcal{F}\{h(t)\} \stackrel{\text{def}}{=} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} h(t) \exp^{-i\omega t} \, \mathrm{d}t \tag{2.37}$$

for an arbitrary function $h(t)$ and $i = \sqrt{-1}$. Equation 2.36 deconvolves the summed correlation of the data and synthetics with the summed autocorrelation of the synthetic data. The final source used during inversion is $s(t) * \mathbf{f}(\mathbf{x}, t)$. The filters can be computed for each source or as an average over the survey. Estimating the source requires N_s PDE solves to estimate the initial set of synthetics. To reduce computational cost, the source is typically updated at regular intervals throughout the inversion as opposed to at every iteration. In

practice, it can be useful to use time- and offset-windowed data for source estimation. The approach remains the same but $\mathbf{u}(\mathbf{x}_r, t)$ is replaced with $\mathbf{W}(\mathbf{x}_r, t)\mathbf{u}(\mathbf{x}_r, t)$, where $\mathbf{W}(\mathbf{x}_r, t)$ is a potentially source and receiver dependent windowing function; a similar substitution is used for the data. Typically, the windowing is used to select early arrivals and short-to-mid offsets. Alternatively, Plessix and Cao (2011) include the source as an inversion parameter and derive a scheme to iteratively update it using gradient updates.

2.2 Numerical implementation

In this section, I provide an overview of my numerical implementation of time-domain elastic FWI. The description is divided into two parts, one for the numerical PDE solver and another for the inversion workflow. We do not describe the details of the numerical methods here as they are already well established.

2.2.1 Wave equation solver

The (isotropic) elastic wave equation is solved using a time-stepping finite-difference scheme. The solver is limited to two spatial dimensions to make computations more tractable. $P-SV$ wave modes are simulated in heterogeneous media using a staggered-grid, finite-difference scheme that is fourth-order accurate in space and second-order accurate in time, $\mathcal{O}(\Delta h^4, \Delta t^2)$ (Yee, 1966; Virieux, 1986; Levander, 1988). The staggered grid defines field variables (e.g. displacement/velocity, stress) at regular $(x + \Delta x)$ and half intervals $(x + \Delta x/2)$, thereby reducing the effective grid spacing and reducing numerical dispersion (Virieux, 1986). Wave equation solvers operate on local computational domains that are significantly smaller than the physical domain of the Earth. To simulate wave propagation in a medium much larger than the computational domain, efficient absorbing boundary conditions are necessary. Absorbing boundary conditions are implemented in the form of convolutional perfectly matched layers which have been demonstrated to be effective at absorbing waves in elastic wave propagation (Komatitsch and Martin, 2007). The code is developed in C for fast numerical computations. An embarrassingly parallel scheme over seismic sources is implemented using MPI. The adjoint-state method requires simultaneous access to the forward and adjoint wavefields at all timesteps to calculate the gradient. The forward wavefield could be stored at all timesteps; however, this becomes memory intensive and may result in I/O bottlenecks for large problems. In my implementation, I calculate the regular wavefield and only store a snapshot of the final timestep along with the boundary elements at all time steps. During the computation of the adjoint wavefield, the forward wavefield is reconstructed in reverse time. This method effectively requires $3N_s$ PDE solves per gradient computation, compared with $2N_s$ PDE solves if the forward wavefield is stored.

2.2.2 Inversion workflow management

A complete FWI implementation requires a number of additional components as well as an efficient wave equation solver. Standard signal processing tools are required for intermediate processing steps (e.g., data muting, filtering, etc.), array processing tools for gradient processing, and robust optimization libraries for gradient-based optimization. In selecting my workflow implementation I had two requirements: 1) Utilize existing libraries/code packages as much as possible. As many of the supplementary routines are not specific to FWI, a range of well developed open-source packages already exist. 2) Exhibit high flexibility and portability to enable rapid prototyping.

To fulfil these criteria I chose the Seisflows inversion framework, an open-source Python workflow tool designed for the SPECFEM class of solvers (spectral-element solvers available at https://geodynamics.org/cig/software/specfem3d/) (Modrak et al., 2018). I developed custom routines to interface Seisflows with my own solver. After full integration, the framework allows me to perform automated inversions with a range of processing and optimization options. Another key feature of Seisflows is its portability. It is able to transition from desktop PCs to HPC compute cluster environments with only minimal changes to the code.

2.3 Multi-parameter inversion

Early applications of FWI in exploration seismology focused on single-parameter inversion of P-wave velocity using the acoustic wave equation. As demonstrated in Equation 2.11, more complex physics in the wave equation are characterized by an increased number of independent physical parameters.

The estimation of multiple independent parameters with limited data poses a significant challenge in multi-parameter FWI. Perturbations to different physical properties in the Earth's subsurface can give rise to similar responses in the data. This introduces trade-off between the parameters during inversion, making it difficult, or impossible to resolve independent parameters uniquely. The model, model parametrization, acquisition geometry, and data, all affect the resolvability of individual parameters (Tarantola, 1986; Pratt et al., 1998; Operto et al., 2013). Information on the resolution of model parameters is contained in the Hessian of the objective function. Parameter trade-offs can be partially corrected by using second-order optimization techniques (e.g. Newton, Gauss-Newton, truncated Newton) (Pratt et al., 1998; Métivier et al., 2013). While second-order optimization methods have been explored in FWI (Epanomeritakis et al., 2008; Métivier et al., 2013; Anagaw and Sacchi, 2014; Castellanos et al., 2015; Pan et al., 2016; Yang et al., 2018a), they are computationally expensive and justifying their additional costs can be difficult. In some

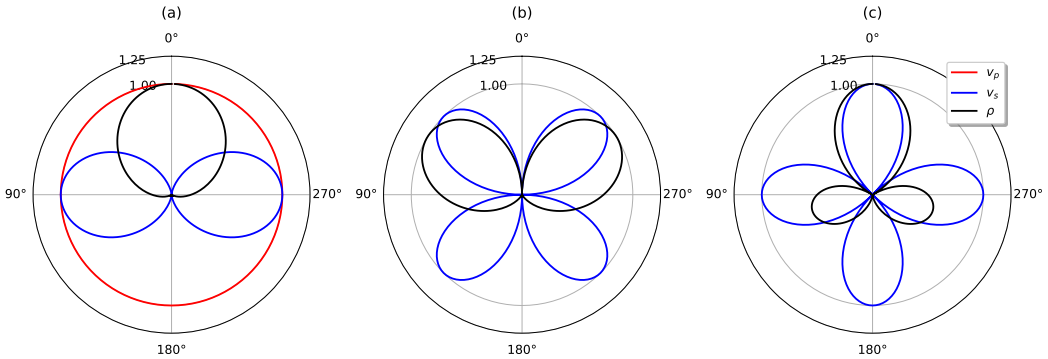

Figure 2.2: Scattering patterns for various wave modes with respect to perturbations in v_p, v_s, and ρ. (a) PP modes. (b) PS/SP modes. (c) SS modes. Angles are scattering angles measured from vertical. Patterns represent amplitudes of scattered wave modes computed using the Born approximation.

scenarios, quasi-Newton methods (e.g. L-BFGS) can achieve comparable inversion results at a reduced cost (Métivier et al., 2013). In lieu of incorporating the Hessian, parameter trade-off can be limited through a combination of data-driven inversion strategies (Shipp and Singh, 2002; Sears et al., 2008; Prieux et al., 2013a,b) and appropriate model parametrization (Tarantola, 1986; Plessix and Cao, 2011; Köhn et al., 2012; Operto et al., 2013; Gholami et al., 2013; Alkhalifah and Plessix, 2014). Understanding parameter trade-off is necessary for accurate model appraisal, particularly when first-order optimization methods are used. A common approach to evaluating the trade-off between parameters is by analyzing scattering patterns as presented in Figure 2.2. Scattering patterns depict the amplitude response of a scattered wavefield with respect to perturbations in different parameters. The responses are plotted as a function of scattering angle. Scattering patterns are derived from high-frequency ray and Born approximations (Wu and Aki, 1985; Tarantola, 1986; Beylkin and Burridge, 1990; Forgues and Lambaré, 1997). Scattering angles for which amplitude responses show significant overlap are interpreted as having greater parameter trade-off. For example, PP scattered modes at narrow scattering angles exhibit considerable overlap for v_p and ρ responses. From an inversion perspective, this suggests that if only narrow scattering angles are available in the data then it will be difficult to decouple v_p and ρ perturbations in the subsurface. Scattering patterns are useful because they have analytical expressions; however, they are somewhat qualitative and do not take into consideration the finite-frequency nature of wave propagation nor the acquisition geometry of a seismic survey.

Parameter trade-offs manifest mathematically in the multi-parameter Hessian. The Hessian carries information pertaining to the strength of parameter trade-offs along with the spatial resolution afforded by the acquisition geometry (Pratt et al., 1998). Forgues and Lambaré (1997) demonstrated that under the ray + Born approximation, scattering patterns (Figure

2.2) are related to the Hessian. Neglecting the Hessian in multi-parameter inversion introduces inaccuracies into the inversion due to erroneous inter-parameter mappings (Operto et al., 2013). The multi-parameter Hessian operator exhibits a block structure and may be expressed in matrix form as

$$\mathbf{H}(\mathbf{x},\mathbf{y}) = \begin{bmatrix} \mathbf{H}_{m_1 m_1}(\mathbf{x},\mathbf{y}) & \dots & \mathbf{H}_{m_1 m_p}(\mathbf{x},\mathbf{y}) \\ \vdots & \ddots & \\ \mathbf{H}_{m_p m_1}(\mathbf{x},\mathbf{y}) & & \mathbf{H}_{m_p m_p}(\mathbf{x},\mathbf{y}) \end{bmatrix}. \tag{2.38}$$

The Newton equations in terms of the multi-parameter Hessian operator are

$$\sum_{j=1}^{N_p} \int \mathbf{H}_{m_i m_j}(\mathbf{x},\mathbf{y}) \delta m_j(\mathbf{y}) \, \mathrm{d}\mathbf{y} = -g_i(\mathbf{x}). \tag{2.39}$$

Equation 2.39 states that the gradient for the *i-th* model parameter is a linear combination of the true model perturbations weighted by the relevant block elements from the Hessian. Due to the expense of Newton based methods, the Hessian in Equation 2.39 is often replaced with a diagonal preconditioning operator to give

$$\int \mathbf{P}_{m_i m_i}(\mathbf{x},\mathbf{y}) \delta m_i(\mathbf{y}) \, \mathrm{d}\mathbf{y} \approx -g_i(\mathbf{x}). \tag{2.40}$$

The lack of off-diagonal contributions ($i \neq j$) in $\mathbf{P}$ introduces inter- parameter mappings that lead to inaccuracies in inverted models. Block-diagonal approximations of the Hessian are a more sophisticated, yet inexpensive, form of preconditioning operator that account for local interparameter trade-off (Korta et al., 2013; Métivier et al., 2015).

To illustrate parameter trade-off, we present a simple example of multi-parameter FWI in Figure 2.3. We estimate a 3 parameter model that consists of spatially inconsistent Gaussian perturbations in P-wave velocity (v_p), S-wave velocity (v_s), and density (ρ); the true model appears in Figure 2.3a-c. In the second row (Figure 2.3d-f), we invert with sources and receivers distributed regularly around the entire boundary of the model. In the last row (Figure 2.3g-i, inversion is performed with a surface acquisition i.e. with sources and receivers distributed only along the top of the model. The anomalies are 5% perturbations from the background model for each parameter. In each inversion, all parameters are estimated simultaneously using 20 NLCG iterations.

In the complete acquisition, dense coverage results in ample subsurface illumination allowing for good recovery of the distinct perturbations. The surface acquisition provides limited illumination of the subsurface due to the geometry restrictions. The acquisition limits the range of scattering angles sampled by the data to narrow and intermediate angles. For this

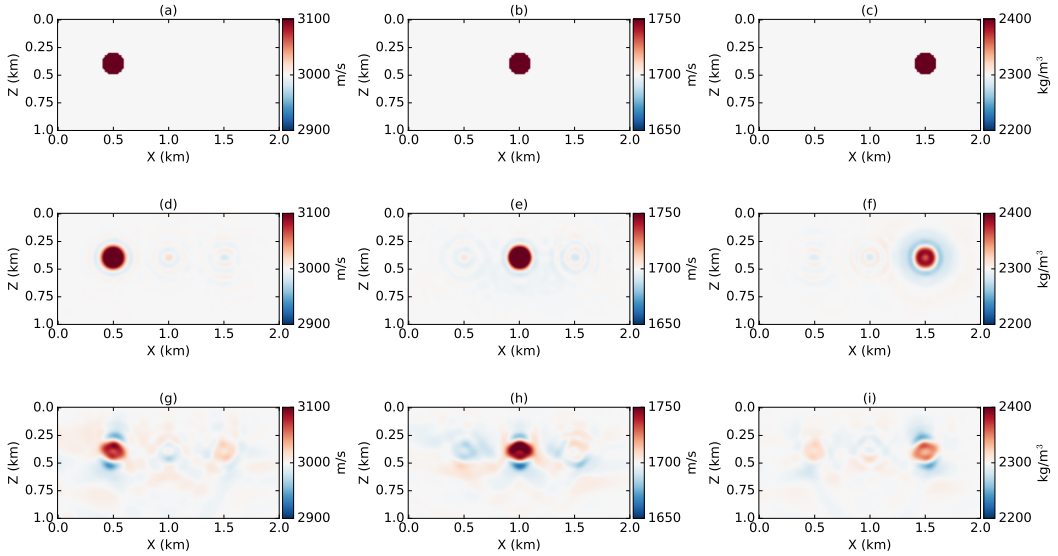

Figure 2.3: Toy inversion test for v_p, v_s and ρ Gaussian anomalies.(a-c) True model. (d-f) Inverted model with full acquisition. (g-i) Inverted model with surface acquisition. The surface acquisition provides restricted subsurface illumination resulting in degraded spatial resolution and parameter separation. The artefacts observed in (g-i) correspond to erroneous mappings, or parameter cross-talk, from other parameters.

range of scattering angles, increased trade-off occurs between P and ρ.

Uniquely resolving multiple parameters is a nuanced problem that is made challenging due to parameter trade-off and limitations of the data. The increased model space results in greater non-uniqueness in the solutions due to a larger null space. Throughout this thesis, we explore some of the challenges discussed in this chapter in more detail.

CHAPTER 3

Source-encoded multi-parameter FWI[1]

3.1 Introduction

Since its conception in the 1980s (Lailly, 1983; Tarantola, 1984a, 1986; Mora, 1987), full waveform inversion (FWI) has matured from a mathematical concept to a viable imaging technique used to estimate physical parameters in the Earth's subsurface. The advent of modern supercomputers coupled with algorithmic advances have led to a flurry of successful applications of FWI in both exploration (Crase et al., 1992; Pratt et al., 1998; Shipp and Singh, 2002; Ravaut et al., 2004; Sears et al., 2008; Virieux and Operto, 2009; Brossier et al., 2009; Krebs et al., 2009; Prieux et al., 2013a) and earthquake seismology (Chen et al., 2007; Tape et al., 2009; Fichtner et al., 2009; French et al., 2013; Zhu et al., 2015). In spite of recent advances, the computational cost of FWI remains a limiting factor for large scale 3D applications on real data. As practitioners continue to accommodate larger datasets, efficient algorithms are crucial to ensuring FWI remains tractable. Potential transitions to more complex physics (e.g. acoustic to elastic) further compound the computational cost. The bulk of FWI's computational expense arises from computing numerical solutions to multiple partial differential equations (PDEs) per source at each iteration. The linear dependence of the cost on the number of sources hampers the scalability of FWI for large datasets. Source encoding effectively reduces the dimensionality of the data by utilizing multiple sources simultaneously rather than independently (Romero et al., 2000; Krebs et al., 2009).

Source encoding was originally proposed by Romero et al. (2000) to reduce the cost of

[1]A version of this chapter is published in Matharu. G., and M. D. Sacchi, 2017, Source encoding in multi-parameter full waveform inversion, Geophysical Journal International, Volume 214, Issue 2, Pages 792-810.

shot-record migration. They substituted individual sources with a smaller number of encoded sources, where each encoded source represented a weighted linear combination of individual sources. The weights —known as encoding functions— were chosen as random phase-shifts and are necessary to reduce cross-talk artefacts in the corresponding migration image. Cross-talk artefacts arise from interactions between different sources in the imaging condition. Romero et al. (2000) significantly reduced the cost of shot-record migration whilst maintaining acceptable image quality.

Following its introduction, source encoding has been explored in a range of applications. Simultaneous sources (without source encoding) were utilized to improve the feasibility of global tomography (Capdeville et al., 2005). Capdeville et al. (2005) demonstrated reduced computational requirements for tomographic inversions of a synthetic dataset; however, the effectiveness of their approach on real data was diminished by missing data. Vigh and Starr (2008) synthesized plane-wave gathers for acoustic FWI using a deterministic form of time-shift encoding. Krebs et al. (2009) generalized source encoding for FWI and presented two significant results. The first, was the adoption of polarity-based encoding functions that had the advantageous property of not increasing simulation time, unlike phase/time-shift based encoding schemes. The second, was that cross-talk artefacts could be almost entirely eliminated from inverted models by randomizing the encoding functions at each iteration. Subsequent studies in source encoded migration/FWI (SEFWI) have explored the properties of cross-talk (Schuster et al., 2011; Ben-Hadj-Ali et al., 2011), strategies for non fixed-spread receivers in marine environments (Routh et al., 2011; Choi and Alkhalifah, 2012), the use of second-order optimization methods (Anagaw and Sacchi, 2014; Castellanos et al., 2015), and stochastic optimization methods in place of source encoding (Haber et al., 2012; van Leeuwen and Herrmann, 2013a). Thus far, applications of source encoding to migration/FWI have focused on mono-parameter inversion under the acoustic approximation. Growing interests in multi-parameter FWI warrant the exploration of source encoding techniques applied to multi-parameter FWI.

The estimation of multiple independent parameters with limited data poses a significant challenge in multi-parameter FWI. Perturbations to different physical properties in the Earth's subsurface can give rise to similar responses in the data. This introduces trade-off between the parameters during inversion, making it difficult, or impossible to resolve independent parameters uniquely. The model, model parametrization, acquisition geometry, and data, all affect the resolvability of individual parameters (Tarantola, 1986; Pratt et al., 1998; Operto et al., 2013). Information on the resolution of model parameters is contained in the Hessian of the objective function. Parameter trade-offs can be partially corrected by using second-order optimization techniques (e.g. Newton, Gauss-Newton, truncated Newton) (Pratt et al., 1998; Operto et al., 2013). While second-order optimization methods have been explored in FWI (Epanomeritakis et al., 2008; Métivier et al., 2013; Anagaw and

Sacchi, 2014; Castellanos et al., 2015; Pan et al., 2016), they are computationally expensive and justifying their additional costs can be difficult. In some scenarios, quasi-Newton methods (e.g. L-BFGS) can achieve comparable inversion results at a reduced cost (Métivier et al., 2013). Schemes that approximate the Hessian have also been developed in an effort to mitigate parameter trade-off (e.g. Tang and Lee (2015); Pan et al. (2018)). In lieu of incorporating the Hessian, parameter trade-off can be limited through a combination of data-driven inversion strategies (Shipp and Singh, 2002; Sears et al., 2008; Prieux et al., 2013a,b) and appropriate model parametrization (Tarantola, 1986; Plessix and Cao, 2011; Köhn et al., 2012; Operto et al., 2013; Gholami et al., 2013; Alkhalifah and Plessix, 2014; Krebs et al., 2016). Understanding parameter trade-off is necessary for accurate model appraisal, particularly when first-order optimization methods are used.

In this chapter, we investigate source encoding in multi-parameter FWI with applications presented for isotropic elastic full waveform inversion. Emphasis is placed on understanding the influence of source encoding on the inversion of multiple parameters. Specifically, we seek to determine how source encoding affects parameter trade-off in multi-parameter inversion. While our study focuses on the isotropic, elastic case, our treatment is independent of a specific parametrization so as to be applicable to the general case of multi-parameter inversion. Our main contributions consist of three distinct areas: (i) An analysis of the source-encoded Hessian and its implications on parameter trade-off in general multi-parameter SEFWI. (ii) Investigations on the efficiency gain and stability of multi-parameter SEFWI. (iii) Demonstrating a limitation of SEFWI that is not overcome with current solutions that are otherwise successful in similar scenarios. To our knowledge, an in-depth study on multi-parameter FWI with source encoding does not exist in the current literature.

The chapter is structured as follows. Section 2 provides a brief review of FWI, source-encoded FWI, and the optimization algorithms associated with either method. Section 3 introduces challenges associated with multi-parameter inversion. The multi-parameter Hessian, with and without source encoding, is examined to determine the influence of source encoding on parameter trade-off. Section 4 presents a series of numerical experiments catered towards testing specific components of source-encoded FWI. We present results describing the efficiency gain and parameter trade-off of SEFWI relative to FWI. Section 5 poses a marine OBC experiment to convey a limitation of SEFWI when presented with data-driven inversion schemes that require time-windowing on the data. Conclusions from the study are presented in section 6.

3.2 Theory

Full waveform inversion can be formulated mathematically as a PDE-constrained optimization problem of the form (Plessix, 2006):

$$\begin{aligned} &\underset{\mathbf{m}}{\text{minimize}} \quad J(\mathbf{m}), \\ &\text{subject to} \quad \mathbf{L}(\mathbf{m})\mathbf{u}(\mathbf{x},t) = \mathbf{s}(\mathbf{x},t), \end{aligned} \tag{3.1}$$

where the functional $J(\mathbf{m})$ is dependent on model parameters $\mathbf{m}(\mathbf{x})$. Time is denoted by $t \in [0,T]$ and $\mathbf{x} \in \Omega \subset \mathbb{R}^d$ denotes spatial coordinates with dimensions $d = 1,2,3$. The linear differential operator $\mathbf{L}(\mathbf{m})$ characterizes the seismic wave equation and can accommodate varying degrees of physical complexity e.g. acoustic, elastic, isotropic/anisotropic etc. The particle displacement $\mathbf{u}(\mathbf{x},t)$ is excited by an external source $\mathbf{s}(\mathbf{x},t)$. For the sake of brevity, we omit the spatial and temporal dependencies of variables after they are first introduced, provided that no ambiguities arise from the omission. The model parameters $\mathbf{m}(\mathbf{x}) = [m_1(\mathbf{x}), m_2(\mathbf{x}), \ldots, m_{N_p}(\mathbf{x})]^T$, represent N_p independent physical properties of the Earth's subsurface; T denotes the transpose. The physics incorporated into the forward modelling operator $\mathbf{L}(\mathbf{m})$ dictate the physical properties comprised in $\mathbf{m}$. In this study, $\mathbf{L}(\mathbf{m})$ refers to the isotropic elastic wave equation in the time domain (Aki and Richards, 2002),

$$\mathbf{L}(\rho,\lambda,\mu) = \rho(\mathbf{x})\frac{\partial^2}{\partial t^2}[\cdot] - \nabla\cdot[\lambda(\nabla\cdot[\cdot])\mathbf{I} + \mu(\nabla[\cdot] + \nabla[\cdot]^T)], \tag{3.2}$$

where $[\cdot]$ is a place-holder for the variable acted upon by $\mathbf{L}(\mathbf{m})$ and $\mathbf{I}$ is the identity operator. The spatial gradient operator is denoted by ∇. The isotropic elastic wave equation is parametrized in terms of density $\rho(\mathbf{x})$ and the Lamé parameters $\lambda(\mathbf{x})$ and $\mu(\mathbf{x})$. The particular choice of $\mathbf{L}(\mathbf{m})$ in Equation 3.2 does not lead to a loss of generality in the forthcoming discussions on FWI and multi-parameter source-encoded FWI.

The misfit functional $J(\mathbf{m})$ —also known as the cost or objective function— quantifies the difference between observed and synthetic data by comparing an observable quantity. The most prevalent choice of objective function for FWI is the least-squares waveform misfit functional

$$J(\mathbf{m}) = \frac{1}{2}\sum_{s=1}^{N_s}\sum_{r=1}^{N_r}\int_T |\mathbf{u}_s(\mathbf{x}_r,t;\mathbf{m}) - \mathbf{d}_s(\mathbf{x}_r,t)|^2 \, \mathrm{d}t. \tag{3.3}$$

The simulated multi-component data $\mathbf{u}_s(\mathbf{x}_r,t;\mathbf{m})$ are recorded at the *r-th* receiver and generated by the *s-th* source $\mathbf{s}_s$ for model $\mathbf{m}$. A similar definition is applicable for the observed data $\mathbf{d}_s(\mathbf{x}_r,t)$. The number of sources and receivers are denoted by N_s and N_r, respectively. The least-squares waveform misfit is a non-linear functional owing to the quadratic term in (Equation 3.3) and the non-linear dependence of $\mathbf{u}$ on $\mathbf{m}$ (Virieux and

Operto, 2009).

Non-linear functionals can be minimized via iterative gradient-based minimization algorithms (Plessix, 2006). Lailly (1983) and Tarantola (1984a) demonstrated this approach in the context of seismic imaging. Solutions to Equation 3.1 can be estimated by iteratively updating the model parameters via

$$\mathbf{m}^{k+1} = \mathbf{m}^k + \nu^k \delta\mathbf{m}^k, \tag{3.4}$$

where k denotes the iteration number, ν^k is a scalar step length, and the model perturbation/update is $\delta\mathbf{m}^k(\mathbf{x}) = [\delta m_1(\mathbf{x}), \delta m_2(\mathbf{x}), \ldots, \delta m_{N_p}(\mathbf{x})]^T$. A suitable ν_k can be estimated using various line-search algorithms (Nocedal and Wright, 2006). In optimization literature, the model update $\delta\mathbf{m}^k$ is referred to as the search or descent direction and can be derived from the gradient of the objective function with respect to the model parameters.

FWI gradient

The gradient of $J(\mathbf{m})$ with respect to $\mathbf{m}$, $\nabla_m J$, can be calculated efficiently using the adjoint-state method; for a complete description of the method, the reader is referred to Tarantola (1986); Mora (1987); Tromp et al. (2005); Plessix (2006); Fichtner et al. (2006). Solving the adjoint-state equation (appendix A) yields the adjoint wavefield $\mathbf{u}^\dagger(\mathbf{x}, t)$. Given $\mathbf{u}$ and $\mathbf{u}^\dagger$, the time-domain expression for the derivative of J with respect to $\mathbf{m}$ is

$$\nabla_m J(\mathbf{x}) = -\sum_{s=1}^{N_s} \int_T \mathbf{u}_s^\dagger(\mathbf{x}, t) \cdot \frac{\partial \mathbf{L}}{\partial \mathbf{m}} \mathbf{u}_s(\mathbf{x}, t)\, \mathrm{d}t. \tag{3.5}$$

Computation of $\nabla_m J$ requires N_s forward simulations and N_s adjoint simulations at each iteration. The computational cost of an FWI iteration therefore grows linearly with N_s. This linear dependence becomes prohibitive when N_s is large. Source encoding effectively reduces the number of sources by considering multiple sources simultaneously rather than independently.

3.2.1 Source-encoded FWI

The linear dependence of $\mathbf{u}$ with respect to $\mathbf{s}$ permits a reformulation of Equation 3.3 to accommodate simultaneous sources (Krebs et al., 2009). The revised misfit functional assumes a fixed-spread acquisition (i.e. fixed receiver positions) and may be expressed as

$$\hat{J}(\mathbf{m}) = \frac{1}{2} \sum_{e=1}^{N_e} \sum_{r=1}^{N_r} \int_T |\hat{\mathbf{u}}_e(\mathbf{x}_r, t; \mathbf{m}) - \hat{\mathbf{d}}_e(\mathbf{x}_r, t)|^2 \, \mathrm{d}t, \tag{3.6}$$

where $\hat{\mathbf{d}}_e$ are the encoded data and $\hat{\mathbf{u}}_e$ are the synthetic data generated by encoded source $\hat{\mathbf{s}}_e$. The number of encoded sources N_e is selected such that $N_e < N_s$. We use a circumflex to signify source-encoded variables or those associated with SEFWI. Source encoding reduces the data volume by a factor N_s/N_e, thereby reducing the number of PDE solves required per iteration. Maximal data compression is achieved when every individual source is combined into a single encoded source.

At this stage, we define a selection criterion used to select individual sources for encoding. Let S denote the set containing all the sources in a given acquisition. We synthesize N_e encoded sources from N_e mutually disjoint subsets of S. Formally, $S = \cup_{i=1}^{N_e} S_i$ where $S_i \cap S_j = \emptyset,\ \{\forall i, j = 1, \ldots, N_e, i \neq j\}$. Mutually disjoint subsets ensure that individual sources are not repeated over multiple encoded sources. This restriction is not essential to the formulation of SEFWI; however, we impose it to simplify the forthcoming treatment of SEFWI. Following this definition, $\hat{\mathbf{s}}_e$ and $\hat{\mathbf{d}}_e$ are formed from the linear combinations,

$$\hat{\mathbf{s}}_e(\mathbf{x}, t) = \sum_{s \in S_e} q_s^e(t) * \mathbf{s}_s(\mathbf{x}, t), \tag{3.7}$$

$$\hat{\mathbf{d}}_e(\mathbf{x}, t) = \sum_{s \in S_e} q_s^e(t) * \mathbf{d}_s(\mathbf{x}, t), \tag{3.8}$$

where $q_s^e(t)$ are source-specific encoding functions for the *e-th* encoded source. Convolution in the time domain is denoted by $*$. Encoding functions are discussed in section 2.2.

SEFWI gradient

The SEFWI gradient is also computed using the adjoint-state method. Source encoding introduces cross-talk artefacts into the gradient that are a consequence of zero-lag correlations (Equation 3.5) between forward and adjoint wavefields that do not correspond to the same source. For the case of pure simultaneous sources, i.e. if $q_s^e(t) = 1\ \{\forall s \in S_e, e = 1, \ldots, N_e\}$, the derivative of $\hat{J}$ with respect to $\mathbf{m}$ is

$$\nabla_m \hat{J}(\mathbf{x}) = -\sum_{e=1}^{N_e} \int_T \hat{\mathbf{u}}_e^\dagger(\mathbf{x}, t) \cdot \frac{\partial \mathbf{L}}{\partial \mathbf{m}} \hat{\mathbf{u}}_e(\mathbf{x}, t)\ \mathrm{d}t, \tag{3.9}$$

$$= \nabla_m J(\mathbf{x}) - \underbrace{\sum_{i=1}^{N_s} \sum_{\substack{j=1 \\ j \neq i}}^{N_s} \int_T \mathbf{u}_j^\dagger(\mathbf{x}, t) \cdot \frac{\partial \mathbf{L}}{\partial \mathbf{m}} \mathbf{u}_i(\mathbf{x}, t)\ \mathrm{d}t}_{\text{Cross-talk term}}. \tag{3.10}$$

The simplification in Equation 3.10 is valid when subsets S_i are mutually disjoint. The second term on the right hand side of Equation 3.10 represents cross-talk artefacts that compromise the accuracy of the desired gradient. For pure simultaneous sources, the cross-talk

artefacts stack as coherent noise over the course of SEFWI iterations resulting in inaccurate models (Romero et al., 2000; Krebs et al., 2009). The influence of cross-talk artefacts can be ameliorated via source encoding.

3.2.2 Source encoding

The frequency-domain representation of the SEFWI gradient allows the role of encoding functions to be more readily understood. The frequency-domain formulation of $\nabla_m \hat{J}$ with general source-encoding is

$$\nabla_m \hat{J}(\mathbf{x}) = -\sum_{e=1}^{N_e} \left[\sum_{i \in S_e} \left\langle Q_i^e(\omega) U_i{}^{\dagger}(\mathbf{x},\omega),\ Q_i^e(\omega) \frac{\partial \mathbf{L}}{\partial \mathbf{m}} U_i(\mathbf{x},\omega) \right\rangle_{\omega} + \underbrace{\sum_{i \in S_e} \sum_{\substack{j \in S_e \\ j \neq i}} \left\langle Q_i^e(\omega) U_i{}^{\dagger}(\mathbf{x},\omega),\ Q_j^e(\omega) \frac{\partial \mathbf{L}}{\partial \mathbf{m}} U_j(\mathbf{x},\omega) \right\rangle_{\omega}}_{\text{Cross-talk term}} \right], \tag{3.11}$$

where $Q_i^e(\omega) = \mathcal{F}\{q_i^e(t)\}$, $U(\mathbf{x},\omega) = \mathcal{F}\{\mathbf{u}(\mathbf{x},t)\}$, $U^{\dagger}(\mathbf{x},\omega) = \mathcal{F}\{\mathbf{u}^{\dagger}(\mathbf{x},t)\}$, and $\mathcal{F}$ is the Fourier transform operator. The inner product $\langle \cdot, \cdot \rangle_{\omega}$ between two arbitrary complex-valued functions f and g, is defined as

$$\langle f, g \rangle_{\omega} := \int_{\omega} \bar{f}(\omega) g(\omega)\ \mathrm{d}\omega. \tag{3.12}$$

Complex conjugation is denoted by a bar above a variable. The origin of Equation 3.11 is available in appendix B. When the encoding functions form an orthonormal basis i.e. $\bar{Q}_i^e(\omega) Q_j^e(\omega) = \delta_{ij}$, $\{\forall i, j \in S_e\}$, Equation 3.11 reduces to the standard FWI gradient. In practice, we seek random encoding functions with the property

$$\mathbb{E}[\bar{Q}_i^e(\omega) Q_j^e(\omega)] = \delta_{ij}. \tag{3.13}$$

Equation 3.13 states that the expected inner product between any two random encoding functions is a Kronecker delta function. Previous studies have used this condition to establish potential encoding functions. Suitable choices include: random time-shifts (Romero et al., 2000; Schuster et al., 2011), plane-wave encoding (Vigh and Starr, 2008), and polarity encoding (Krebs et al., 2009). Polarity encoding uses encoding functions of the form $Q_i(\omega) = r_i$, where the discrete random variable r_i takes values of +1 or -1 with equal probability.

Cross-talk artefacts can be further reduced by randomizing the encoding functions at regular intervals. Altering the source encoding after every iteration has been demonstrated to be

optimal (Krebs et al., 2009). Randomizing the encoding functions results in the cross-talk terms stacking as incoherent noise over the course of SEFWI.

3.2.3 Gradient-based optimization

In the vicinity of an initial model $\mathbf{m}^0$, $J(\mathbf{m}^0)$ can be approximated by a locally quadratic function following a second-order Taylor expansion (Pratt et al., 1998). The perturbation $\delta\mathbf{m}$ that minimizes the quadratic approximation is obtained by solving the Newton system of equations, represented symbolically as

$$\mathbf{H}\delta\mathbf{m} = -\mathbf{g}, \tag{3.14}$$

where $\mathbf{g} = \nabla_m J$ and $\mathbf{H} = \nabla_m^2 J$ denote the gradient and Hessian of $J(\mathbf{m}^0)$, respectively. Henceforth, we adopt letter symbols for the FWI/SEFWI gradients and Hessians in favour of readability. The multi-parameter Hessian is explored in section 3.1.

In this study we consider first-order techniques only; namely, steepest descent (SD), nonlinear conjugate gradients (NLCG), and the quasi-Newton L-BFGS method. L-BFGS generates approximations to the inverse Hessian from previous gradients (Liu and Nocedal, 1989). First-order algorithms neglect the Hessian and rely solely on gradient information to generate search directions. Details for each algorithm can be found in Nocedal and Wright (2006).

Optimization algorithms for source-encoded FWI

The source-encoded Hessian $\hat{\mathbf{H}} = \nabla_m^2 \hat{J}$ can be altered considerably when the source encoding is randomized. Applications of NLCG and L-BFGS should be amended to account for this. In NLCG, the conjugacy condition cannot be guaranteed and therefore search directions are not assured to be conjugate pairs. Moghaddam et al. (2013) proposed a heuristic alternative that formed search directions as a weighted sum of prior gradients. Their approach demonstrated higher convergence rates relative to SD when applied to acoustic SEFWI.

We implement hybrid forms of NLCG and L-BFGS that have been utilized for source-encoded migration/FWI in the acoustic case (Huang and Schuster, 2012; Castellanos et al., 2015). The hybrid algorithms amount to regular-restart versions of their conventional counterparts. After every *M-th* iteration, the optimization history is reset and the source encoding is randomized. The source encoding does not vary between restart intervals. As the hybrid algorithms do not randomize the source encoding at every iteration, cross-talk is expected to be more prominent for these algorithms. The restart interval should be chosen to

be as small as possible (to suppress cross-talk), but large enough to allow the optimization algorithm to perform effectively. For example, NLCG requires gradients from the current and prior iteration, so the restart interval should be at least 2. The L-BFGS method generates search directions from M previous gradients, with M typically ranging between 3-15 (Liu and Nocedal, 1989). The performance of L-BFGS is problem dependent and also varies with M (Nocedal and Wright, 2006), making it difficult establish a guideline for selecting the restart interval. Extensive tests to assess the performance of the hybrid schemes for various restart intervals have not been conducted. In this study, we deploy SEFWI with SD and restart variants of NLCG and L-BFGS. We do not differentiate between regular and restart versions of NLCG/L-BFGS in the text. The reader may assume that NLCG/L-BFGS applied to SEFWI corresponds to the restart versions described in this section.

Model regularization

FWI is an ill-posed problem meaning an infinite number of models can fit the data equally well (Virieux and Operto, 2009). Model regularization, included explicitly into the objective function, serves to stabilize the inversion and make it more well-posed. Furthermore, model regularization constrains updates by imposing prior assumptions on the model. In this study, we implement a form of Tikhonov regularization that penalizes deviatoric perturbations from a prior model,

$$R(\mathbf{m}) = \frac{\gamma}{2} \sum_{p}^{N_p} \|m_p - m_p^{prior}\|^2. \tag{3.15}$$

A tunable hyperparameter γ controls the contribution of the regularization term relative to the data misfit. The prior model in Equation 3.15 is taken as the initial model input to FWI/SEFWI.

3.3 Multi-parameter inversion

Prior applications of source encoding in migration and FWI have primarily focused on single parameter inversion under the constant-density, acoustic approximation (Romero et al., 2000; Vigh and Starr, 2008; Krebs et al., 2009; Schuster et al., 2011; Dai et al., 2012; Anagaw and Sacchi, 2014; Castellanos et al., 2015). While Capdeville et al. (2005) performed elastic inversion, it was restricted to a single-parameter (S-wave velocity). Krebs et al. (2016) carried out a synthetic parameter resolution study for FWI in tilted transversely-isotropic visco-elastic media. The study utilized source encoding to improve the efficiency of inversions, but did not discuss the relationship between source encoding and parameter trade-off. When considering anisotropic or elastic representations of the Earth, multiple independent

model parameters are required to characterize the subsurface. An isotropic elastic medium is adequately described by 3 independent parameters; a potential parametrization is in terms of density (ρ) and the Lamé parameters (λ, μ).

An ideal model parametrization consists of a set of physical parameters that are uniquely resolvable. The extent to which a model perturbation can be resolved uniquely, is dictated by the choice of model parametrization, acquisition geometry, background model, and bandwidth of the data (Tarantola, 1986; Plessix and Cao, 2011; Operto et al., 2013; Gholami et al., 2013; Alkhalifah and Plessix, 2014). A poor choice of model parametrization or inadequate subsurface illumination can lead to ambiguities between different parameters (Plessix and Cao, 2011; Operto et al., 2013). Parameter trade-off is the phenomena where changes in different parameters elicit similar responses in the data. A classic example is the velocity-depth ambiguity associated with reflection travel times. To further complicate matters, perturbations to parameters that lie in the null space of the problem will not register in the data making them unresolvable (Alkhalifah and Plessix, 2014).

Within a given parametrization, certain parameters have a greater influence on the data than others. A good parametrization prioritises the accurate reconstruction of parameters that most strongly influence the kinematics of the data (e.g. Plessix and Cao (2011); Gholami et al. (2013); Alkhalifah and Plessix (2014)). For example, P-wave velocity controls the kinematics of compressional waves, whereas density primarily influences reflection amplitudes. A parametrization that allows for the broadband reconstruction of P-wave velocity should be favoured. Tarantola (1986) compared scattering patterns derived from the Born approximation to assess parameter trade-off and resolution. The study concluded that a parametrization of density (ρ), P-wave velocity (α), and S-wave velocity (β) was suitable for broadband reconstruction of α while limiting parameter trade-offs.

3.3.1 Multi-parameter Hessian

Parameter trade-offs manifest mathematically in the multi-parameter Hessian. The Hessian carries information pertaining to the strength of parameter trade-offs along with the spatial resolution afforded by the acquisition geometry. Neglecting the Hessian in multi-parameter inversion introduces inaccuracies into the inversion due to erroneous inter-parameter mappings (Operto et al., 2013). The multi-parameter Hessian operator exhibits a block structure and may be expressed in matrix form as

$$\mathbf{H}(\mathbf{x},\mathbf{y}) = \begin{bmatrix} \mathbf{H}_{m_1 m_1}(\mathbf{x},\mathbf{y}) & \dots & \mathbf{H}_{m_1 m_p}(\mathbf{x},\mathbf{y}) \\ \vdots & \ddots & \\ \mathbf{H}_{m_p m_1}(\mathbf{x},\mathbf{y}) & & \mathbf{H}_{m_p m_p}(\mathbf{x},\mathbf{y}) \end{bmatrix}. \tag{3.16}$$

The Newton equations in terms of the multi-parameter Hessian operator are

$$\sum_{j=1}^{N_p} \int \mathbf{H}_{m_i m_j}(\mathbf{x}, \mathbf{y}) \delta m_j(\mathbf{y}) \, \mathrm{d}\mathbf{y} = -g_i(\mathbf{x}). \tag{3.17}$$

Equation 3.17 states that the gradient for the *i-th* model parameter is a linear combination of the true model perturbations weighted by the relevant block elements from the Hessian. Due to the expense of Newton based methods, the Hessian in Equation 3.17 is often replaced with a diagonal preconditioning operator to give

$$\int \mathbf{P}_{m_i m_i}(\mathbf{x}, \mathbf{y}) \delta m_i(\mathbf{y}) \, \mathrm{d}\mathbf{y} \approx -g_i(\mathbf{x}). \tag{3.18}$$

The lack of off-diagonal contributions ($i \neq j$) in $\mathbf{P}$ introduces inter parameter mappings that lead to inaccuracies in inverted models. When inter-parameter mappings are not corrected, it is important to understand their nature for proper model appraisal. A concern arises when using first-order optimization for SEFWI. The cross-talk artefacts in the SEFWI gradient will map across multiple parameters manifesting as additional parameter trade-offs. The source-encoded, multi-parameter Hessian is examined in the following section to better understand the behaviour of parameter trade-off in SEFWI.

3.3.2 Source-encoded multi-parameter Hessian

In the frequency domain, the source-encoded Gauss-Newton Hessian ($\hat{\mathbf{H}}^a$) can be written as

$$\hat{\mathbf{H}}^a(\mathbf{x}, \mathbf{y}) = \sum_{e=1}^{N_e} \sum_{r=1}^{N_r} \left\langle \sum_{i \in S_e} Q_i^e(\omega) \frac{\partial \mathbf{u}_i(\mathbf{x}_r, \omega)}{\partial \mathbf{m}(\mathbf{x})}, \sum_{j \in S_e} Q_j^e(\omega) \frac{\partial \mathbf{u}_j(\mathbf{x}_r, \omega)}{\partial \mathbf{m}(\mathbf{y})} \right\rangle_\omega, \tag{3.19}$$

where $\frac{\partial \mathbf{u}_i(\mathbf{x}_r, \omega)}{\partial \mathbf{m}(\mathbf{x})}$ are the Fourier transformed Fréchet derivatives. Equation 3.19 can be simplified to

$$\hat{\mathbf{H}}^a(\mathbf{x}, \mathbf{y}) = \mathbf{H}^a + \sum_{e=1}^{N_e} \sum_{r=1}^{N_r} \sum_{i \in S_e} \sum_{\substack{j \in S_e \\ j \neq i}} \left\langle Q_i^e(\omega) \frac{\partial \mathbf{u}_i(\mathbf{x}_r, \omega)}{\partial \mathbf{m}(\mathbf{x})}, Q_j^e(\omega) \frac{\partial \mathbf{u}_j(\mathbf{x}_r, \omega)}{\partial \mathbf{m}(\mathbf{y})} \right\rangle_\omega. \tag{3.20}$$

Once again we assumed that the subsets S_i are mutually disjoint. A short derivation for Equation 3.20 is presented in appendix C. The cross-talk terms in the source-encoded Hessian are comparable to those in the source-encoded gradient (Equation 3.11). This implies that the cross-talk in $\hat{\mathbf{H}}^a$ can also be attenuated by selecting orthonormal encoding functions

(Equation 3.13). The symbolic representation of the Newton equations in SEFWI is,

$$(\mathbf{H}^a + \mathbf{H}^c)\delta\mathbf{m} = -(\mathbf{g} + \mathbf{g}^c), \tag{3.21}$$

where $\mathbf{H}^c$ and $\mathbf{g}^c$ are the cross-talk components of the source-encoded Hessian and gradient, respectively. The operator $\mathbf{H}^c$ maps model perturbations into $\mathbf{g}^c$. If $\delta\mathbf{m}$ is computed using first-order gradient techniques, i.e. by neglecting $\hat{\mathbf{H}}^a$, estimates of $\delta\mathbf{m}$ will exhibit erroneous inter-parameter mappings associated with $\mathbf{g}^c$. To verify that source encoding can be used to suppress cross-talk in the Hessian, we perform an analysis that involves probing the multi-parameter Hessian.

3.3.3 Hessian probing

In the vicinity of the true model, the resolvability of a model perturbation can be assessed by computing

$$\mathbf{H}^{-g}\mathbf{H}\delta\mathbf{m}_{true} = \delta\mathbf{m} \tag{3.22}$$

where $\mathbf{H}^{-g}$ is the generalized inverse of the Hessian, $\delta\mathbf{m}_{true}$ is a true model perturbation, and $\delta\mathbf{m}$ is an estimated model perturbation (Fichtner and Trampert, 2011b). The term $\mathbf{H}^{-g}\mathbf{H}$ acts as a resolution operator and $\mathbf{H}^{-g}\mathbf{H}\delta\mathbf{m}_{true}$ is a point spread function (PSF) that describes how model perturbations are smeared in space. On its own, $\mathbf{H}$ can be viewed as a conservative approximation to the true resolution operator. Likewise, $\mathbf{H}\delta\mathbf{m}_{true}$ provides an estimate of the true PSF (Fichtner and Leeuwen, 2015). Henceforth, we use the term PSF to refer to the approximate PSF $\mathbf{H}\delta\mathbf{m}$. PSFs can be interpreted as resolution proxies in the vicinity of the true model or, more generally, as weighted row averages of $\mathbf{H}$ in the discrete case. While the Hessian itself is expensive to compute, Hessian-vector products can be computed efficiently using second-order adjoints (Fichtner and Trampert, 2011a) or finite-difference approximations (Zhu et al., 2015). We use the finite-difference approach as it is more convenient to implement with our current code.

For fixed $\delta\mathbf{m}$, we define a realization of $\hat{\mathbf{H}}\delta\mathbf{m}$ as the PSF computed for a particular set of random encoding functions. When the encoding functions satisfy Equation 3.13, the expected PSF in SEFWI satisfies

$$\mathbb{E}[\hat{\mathbf{H}}\delta\mathbf{m}] = \mathbb{E}[\hat{\mathbf{H}}]\delta\mathbf{m} = \mathbf{H}\delta\mathbf{m}. \tag{3.23}$$

The convergence of $\hat{\mathbf{H}} \rightarrow \mathbf{H}$ was noted by Tang (2009) and used for efficient access to the Hessian in mono-parameter acoustic migration/FWI. The expected PSF is estimated from

the ensemble average taken over N random realizations of $\hat{\mathbf{H}}\delta\mathbf{m}$,

$$\mathbb{E}[\hat{\mathbf{H}}\delta\mathbf{m}] \approx \frac{1}{N}\sum_{i=1}^{N}\hat{\mathbf{H}}^{(i)}\delta\mathbf{m}. \tag{3.24}$$

Schuster et al. (2011) defined the cross-talk signal-to-noise ratio (c-SNR) as a measure of cross-talk noise in source-encoded migration images. An analogous definition for Hessian-vectors products is given by

$$\text{c-SNR} = \frac{\|\mathbf{H}\delta\mathbf{m}\|}{\|\sum_{i=1}^{N}\hat{\mathbf{H}}^{(i)}\delta\mathbf{m} - \mathbf{H}\delta\mathbf{m}\|}. \tag{3.25}$$

The L_2 norm ($\|\cdot\|$) follows a standard definition

$$\|\psi\| := (\int_{\Omega}|\psi(\mathbf{x})|^2\ \mathrm{d}\mathbf{x})^{\frac{1}{2}}, \tag{3.26}$$

for an arbitrary real-valued function $\psi(\mathbf{x})$. Schuster et al. (2011) demonstrated that the c-SNR grows $\propto \sqrt{N}$. The effect of stacking random realizations is mimicked by randomizing encoding functions at each iteration of SEFWI.

As a numerical test, we compute expected PSFs for perturbations applied to a homogeneous model. The test focuses on the two parameter case, where $\mathbf{m}(\mathbf{x}) = [\alpha(\mathbf{x}), \beta(\mathbf{x})]^T$. 16 sources and 50 receivers are evenly distributed along the surface of a model that is discretized on a 100 x 100 grid with a spacing of 10 m. We use model perturbations of the form

$$\delta\mathbf{m}(\mathbf{x}) = [\delta\alpha(\mathbf{x}), \delta\beta(\mathbf{x})]^T = \begin{cases} [c, 0]^T, & \text{for } \mathbf{x} = \mathbf{x}_c \\ [0, 0]^T, & \text{otherwise} \end{cases}$$

where $\mathbf{x}_c$ is the central grid point and c is a constant value, taken as 1% of the background model in this example. Perturbations are applied to one parameter at a time allowing us to target individual block elements of the Hessian operator (Equation 3.16). We compute similar PSFs with the sequential-source Hessian for reference.

Figure 3.1 depicts PSFs associated with $\mathbf{H}$ and $\hat{\mathbf{H}}$. For a small number of random realizations, prominent cross-talk artefacts are apparent in the expected PSFs. As the number of random realizations increases, $\mathbb{E}[\hat{\mathbf{H}}\delta\mathbf{m}]$ increasingly resembles $\mathbf{H}\delta\mathbf{m}$. For 64 realizations, the expected PSF is almost identical to the reference PSF. We notice some spurious oscillations that persist in $\mathbf{H}_{\alpha\beta}$ (Figure 3.1e) even after 64 realizations. These oscillations are attributed to boundary related artefacts that stem from the relatively small grid size. Figure 3.2 displays the growth of c-SNR as a function of random realizations in an ensemble. The PSFs associated with each block component of $\hat{\mathbf{H}}$ exhibit similar convergence behaviour and

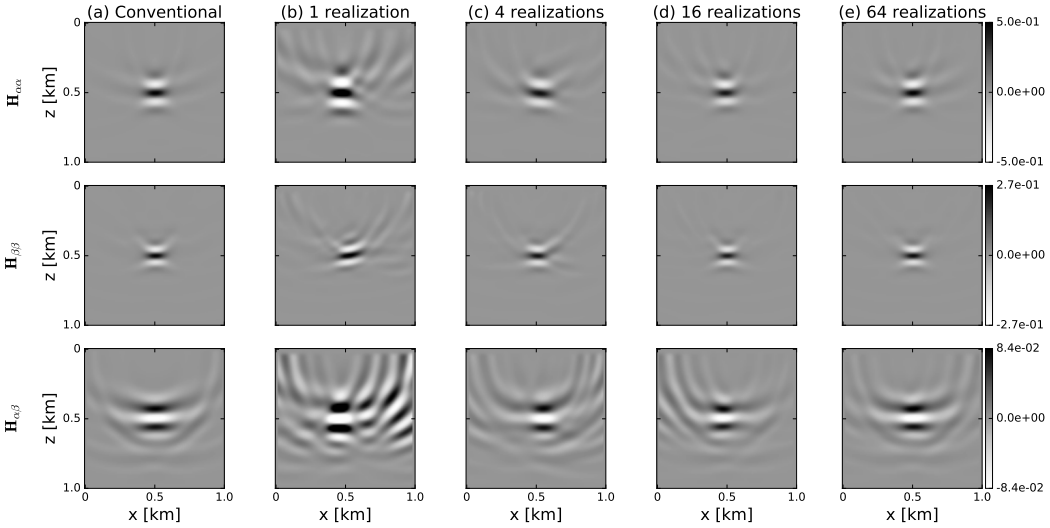

Figure 3.1: Point spread functions for a point scatterer ($x = z = 0.5$ km) generated using different block components of the Hessian (with and without source encoding): (top row) $\mathbf{H}_{\alpha\alpha}$, (middle row) $\mathbf{H}_{\beta\beta}$, (bottom row) $\mathbf{H}_{\alpha\beta}$. (a) PSFs computed using the Hessian without source encoding. (b-e) Ensemble averaged PSFs for a varying number of realizations of the source-encoded Hessian. With an increasing number of realizations the cross-talk artefacts are suppressed and the expected PSFs approach the equivalent PSF obtained without source encoding.

approximately follow the $\propto \sqrt{N}$ relation defined by Schuster et al. (2011). The bias in the $\mathbf{H}_{\alpha\beta}$ PSF is attributed to the propagation of numerical errors mentioned previously.

Certain conclusions can be drawn from Figs. 3.1 and 3.2. Since the cross-talk terms in the Hessian can be attenuated through SEFWI iterations, the parameter trade-off in SEFWI should be comparable to that of FWI. Furthermore, as the cross-talk terms in $\hat{\mathbf{H}}$ are attenuated at the same rate as those in $\hat{\mathbf{g}}$, the number of iterations need not be amended to correct for inter-parameter mappings of cross-talk noise. Interpreting PSFs as resolution proxies corroborates the notion that FWI and SEFWI have comparable resolution.

3.3.4 Trade-off and the number of inversion parameters

The property in Equation 3.23 suggests that SEFWI should have the same sensitivity to model parametrization as FWI. As the expression is not restricted to a particular parametrization, the previous statement should be true for any model parametrization, including when a change in parametrization occurs due to changes in the modelling physics. The characteristics of parameter trade-off can be affected by the inversion scheme as well as changes to the model parametrization. Multi-parameter FWI can invert for independent model parameters sequentially, simultaneously, or through some combination of the two. While simultaneous inversion reduces the overall number of inversions required, sequential schemes may be favourable as they can be designed to mitigate non-linearities in FWI. Sequential inversion strategies typically impose hierarchy on the importance of individual model parameters. For example, parameters that influence the kinematics (e.g. P-wave velocity) of the data hold precedence over those that control the dynamics (e.g. density). By sequentially inverting parameters, or parameter subsets, inversions can insert increasingly complex features into the model in a progressive and controlled manner. Studies performing elastic inversion on ocean-bottom cable (OBC) data have used this approach with success (Shipp and Singh, 2002; Sears et al., 2008, 2010; Prieux et al., 2013b); a similar example is explored with source encoding in section 5. A drawback of sequential inversions is that they can introduce strong artefacts due to inter-parameter mappings. If trade-off artefacts are sufficiently strong, subsequent stages of the inversion may fail due to the inaccurate models recovered in earlier stages (Operto et al., 2013; Prieux et al., 2013b). Sequential inversion is a series of one-dimensional minimizations over the search space. A poor starting model or errors introduced by parameter trade-offs can result in the inversion converging to a local minimum.

The toy example presented in Figure 3.3 illustrates the variable behaviour of parameter trade-off with scheme. The true model contains 3 uncorrelated Gaussian anomalies in α, β, and ρ. Sources and receivers are placed along the boundaries of the model. All three

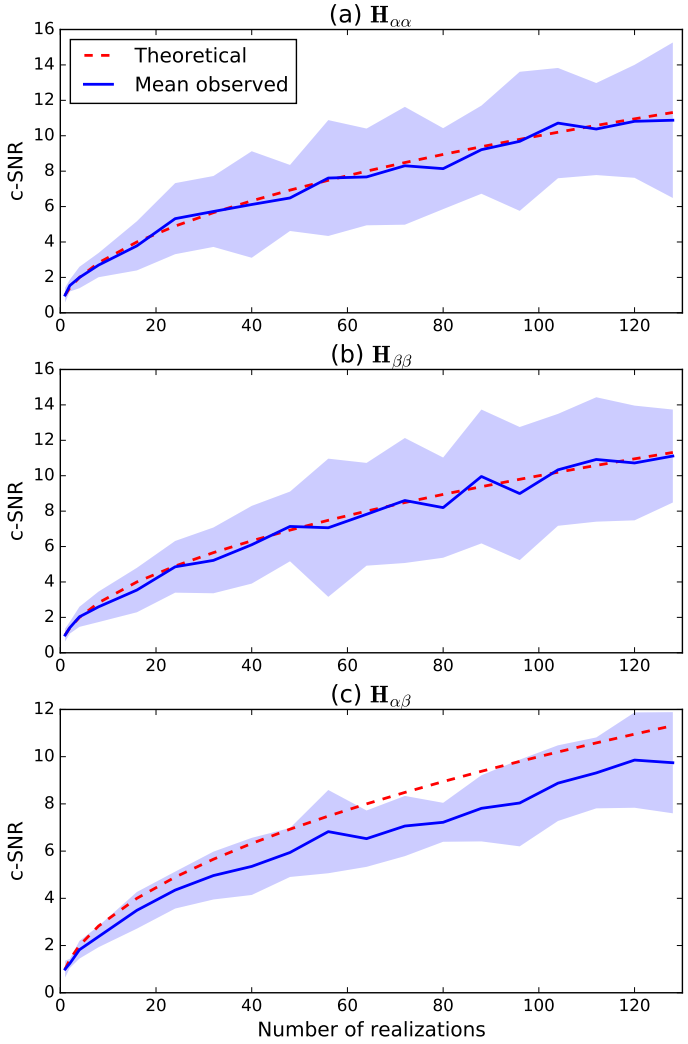

Figure 3.2: PSF c-SNR as a function of the number of random realizations in an ensemble. The mean c-SNR (solid blue line) for 20 independent trials is plotted with errors bars that represent one standard deviation. Each panel corresponds to the PSF c-SNR associated with a particular block component of the source-encoded Hessian: (a) $\mathbf{H}_{\alpha\alpha}$, (b) $\mathbf{H}_{\beta\beta}$, (c) $\mathbf{H}_{\alpha\beta}$. Mean c-SNR grows approximately $\propto \sqrt{N}$ (red dashed line) (Schuster et al., 2011). Each panel is normalized to have c-SNR=1 at the first iteration.

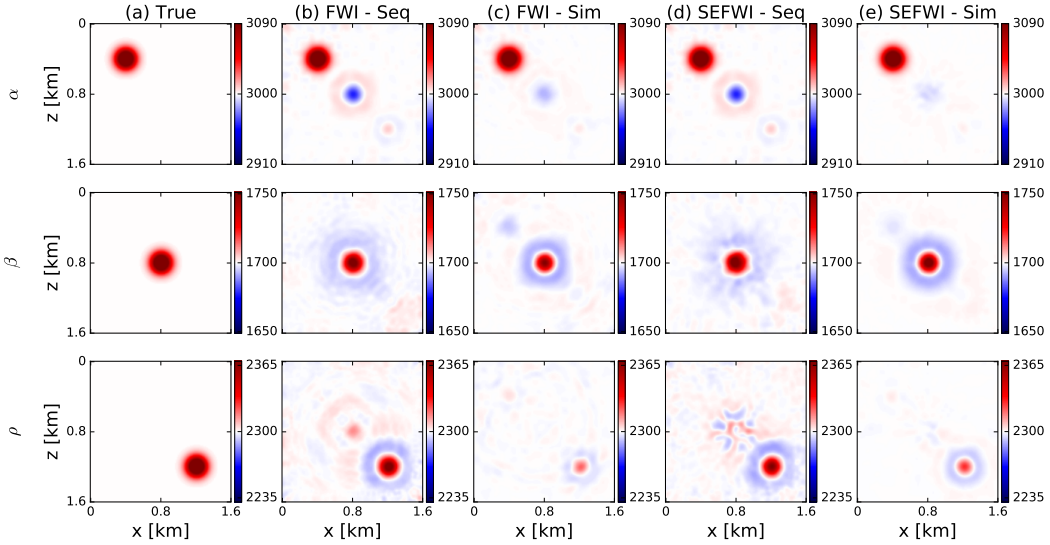

Figure 3.3: Inversion results for (b, d) sequential and (c, e) simultaneous inversion of multiple parameters using FWI and SEFWI. (a) True model containing three spatially inconsistent Gaussian anomalies in α, β, and ρ. (b) Sequential FWI model. (c) Simultaneous FWI model. (d) Sequential SEFWI model. (e) Simultaneous SEFWI model. Parameter trade-off varies depending on the inversion strategy. For a common strategy, FWI and SEFWI exhibit similar parameter trade-off.

parameters are inverted by applying FWI and SEFWI to two different inversion schemes. The first scheme employs a sequential inversion strategy whereas the second performs a simultaneous inversion. The sequential inversion inverts for α first, β second, and ρ last. Each successive stage of a sequential inversion uses the final model from the prior stage as an initial model. The parameter order reflects a progression used in real applications (Sears et al., 2008; Prieux et al., 2013b), but is otherwise subjective.

Synthetic inversions terminate after 100 iterations or when the line search (up to 15 trial steps) fails. When parameters are inverted simultaneously, both FWI and SEFWI reach the maximum number of iterations; the relative misfit reduction is $J(\mathbf{m}^{100})/J(\mathbf{m}^{0}) = 3 \times 10^{-3}$ for FWI. In sequential FWI trials, the α and ρ inversions terminate early due to line search failures; these are attributed to prominent trade-off artefacts in the inverted models. All SEFWI trials reach the maximum number of iterations. Increasing the maximum number of iterations does not result in the same inverted models for the simultaneous and sequential trials. This suggests that the two schemes converge to different points in the search space.

The variable parameter trade-offs are evident from a comparison of Figs. 3.3(b) and 3.3(c). In the sequential case, the inverted α model contains strong artefacts associated with the β and ρ anomalies. Similar artefacts appear in the simultaneous example, but with diminished

amplitudes. For the β and ρ models, the artefacts appear to be stronger in the simultaneous inversion case. Overall, the trade-off behaviour is consistent with eqs. (3.17) and (3.18). Given a common inversion scheme, FWI and SEFWI have comparable trade-off characteristics, as observable through a comparison of Figs. 3.3(b) and 3.3(d) along with Figs. 3.3(c) and 3.3(e).

3.4 Numerical experiments

We conduct a series of numerical experiments to interrogate specific components of the SEFWI algorithm. Initially, we test the efficiency gain offered by SEFWI when coupled with first-order optimization algorithms. In a subsequent test, we seek to verify the claim that parameter trade-off in SEFWI is comparable to FWI. Tests with noisy data and early termination are performed to test the stability of the algorithm.

3.4.1 Inversion procedure

The inversion procedure described in this section is applicable to every experiment unless stated otherwise. 'Observed' and synthetic data are generated using 2D time-domain, *P-SV* finite difference modelling (fourth order in space, second order in time) (Virieux, 1986; Levander, 1988). Convolutional perfectly matched layers are implemented to simulate absorbing boundaries at the edges of the numerical grid (Komatitsch and Martin, 2007). The free surface is replaced with an absorbing boundary to avoid generating surface waves in the data. Source inversion is not performed and the true source wavelet is assumed to be known. We acknowledge that this is not realistic but we make the assumption to reduce the number of variables in the inversions.

Elastic models are parametrized in terms of seismic velocities and density. Density is not included as an inversion parameter and is updated empirically via Gardener's relation ($\rho = 310\alpha^{0.25}$) (Gardner et al., 1974), where appropriate. Nondimensionalization is applied to the inversion parameters via rescaling of the form $m'_p = m_p/m_0$ (Prieux et al., 2013a). The scaling values m_0 are taken as the mean values of the starting models. The gradient associated with the nondimensionalized parameters is $g'_p = m_0 g_p$. Inversion results are presented in terms of physical parameters. A square-root of depth preconditioner is implemented to compensate for inadequate illumination in deeper regions of the model. We forego Hessian based preconditioners due to the differences between $\mathbf{H}$ and $\hat{\mathbf{H}}$.

Restart versions of NLCG and L-BFGS are restarted after every 3 and 5 iterations, respectively. The Polak-Ribière scheme is used for NLCG (E. and Ribiere, 1969). Trials

using SD and NLCG are deployed with a bracketing line search, whereas conventional and restart versions of L-BFGS use a backtracking line search. In conjunction with L-BFGS, the backtracking line search can provide step-lengths at almost no additional cost (Modrak and Tromp, 2016). The line search satisfies the Armijo condition (first Wolfe condition) (Nocedal and Wright, 2006). We do not require the curvature condition (second Wolfe condition) be satisfied as it requires additional gradient computations for each trial step. Source-encoded inversions are performed 5 times to account for the variability introduced by the random source encoding. To enable fair comparisons between the various algorithms, termination conditions for the inversions must be specified. Termination criteria based on the relative misfit reduction ($J(\mathbf{m}^k)/J(\mathbf{m}^0)$) or gradient norms can be problematic when source encoding is used. In SEFWI, the objective function changes each time the encoding is randomized; therefore, the relative misfit reduction may not be a reliable indicator of convergence. Alternatively, a criterion that monitors the changes in misfit/gradient norm over a few iterations could be used (e.g. Castellanos et al. (2015)). Instead, we opt to terminate inversions after a predetermined number of iterations. The maximum number of iterations is selected based upon preliminary trials. To account for the fact that different algorithms converge at different rates, performance comparisons are evaluated at a target convergence point.

3.4.2 Diagnostic quantities

Before proceeding to the examples, we define diagnostic quantities that enable comparisons between SEFWI and FWI. The efficiency gain η is defined as $\eta = NS/\widehat{NS}$ and represents the ratio between the total number of simulations performed in FWI (NS) and SEFWI ($\widehat{NS}$). A simulation refers to any numerical solution of the forward or adjoint wave-equation during gradient computations or the line search. The variable convergence behaviour of each algorithm is implicitly represented by η. The efficiency gain is evaluated at a common convergence point for each algorithm. Specifically, η is computed once a target α model error is reached. The relative model error is defined for each independent parameter as

$$m_p^{err} = \frac{\|m_p^* - m_p^k\|}{\|m_p^*\|}, \tag{3.27}$$

where m_p^* is the true model for parameter p. Repeated trials in SEFWI are used to compute the mean and local covariances of inverted models in SEFWI; we use 5 independent trials for SEFWI inversions in this study. The mean $\tilde{m}_p(\mathbf{x})$ is computed over N independent trials (Castellanos et al., 2015),

$$\tilde{m}_p(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N} m_p^{(i)}(\mathbf{x}). \tag{3.28}$$

Table 3.1: SEG/EAGE overthrust inversion results. A comparison of the computational resources required by FWI and SEFWI to achieve $\alpha^{err} = 0.65$. Efficiency gain (η) describes the ratio between the total number of simulations required by FWI and SEFWI. As an additional comparison, efficiency gain is computed relative to the most efficient FWI implementation (FWI with L-BFGS).

	Iterations		No. simulations		η	
Optimization	FWI	SEFWI	FWI	SEFWI	vs FWI	vs FWI (L-BFGS)
SD	59	98	44160	568	77.7	8.6
NLCG	33	48	16416	341	48.1	14.4
L-BFGS	24	76	4896	266	18.4	18.4

The diagonals of the covariance matrix, are obtained via

$$\Sigma_{pq}(\mathbf{x},\mathbf{x}) = \sum_{i=1}^{N} \frac{(m_p^{(i)}(\mathbf{x}) - \tilde{m}_p(\mathbf{x}))(m_q^{(i)}(\mathbf{x}) - \tilde{m}_q(\mathbf{x}))}{N-1}. \tag{3.29}$$

In the absence of noise, the covariances act as a proxy for the cross-talk noise. The previous statement assumes that the mean inverted model is largely devoid of cross-talk artefacts.

3.4.3 Efficiency gain

Inversions are performed on a 20 km x 4.5 km, 2D section of the 3D acoustic SEG/EAGE overthrust model (Aminzadeh et al., 1997). Scaling relations are used to synthesize density ($\rho = 310\alpha^{0.25}$) and S-wave velocity ($\beta = \alpha/\sqrt{3}$) models. Starting models are obtained by convolving the true models with a Gaussian kernel ($\sigma_x = \sigma_z = 700$ m). True and initial α models are displayed in Figure 3.4. The seismic experiment consists of 96 explosive sources ($\Delta x_s = 200$ m, $z_s = 25$ m) recorded at 264 multi-component receivers ($\Delta x_r = 75$ m, $z_r = 25$ m). All 96 sources are combined into a single encoded source in SEFWI. The source wavelet is a Ricker wavelet with a dominant frequency of 5 Hz, corresponding to a bandwidth of 0-15 Hz. Inversions are performed using SD, NLCG, and L-BFGS and terminate after 100 non-linear FWI/SEFWI iterations. We invert the full bandwidth data as multi-scale methods are not required in this case. Model regularization did not lead to a discernible improvement in the inverted models for this experiment and is therefore not included. Truncating inversions by iteration number acts as an implicit form of regularization that limits overfitting (e.g. Hansen (1998)).

The convergence behaviour of FWI and SEFWI is summarized in Table 3.1 and Figure 3.5. For a particular optimization algorithm, SEFWI exhibits slower convergence, per iteration, in data misfit and model error when compared to FWI. The slower convergence is attributed to the presence of cross-talk noise in the gradient and is well established from prior studies

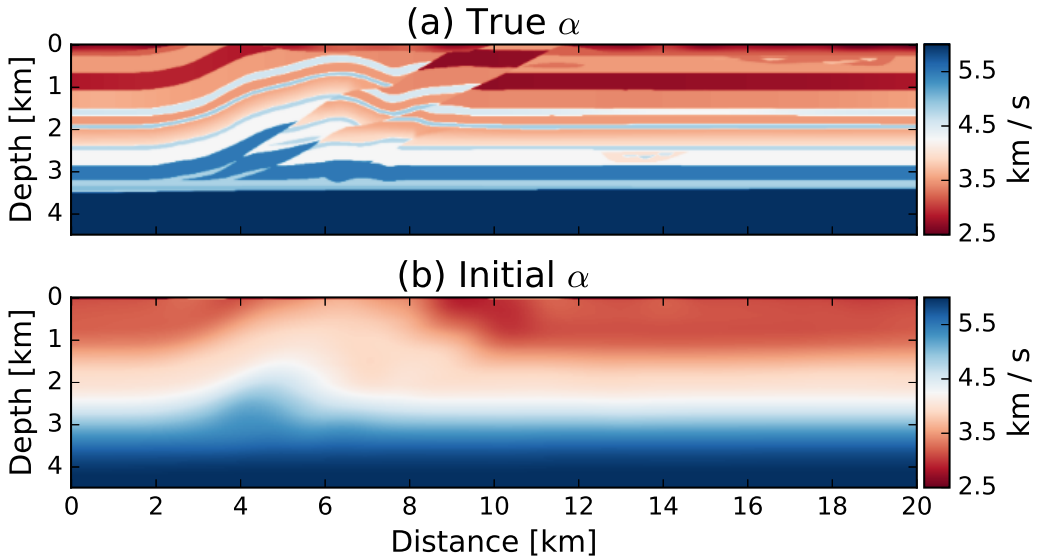

Figure 3.4: SEG/EAGE overthrust model. (a) True α model. (b) Initial α model. Empirical scaling relations, with respect to α, are used to synthesize the corresponding ρ and β models.

(Moghaddam et al., 2013; Anagaw and Sacchi, 2014; Castellanos et al., 2015). Efficiency gain and other performance comparisons are evaluated at a target model error $\alpha^{err} = 0.65$, which is the final model error for the slowest converging algorithm (SEFWI-SD). At the target model error, the largest efficiency gain is demonstrated by SD SEFWI ($\eta = 77.7$). NLCG and L-BFGS offer reduced efficiency gains of $\eta = 48.1$ and $\eta = 18.4$, respectively. The reduced effectiveness of NLCG/L-BFGS is due to an increased sensitivity of the algorithms to cross-talk noise; similar observations were documented by Castellanos et al. (2015). Despite exhibiting reduced efficiency gains, NLCG and L-BFGS still outperform SD in SEFWI, requiring fewer iterations and simulations to reach the target model error (Figure 3.5b). L-BFGS requires a greater number of iterations (76) than NLCG (48) to reach the desired model error; however, the more efficient line search results in fewer overall simulations despite the disparity in iterations. Relative to our most efficient FWI implementation (L-BFGS with a backtracking line search), the efficiency gain of SEFWI algorithms is more modest. This is an indication that the computational gain provided by SEFWI can be partially offset by more sophisticated optimization algorithms available in FWI.

Figs. 3.6 and 3.7 display mean inverted models and diagonal covariances. Figure 3.6 displays the evolution of $\tilde{\alpha}$, $\Sigma_{\alpha\alpha}^{1/2}$, and $|\Sigma_{\alpha\beta}|^{1/2}$ over the course of SD SEFWI iterations. The standard deviation of β ($\Sigma_{\beta\beta}^{1/2}$) is not included, but follows trends consistent with the other terms of the diagonal covariance matrix. The diagonal covariances reduce in magnitude at later iterations

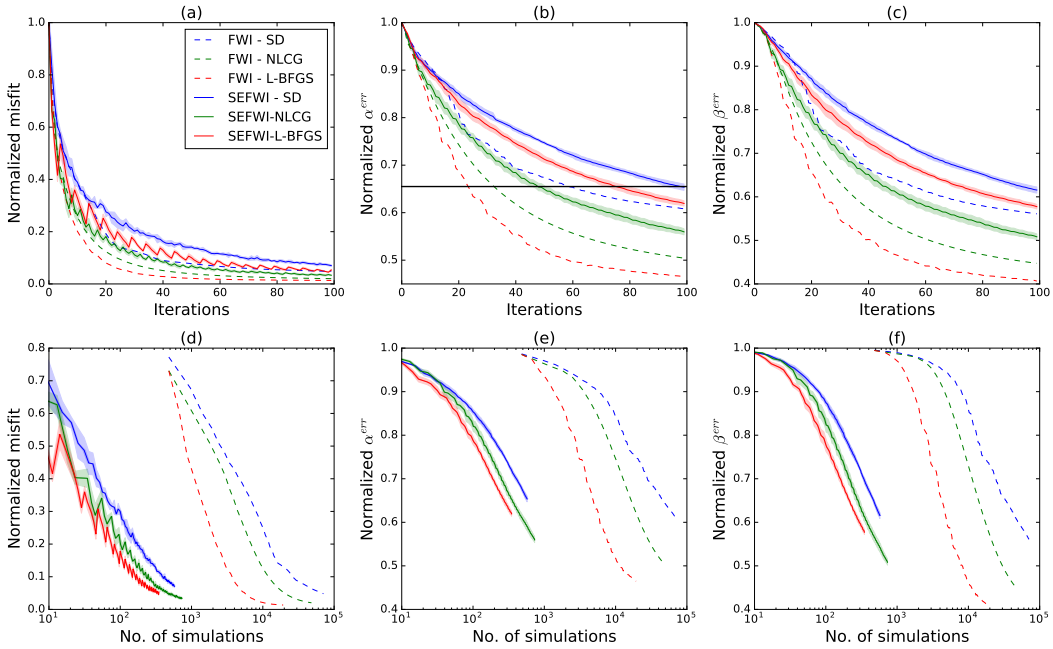

Figure 3.5: Convergence behaviour of FWI/SEFWI algorithms. (a, d) Normalized misfit. (b, e) α model error. (c, f) β model error. Each property is displayed as a function of iterations (top row) and number of simulations (bottom row). Dashed and solid coloured lines display results for FWI and SEFWI, respectively. For SEFWI, lines correspond to mean values of misfit/model error from 5 random trials; error bands represent one standard deviation. FWI exhibits higher per iteration convergence rates at the expense of a greater per iteration cost. The horizontal black line in panel (b) is the target model error used to compare algorithms in Table 1.

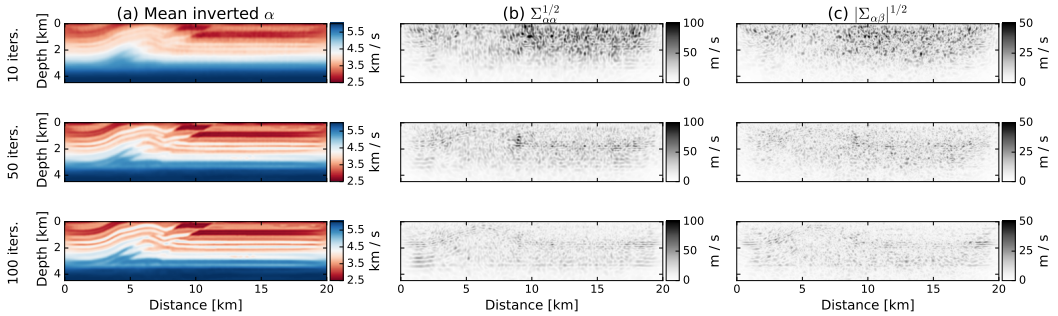

Figure 3.6: SEFWI inversion results for the overthrust model after 10 (top row), 50 (middle row), and 100 SD iterations (bottom row). (a) Mean α model. (b) $\Sigma^{1/2}_{\alpha\alpha}$. (c) $|\Sigma_{\alpha\beta}|^{1/2}$. The diagonal covariances decrease in magnitude as the iteration number increases implying that cross-talk artefacts are being attenuated.

indicating that the cross-talk artefacts are being suppressed. Figure 3.7 displays a similar comparison, but compares $\tilde{\alpha}$, $\Sigma^{1/2}_{\alpha\alpha}$, and $|\Sigma_{\alpha\beta}|^{1/2}$ after 100 iterations for the 3 different optimization methods. While SD exhibits diagonal covariances with lower magnitudes, the models inverted using NLCG/L-BFGS are better resolved, indicated by the lower model errors in Figure 3.5. The magnitude of the diagonal covariances is tied to the frequency with which the source-encoding is randomized for each algorithm. In these trials the restart intervals are 1, 3, and 5 for SD, NLCG, and L-BFGS, respectively. The restart version of NLCG appears to provide the best compromise between model resolution and mitigating cross-talk artefacts in the final model.

3.4.4 Parameter trade-off

The Marmousi II model is a fully elastic synthetic model with multiple hydrocarbon layers and complex faulting (Martin et al., 2006). Shallow shale layers in the original model exhibit low shear wave velocities (300-400 m/s) that require fine grid spacing to avoid dispersion related artefacts in the data. Reduced grid spacing increases the computational cost of forward/adjoint simulations due to a larger computational domain and considerations of numerical stability. To reduce the computational burden, S-wave velocities in the shale layers are replaced by $\beta = \alpha/\sqrt{3}$; density in these layers is rescaled via Gardner's relation. Adjusting the shale layers alone preserves heterogeneities exclusive to α or β. In this case, an exclusive heterogeneity refers to instances where one parameter demonstrates a significant perturbation from background, while the other parameter does not. Some examples are identified with white arrows in Figure 3.8. Heterogeneities exclusive to the α/β models serve as positional markers that are used to examine parameter trade-off. The water layer in the original model is removed to simulate land acquisition.

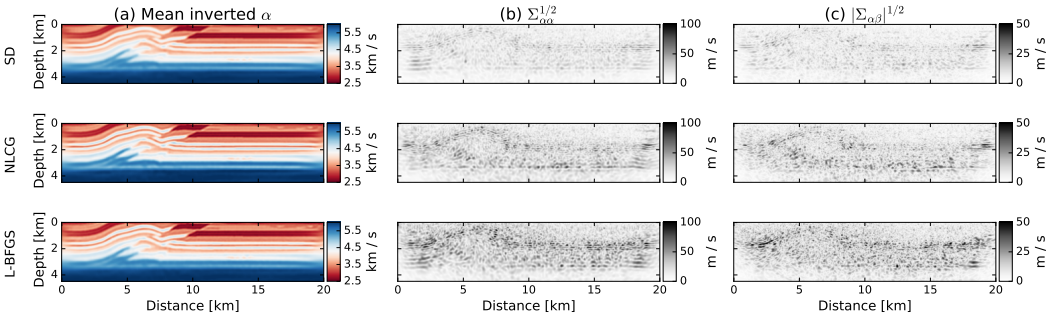

Figure 3.7: Final SEFWI inversion results after 100 iterations using SD (top row), NLCG (middle row), and L-BFGS (bottom row). (a) Mean α model. (b) $\Sigma_{\alpha\alpha}^{1/2}$. (c) $|\Sigma_{\alpha\beta}|^{1/2}$. For SD, the source-encoding is randomized at every iteration, whereas NLCG and L-BFGS randomize the source-encoding every 3 and 5 iterations, respectively. The amplitudes of the diagonal covariances reflect the strength of cross-talk artefacts, which in turn relate to the frequency at which the source-encoding is reset. Larger reset intervals are associated with more prevalent cross-talk artefacts.

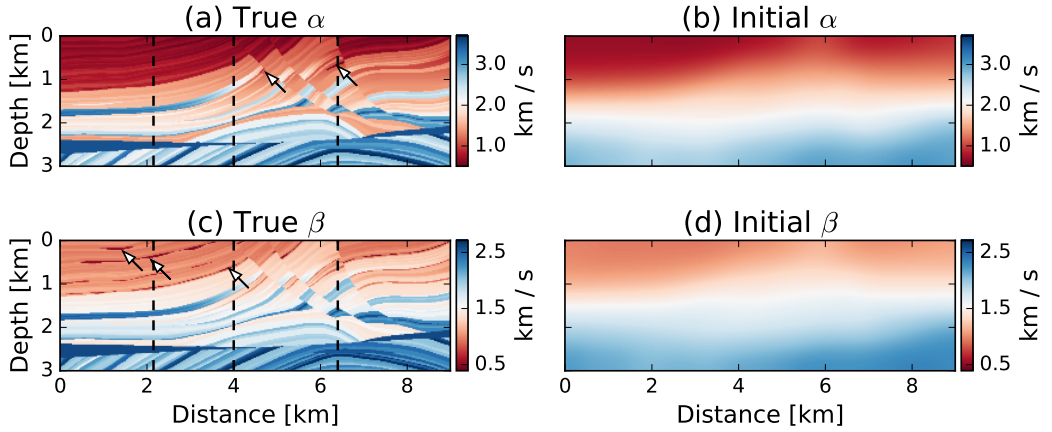

Figure 3.8: Modified Marmousi II model. (a) True and (b) initial α models. (c) True and (d) initial β models. The original β model has been altered to increase shear wave velocities in the shale layers. A heterogeneous ρ model is used, but not displayed. White arrows identify hydrocarbon reservoirs as perturbations from background in α and β. The dashed vertical lines designate the pseudo well logs used in Figure 3.10.

Initial models (Figure 3.8) are derived by convolving the true models with a Gaussian kernel ($\sigma_x = \sigma_z = 800$ m). The seismic experiment is composed of 112 explosive sources ($\Delta x_s = 80$ m, $z_s = 10$ m) and 296 receivers ($\Delta x_r = 30$ m, $z_r = 10$ m). 112 sources are reduced to $N_e = 2$ encoded sources, with each encoded source containing 56 individual sources. The source wavelet is a Ricker wavelet with a dominant frequency of 10 Hz. A 3 Hz highpass filter is applied to the data and source to remove some of the low frequency information. The starting model is sufficiently far from the true model that full-bandwidth FWI fails and converges to a local minimum in the objective function. The multi-scale approach of Bunks et al. (1995) is implemented to circumvent cycle-skipping. The frequency bands used for inversion are informed by the selection criteria of Sirgue and Pratt (2004). Inversions are performed using low-pass cutoff frequencies of 3 Hz, 5 Hz and, 8 Hz. The inversions start at 3 Hz because the filters do not have sharp cutoffs; therefore, some low-frequency information persists in the data. Some frequencies above the cutoff frequency prematurely appear in the inversion for the same reason. Multi-scale inversions are performed using SD to allow for a more direct comparison. The source encoding is randomized at every iteration and inversions are terminated after 75 SD iterations at each scale. We include damping regularization in the form of Equation 3.15 with $\gamma = 1 \times 10^{-4}$.

The final inverted α and β models are displayed in Figure 3.9. The SEFWI example corresponds to 1 of the 5 random trials conducted. Qualitatively, both inversion methods produce comparable results and no cross-talk artefacts are noticeable in the SEFWI models. Hydrocarbon layers, indicated by arrows in Figure 3.8, appear to be well resolved in α and β, with no perceptible trade-off between parameters. Further confirmation is provided by pseudo well-logs taken at different points in the model (Figure 3.10). Conventional FWI provides a marginal improvement in the estimation of true model perturbations.

Sensitivity to random noise

To test the sensitivity of SEFWI to noise, we add random noise to the Marmousi II data. Gaussian white noise arrays are generated for each component of each shot record. The variance of the noise array is set by selecting a desired signal-to-noise ratio (SNR), defined as

$$\text{SNR (dB)} = 10 \log_{10} \left(\frac{a_{rms}^2}{\sigma^2} \right), \tag{3.30}$$

and solving for the variance σ^2, where a_{rms}^2 is the root mean square amplitude of the shot record. For any given shot, the noise arrays of both components (x, z) have equal variance which results in different SNRs for the two components. We pick a_{rms}^2 from the z-component data and refer to the SNR of the z-component data in the text. The noisy dataset used for inversion has SNR = 10 dB. The inversion procedure follows the noise-free example, but with

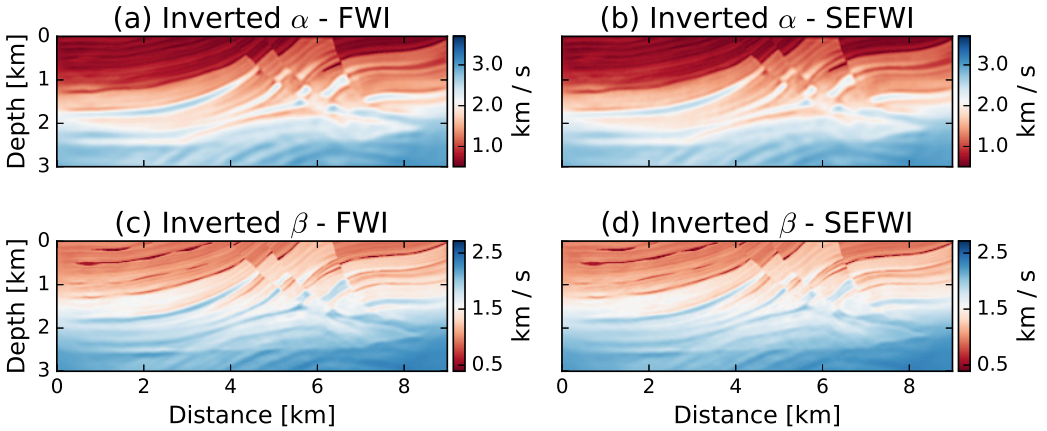

Figure 3.9: Final Marmousi II models after a multi-scale inversion. (a, c) FWI models. (b, d) SEFWI models. SEFWI attains models with similar resolution to those from FWI. SEFWI models do not exhibit any discernible parameter trade-off originating from cross-talk

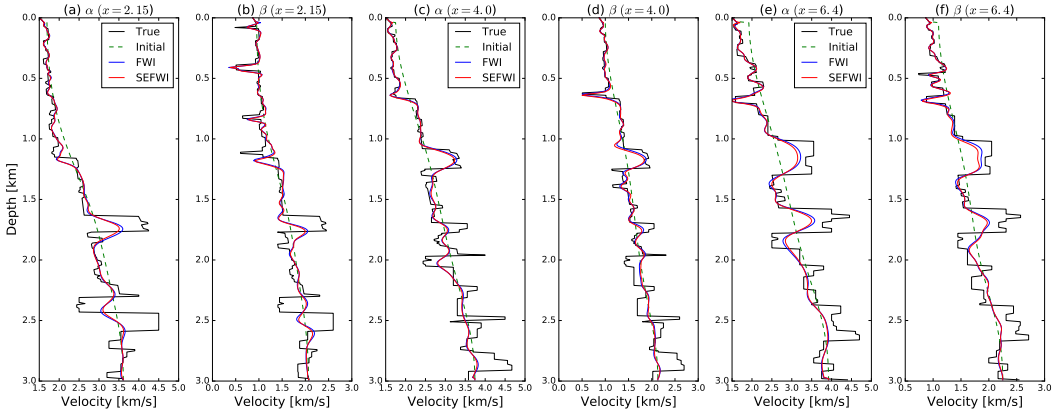

Figure 3.10: Pseudo well logs of α and β taken at (a, b) x = 2.5 km, (b, d) 4.0 km, and (e, f) 6.4 km. FWI Models display marginally better amplitude recovery at intermediate depths. Perturbations distinct to α or β do not appear to map into the other parameter, suggesting that the parameters are well resolved with both methods.

an increase to the regularization hyperparameter ($\gamma = 1 \times 10^{-3}$) to damp high-frequency contributions from noise in the model update.

Figure 3.11 (b) displays the mean inverted models and the diagonal covariances. The inversion results are largely similar to the noise-free case. A small increase in the magnitudes of the diagonal covariances is observed.

Sensitivity to early termination

Model updates that occur during the the later stages of FWI i.e. when the data misfit has largely flattened out, can still generate appreciable reductions in the model error despite producing only small reductions in the data misfit (e.g. Figure 3.5a-c). In realistic applications of 3D FWI, it may not be feasible to extend an inversion to a large number of iterations due to considerations of time or computational expense. In such cases, practitioners may terminate the inversion after a set number of iterations, before the optimization has truly converged. Early termination may also be prompted by strong noise in the data. In the presence of noise, the least-squares waveform misfit will converge to L_2 norm of the noise. Once the data misfit has flattened out, it is difficult to ascertain whether subsequent model updates are fitting the data or the noise. Early termination serves as a precautionary measure to prevent overfitting the data. In SEFWI, the later iterations are valuable as they further reduce the imprint of cross-talk artefacts.

We perform a test to investigate the effect of early termination in SEFWI. The inversion procedure follows the earlier multi-scale inversion performed on noise-free Marmousi II data, but with the number of SD iterations reduced from 75 to 30 at each scale. The mean inverted models and the diagonal covariances are depicted in Figure 3.11(c). Early termination does not appear to destabilize the inversion, rather it demonstrates two predictable results. Firstly, the mean inverted models are not as well resolved as in equivalent inversions run to a greater number of iterations (Figure 3.11a). Secondly, early termination increases the magnitude of the diagonal covariances, consistent with the expectation of increased cross-talk artefacts.

3.5 Limitations - Data driven inversion

The final synthetic example highlights a potential limitation of SEFWI when confronted with data-driven inversion schemes. Data-driven schemes perform a series of inversions in a hierarchical manner using various portions of the data, typically as means to mitigate the non-linearity of FWI by gradually introducing increased resolution/complexity into the model (e.g. Shipp and Singh (2002); Brossier et al. (2009)). This approach has been

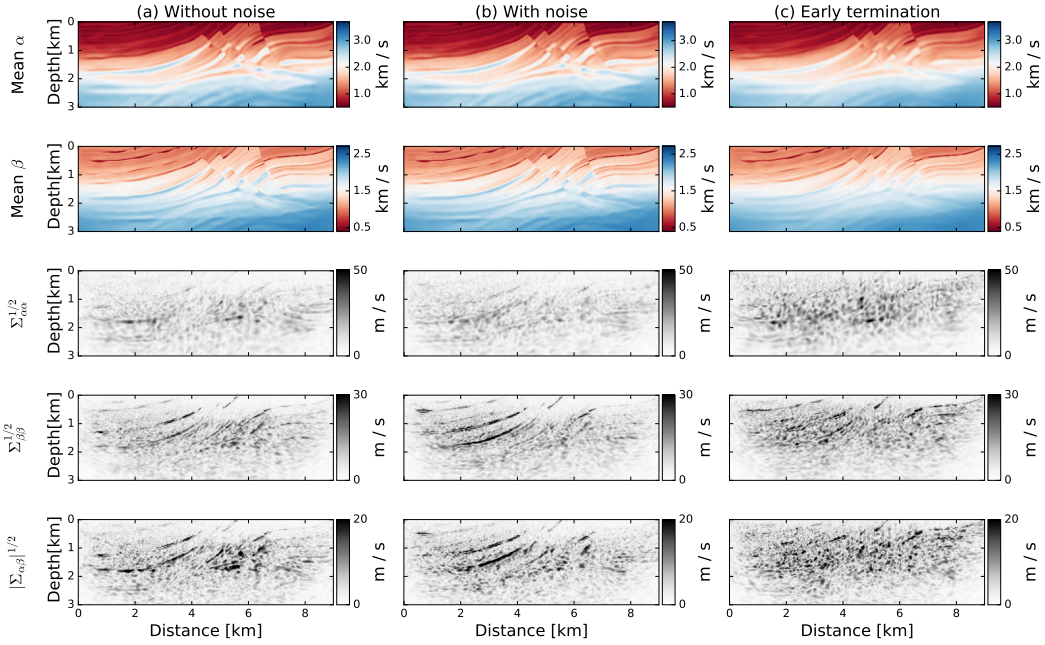

Figure 3.11: Mean velocity models and diagonal covariances for SEFWI: (a) noise-free data, (b) with noisy data (SNR=10 dB), and (c) early termination (noise-free data). 75 SD iterations, per scale, are used in (a) and (b) compared to only 30 for (c). Early termination produces models that are less resolved with greater cross-talk than those of (a).

successfully adopted to perform elastic inversion using ocean-bottom cables in a marine environment (Shipp and Singh, 2002; Sears et al., 2008, 2010; Prieux et al., 2013b). Sears et al. (2008) proposed a 3 stage inversion that imposed hierarchy on the model parameters and data components. The first stage conducts a broadband estimation of α from short- and wide-angle P-wave data recorded on hydrophones or z component OBC data. The inversion for β is divided into two stages to ensure that the intermediate wavelength structure is recovered prior to including shorter wavelength features into the model thereby mitigating the potential for cycle skipping. The amplitude variation with offset (AVO) of wide-angle P-waves can be used to update the intermediate wavelengths of the β model whereas PS waves primarily contribute to the short wavelength structure (Sears et al., 2008). However, as PS-waves exhibit greater sensitivity to changes in β than P-wave AVO (Tarantola, 1986; Ji et al., 2000), separating the two sources of information is important to prevent PS-waves from dominating the inversion. Stage 2 estimates intermediate wavelength structure of β from wide-angle P-wave data recorded on z component OBC data. The third and final stage fits PS-wave data recorded on x component OBC data to update the short wavelength structure of β. In this section, we apply the data-driven strategy of Sears et al. (2008) to a marine version of the overthrust model. We demonstrate that SEFWI fails in this setting because the assumption of a fixed-spread is not realized owing to time-windowing applied on the data.

A marine example is created by modifying the overthrust model presented in section 4.3. A 500 m water layer is added to the top of the model. 96 sources ($\Delta x_s = 200$ m, $z_s = 25$ m), placed just beneath the sea surface, are recorded at 264 multi-component ocean-bottom nodes ($\Delta x_r = 75$ m, $z_r = 525$ m). Low frequencies are removed from the data and source wavelet by applying a 5 Hz lowcut filter. The removal of low frequencies further promotes the use of a hierarchical strategy as it ensures that low/intermediate wavelength information cannot be obtained from low-frequency PS-waves. The starting models are obtained by smoothing the true models, excluding the water layer, with a Gaussian kernel ($\sigma_x = \sigma_z = 1000$ m). A multi-scale approach (over frequencies) did not alter the success or failure of an inversion nor did it noticeably improve the results in these trials; therefore, we invert the full bandwidth data. Preliminary inversions showed strong sensitivity to the near surface structure. To ensure accurate sea-floor mode conversions, we assumed that the upper 250 m of the seafloor was known and kept it fixed during inversions.

Figure 3.12 displays FWI models inverted without a hierarchical inversion scheme. The parameters α and β are estimated simultaneously using x and z component OBC data in an inversion that terminates after 50 L-BFGS iterations. While the α model is recovered well, the β model converges to a local minimum (Figure 3.12b), evident from the poorly reconstructed deeper layers. The failure of the inversion is attributed to PS-mode converted waves dominating the inversion and adding short wavelength features before the intermediate

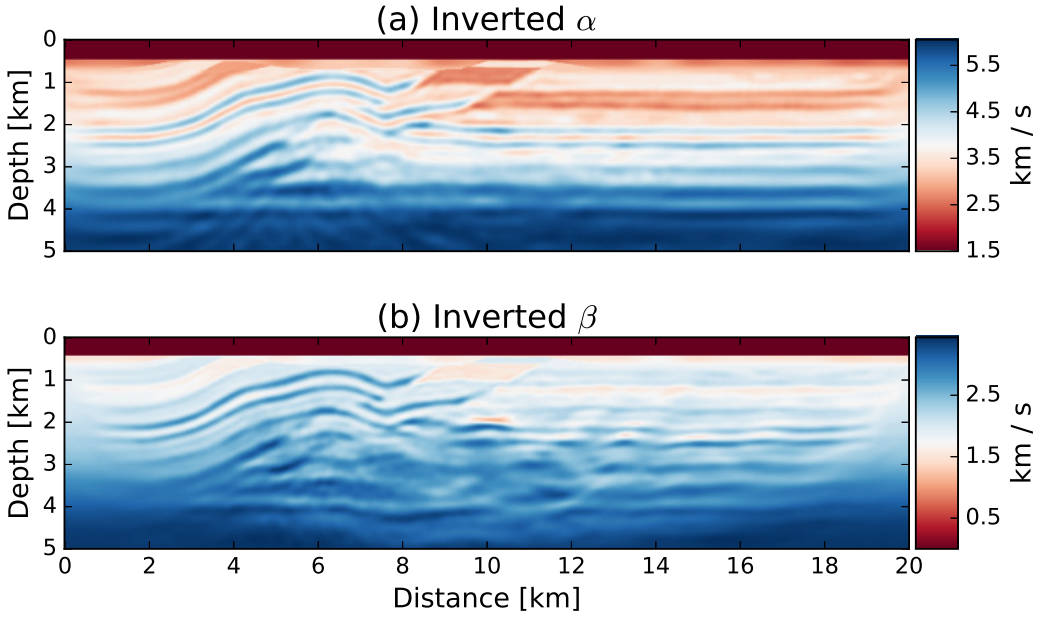

Figure 3.12: Final FWI models following a simultaneous inversion of α and β from x and z component OBC data. (a) Final α and (b) β models after 50 L-BFGS iterations. The β model has converged to a local minimum.

wavelengths are suitably recovered. To circumvent the non-linearity encountered here, we employ the data-driven scheme of Sears et al. (2008) as described earlier.

Stage 1 estimates α by fitting x and z components with no time windowing applied to the data. Stage 2 is restricted to wide-angle P-wave data after muting short offsets (< 4 km) and late arrivals (linear slope mute) in z component OBC data. Stage 3 data focuses on PS-waves by applying a linear slope mute to early arrivals in x component OBC data. Each stage of the inversion terminates after 50 SD iterations and uses the final model from the preceding stage as an initial model. Examples of the time-windowed input data for stages 2 and 3 are available in Figure 3.13, alongside corresponding synthetics (Figs. 3.13b and f), initial (Figs. 3.13c and g) and final residuals (Figs. 3.13d and h). The residual energy is slightly reduced after each stage. The FWI models after each stage are displayed in Figure 3.14. An acceptable α model is recovered after stage 1 at which point density is updated using Gardener's relation. Stage 2 captures some features of the fault zones and begins to distinguish between the different layers in the β model. Following stage 3, the deeper layers in β are mostly continuous and flat, demonstrating better resolution than the earlier simultaneous inversion result. We note that some of the shallow layers are not completely recovered in either the α or β model.

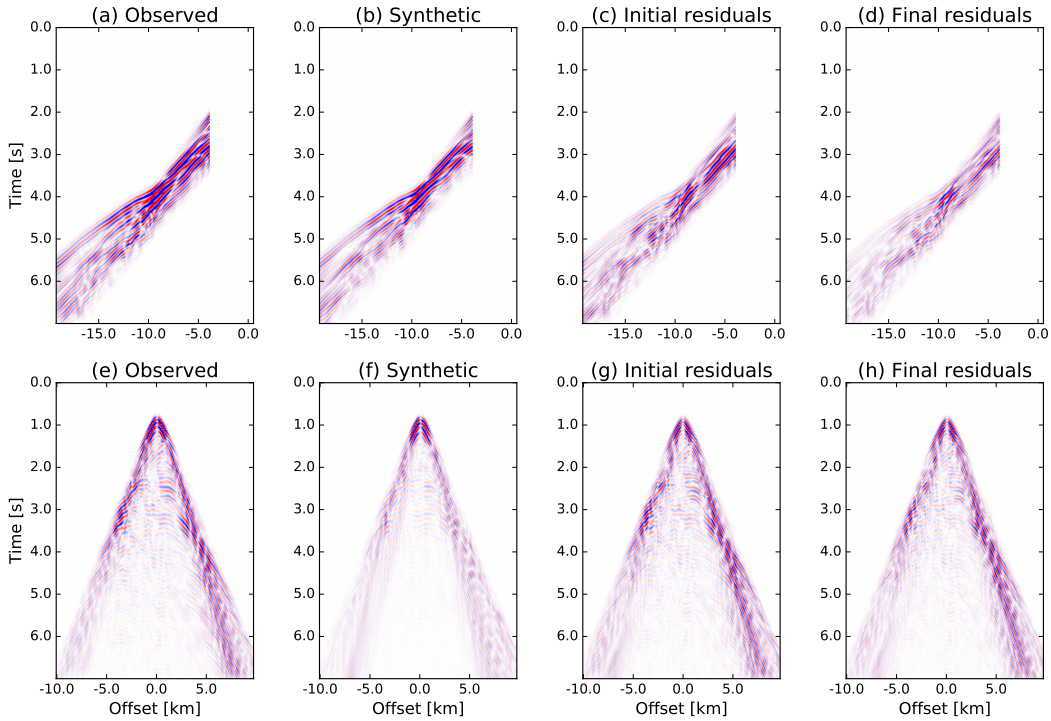

Figure 3.13: Time-windowed data and residuals from (a-d) stage 2 and (e-h) stage 3 of the marine overthrust example. Stage 2 inverts for intermediate length scales of the β model by fitting amplitude variations of wide-angle P-wave data. Stage 3 inverts for short wavelength β structure by fitting PS waves. Shot records from a single stage are plotted using the same scaling.

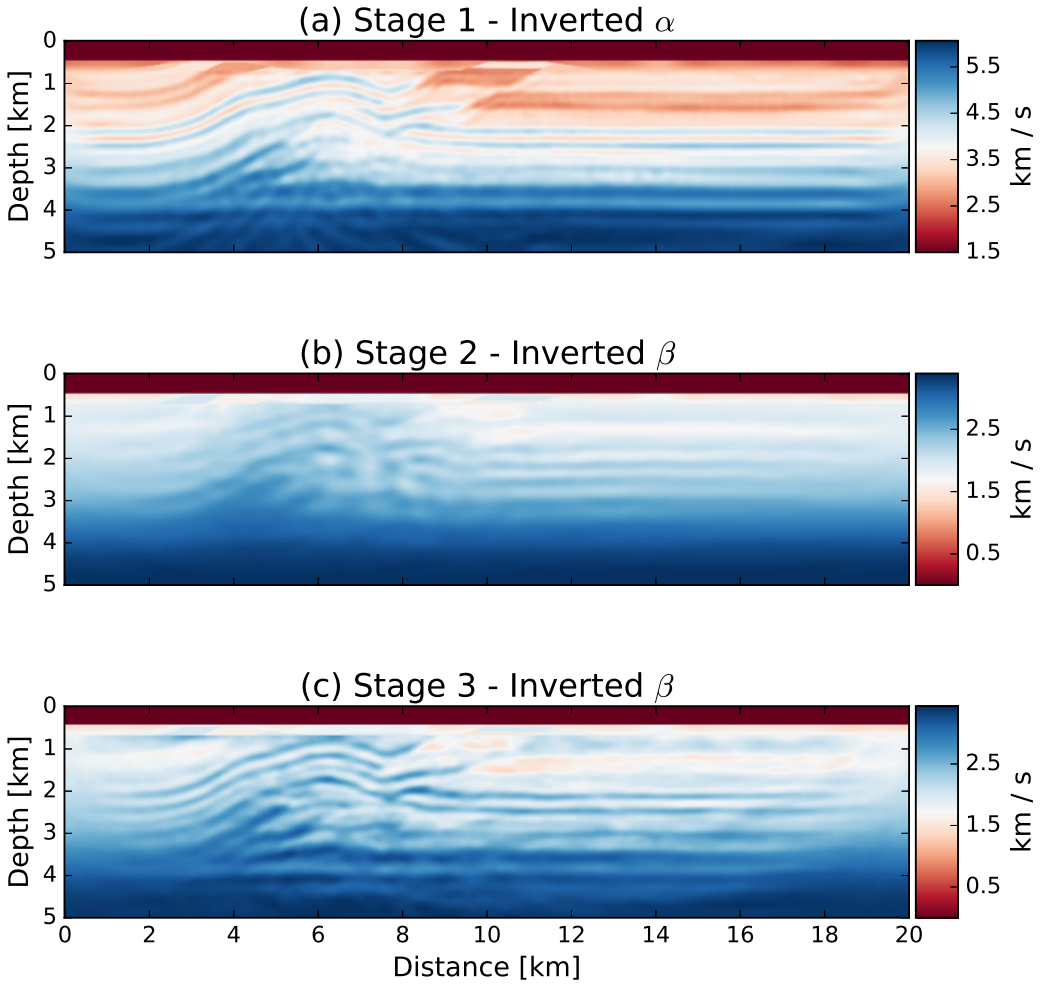

Figure 3.14: FWI models after each stage of a data-driven inversion of OBC data (Sears et al., 2008). (a) Final α model inverted from x and z component OBC data. (b) Intermediate wavelength β model inverted from wide-angle P-waves recorded on the z component. (c) Final β model inverted using PS-waves recorded on the x component.

We attempt an equivalent inversion using SEFWI where the input data are the time-windowed data as depicted in Figure 3.13. A comparison of the encoded data (Figs. 3.15a and e) and synthetics (Figs. 3.15b and f) reveals a problem. Since the synthetic wavefield is computed as an encoded wavefield, we cannot access independent shot records. Consequently, we cannot apply the time windowing that was applied to the data, to the synthetic data. The encoded synthetics include portions of the wavefield that are not present in the encoded data. This limitation ultimately stems from the fixed-spread acquisition assumption in SEFWI. Certain sources may not be recorded across all the receivers within an acquisition due to reasons including, but not limited to, poor/failed measurements, non-stationary acquisition geometries (e.g. towed streamer), or processing requirements on the data (e.g. time windowing). To further illustrate this problem, we compare the encoded residuals with the waveform adjoint source. The encoded residuals refer to the result obtained by encoding the waveform residuals computed for independent shot records (computed separately here). Re-injecting the encoded residuals would produce the FWI gradient plus crosstalk noise as described in Equation 3.10. During adjoint modelling, the waveform adjoint source back-propagates synthetic data in places where no data exists. These undesired contributions introduce errors into the gradient and subsequently the inversion.

A similar situation is encountered when source encoding is applied to towed-streamer data, a scenario in which the observed and synthetic data have incompatible acquisition geometries. Routh et al. (2011) and Choi and Alkhalifah (2012) used the normalized cross correlation (NCC) in lieu of the waveform misfit to circumvent this issue. Maximizing the normalized cross correlation is akin to minimizing a normalized form of the waveform misfit (Choi and Alkhalifah, 2012). The advantage becomes apparent by observing the NCC adjoint source (Figs. 3.15d and i). The adjoint source resembles an instance of the waveform residual in which the amplitude of the synthetic wavefield is controlled by the similarity between the observed and synthetic data. Due to the weighting, the contributions from traces where data are not available are downweighted relative to traces where data exists. This effect is noticeable in Figure 3.15(d) where the NCC adjoint source more closely resembles the encoded residual. The L1 waveform misfit is a robust norm that is less sensitive to outliers than the L2 waveform misfit (Crase et al., 1992; Brossier et al., 2010). We attempt to treat the undesired components in the waveform residual as outliers by employing the L1 waveform misfit. Both alternative objective functions are utilized in the upcoming SEFWI trials.

Each stage of SEFWI performs 50 SD iterations. The 96 independent sources are combined into 16 encoded sources each containing 6 individual sources. The inverted FWI models from the previous stage are used as initial models in SEFWI to ensure a consistent starting point for all inversions. All 3 misfit functionals expectedly invert α during stage 1 as no time-windowing was applied to the data. Stage 1 results are not presented as they

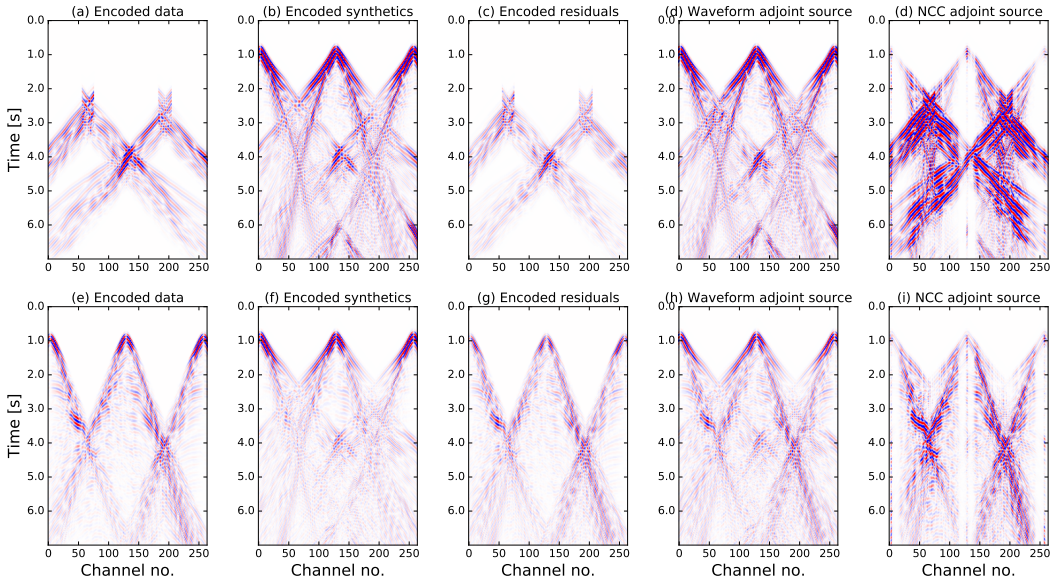

Figure 3.15: Encoded data and adjoint sources for (a-d) stage 2 and (e-i) stage 3. For illustration purposes, only 3 sources are encoded in panels (a-i). The encoded data are obtained by encoding the time-windowed data observed in Figure 3.13. Similar time-windowing cannot be applied to the synthetics as the wavefield is computed with encoding in place. The encoded residuals (c, d) represent the encoded waveform residuals computed when the synthetic wavefield is available for each independent source. The waveform adjoint source (d, h) contains undesired contributions from synthetic wavefield. The normalized cross correlation adjoint source more closely resembles the encoded residuals (Routh et al., 2011; Choi and Alkhalifah, 2012).

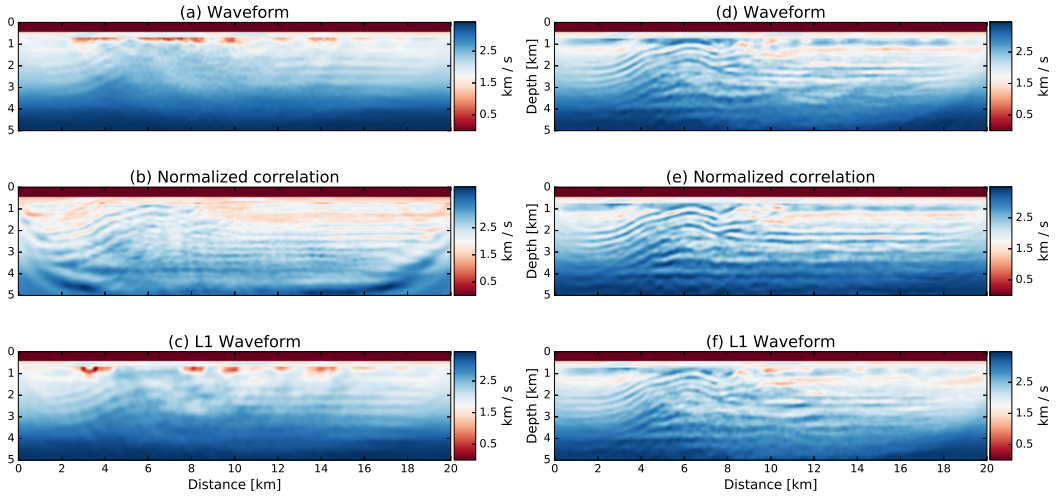

Figure 3.16: SEFWI β models after (a-c) stage 2 and (d-f) stage 3 for different misfit functionals. Each stage is initialized with the final FWI model from the previous stage. The normalized cross correlation fares better than the alternatives, but no choice of misfit produces acceptable inversion results.

do not offer additional insight over earlier examples. Inversion results for stages 2 and 3 are displayed in Figure 3.16. None of the inversion results from the latter two stages are deemed successful as, in all instances, the model error of the final inverted model is worse than the starting model. For this example, the two variants of the waveform misfit do not show appreciable differences in inversion results. As evidenced by the failure of the L1 waveform objective function, it is not suitable to treat the undesired components of the waveform residual as outliers. The algorithm is not able to distinguish between wanted/unwanted signal and therefore fails. The normalized cross correlation generally outperforms the other two objective functions tested. Two factors may explain the inability of the NCC to properly invert for β. Firstly, the normalization causes a loss of sensitivity to amplitude variations which is counter-productive considering that stage 2 is predicated on fitting P-wave AVO. Secondly, while the NCC adjoint source damps unwanted noise in the residuals, the remaining noise is too prominent and contributes to the failure of the inversion.

Current forms of SEFWI fail to accommodate time-windowing steps that can be essential to ensuring the success of an inversion. In place of alternative objective functions, we offer speculation on how the fixed-spread limitation might be overcome by applying an intermediate processing step that deblends the synthetic data. Hypothetically, if deblending could be applied to the synthetic data (e.g. Ibrahim and Sacchi (2014)), it would provide access to the individual shot records. Following deblending, the desired time-windowing

could be applied to the separated shot records. The time-windowed, deblended synthetics would be used to compute encoded residuals as portrayed in Figure 3.15. Replacing the waveform adjoint source with the encoded residuals would result in a meaningful gradient as in Equation 3.10. A strategy centred around deblending would require an encoding scheme that is not overly aggressive i.e. the number of sources in an encoded source should be low enough to fall into the regime where deblending is possible. Restricting the number of sources in an encoded source imposes an upper limit on the efficiency gain, thereby diminishing the appeal of SEFWI. When time-windowing is a requirement of the workflow, stochastic optimization algorithms (e.g. Haber et al. (2012); van Leeuwen and Herrmann (2013a)) are likely a more suitable means to improving the efficiency of FWI. Stochastic algorithms achieve their computational efficiency by operating over a small subset of the complete dataset. In FWI, this amounts to subsampling the sources used for inversion at any iteration. Stochastic algorithms require minimal changes to the conventional FWI algorithm. As source encoding is not required, time-windowing can readily be applied. We do not pursue the deblending concept or comparisons with stochastic optimization techniques further as they are beyond the scope of this study.

3.6 Conclusions

Source encoding has been applied to elastic isotropic full waveform inversion. The theory of source-encoded FWI was extended to the general multi-parameter case, with an emphasis placed on understanding the influence of source encoding on parameter trade-off. The behaviour was determined by analysing the source-encoded multi-parameter Hessian. The convergence of the expected source-encoded Hessian towards the conventional FWI Hessian, was verified via Hessian probing techniques. When cross-talk artefacts are suppressed, the properties of the source-encoded Hessian ensure that the parameter trade-off in SEFWI is comparable with FWI. Furthermore, SEFWI exhibits similar sensitivities to model parametrization and inversion schemes as FWI.

Additional numerical tests evaluated the performance and stability of SEFWI. In particular, tests sought to assess the efficiency gain, parameter trade-off, and the suitability of SEFWI for data-driven workflows. For all optimization algorithms, SEFWI required significantly fewer overall simulations (than FWI) to reach a target model error. The efficiency gain was on the order of the number of individual sources in an experiment. A hybrid-CG algorithm was judged to have outperformed SD or L-BFGS alternatives. While SD generated SEFWI models with the lowest variances, both CG and L-BFGS converged to a smaller model error in the same number of iterations. The increased model variance observed with hybrid CG/L-BFGS algorithms is due to less frequent randomization of the source encoding. A test on

the Marmousi II model corroborated the claim that the parameter trade-off in SEFWI is comparable to that of FWI. Spatially inconsistent $P-$ and S-wave velocity models were well resolved in both methods. The presence of noise or early termination did not exacerbate the inversion results significantly. Early termination resulted in models that were less resolved and exhibited greater diagonal covariances.

Our results indicate that source encoding is feasible in multi-parameter FWI; however, there are concerns that hinder the use of source encoding in real data applications. The fixed-spread acquisition assumption impedes the use of source encoding in data-driven workflows that require extensive data pre-processing (e.g. time or offset windowing), as demonstrated by a 3 stage inversion of OBC data. Stages that required time-windowing on the data failed to produce acceptable models when tested with the waveform misfit. The normalized cross correlation and L1 waveform objective functions were not able to compensate for the time-windowing and also failed. Out of the three objective functions tested, the normalized cross correlation offered the most reasonable results. Ultimately, the applicability of source encoding on real data, multi-parameter FWI is entirely dependent on the dataset. Alternative misfit functionals are not always applicable; therefore, future work on source encoding should pursue techniques that can accommodate processing steps into SEFWI. Source encoding may be useful for accessing the multi-parameter Hessian in an economical manner (Tang, 2009) or to produce preconditioners for FWI/migration (Tang and Lee, 2010). In addition, source-encoding could be used to reduce the cost of the inner conjugate gradient iteration of truncated Newton methods (e.g. (Castellanos et al., 2015)).

CHAPTER 4

A subsampled truncated-Newton method for multi-parameter full waveform inversion[1]

4.1 Introduction

Multi-parameter full-waveform inversion (FWI) can benefit from Newton-based optimization algorithms that account for the Hessian; potential advantages include improved convergence rates, resolution, and mitigation of parameter trade-off (Pratt et al., 1998; Operto et al., 2013). However, explicit computation of the Hessian is not feasible on current hardware for large-scale FWI problems due to its computational cost. This limitation has motivated the pursuit of computationally inexpensive approximations to the Hessian to augment the performance of FWI. Hessian approximations are generally derived from knowledge about its structure. Pratt et al. (1998) established the Hessian in FWI as a dense, diagonally-dominant banded matrix. The banded structure arises from the finite-frequency nature of seismic data. The Hessian acts as a convolutional operator that, when applied to a vector, smooths it spatially. Conversely, the inverse Hessian behaves as a focusing operator that improves resolution. We present a brief overview of some common Hessian approximations and describe their potential limitations.

The pseudo-Hessian is a popular preconditioner that approximates the Hessian as a diagonal matrix (Shin et al., 2001). Block diagonal extensions of the pseudo-Hessian for multi-parameter problems have also been explored (Innanen, 2014; Métivier et al., 2015; Wang et al., 2016). Diagonal approximations act as spatial weighting operators, typically used to rescale gradients to compensate for geometrical spreading or inadequate subsurface

[1] A version of this chapter is published in Matharu. G., and M. D. Sacchi, 2019, A subsampled truncated-Newton method for multi-parameter full-waveform inversion, Geophysics, 84, R333-R340.

illumination. In neglecting the banded-diagonal structure of the Hessian, diagonal approximations fail to account for the finite-frequency nature of the Hessian and therefore do not improve focusing in the gradient. Approximating the action of the Hessian within a limited spatial window has been proposed as a strategy to incorporate the banded structure of the Hessian (Valenciano et al., 2006; Tang and Lee, 2015; Feng et al., 2018). In such methods, the Hessian is replaced by a series of non-stationary filters computed at certain points in a model. Filter construction is an added expense as it requires sampling elements from the Hessian. The effectiveness of these methods relies on the action of the Hessian being limited to small spatial windows about various grid points in the model. When violated, the size of the windows can become large thus requiring more samples from the Hessian. Inexact or truncated Newton (TN) methods do not approximate the Hessian, but rather approximate the solution of a linear system of equations (featuring the Hessian) using the conjugate gradient method (CG) (Akcelik et al., 2002; Epanomeritakis et al., 2008; Métivier et al., 2013). Each CG iteration computes the action of the Hessian on a vector at a cost comparable to that of two gradient computations (Fichtner and Trampert, 2011a). The advantages of second-order algorithms can become marginal in light of their increased cost (Métivier et al., 2013). Gradients and Hessian-vector products in FWI, typically computed using adjoint-state methods, require a number of PDE solves that grows linearly with the number of sources in a data set. The linear dependence prompts the consideration of strategies that reduce the dimensions of the data thereby lessening the computational burden of FWI.

Two proven data reduction strategies in FWI are source subsampling/decimation and source encoding. The former employs subsets of the complete data set within FWI (eg., van Leeuwen and Herrmann (2013a); Warner et al. (2013)) whereas the later employs simultaneous or encoded sources that represent weighted linear combinations of multiple sources. The linearity of the wave equation with respect to the source, allows combined wavefields to be simulated using encoded sources. The number of encoded sources is generally much smaller than the number of independent sources in a survey. A drawback associated with source encoding is the introduction of cross-talk artefacts to the FWI gradient; however, these can be ameliorated by selecting appropriate encoding functions (Romero et al., 2000; Krebs et al., 2009). Source encoding assumes a fixed-spread acquisition making it incompatible with data-driven inversion strategies (e.g., time/offset windowing schemes) that are, at times, necessary to navigate the non-linearities of multi-parameter FWI (e.g., Sears et al. (2008); Matharu and Sacchi (2018)). Source subsampling techniques are not subject to this limitation, motivating us to examine stochastic second-order optimization methods. Second-order optimization with source encoding has been explored in earlier studies (Anagaw and Sacchi, 2014; Castellanos et al., 2015).

This chapter investigates a subsampled truncated Newton (STN) approach to multi-parameter

(isotropic) elastic FWI. Hessian-vector products are computed for a uniformly, or non-uniformly, sampled subset of the sources. The performance of STN is benchmarked against LBFGS and TN algorithms in a series of synthetic inversions. Generally, we find the performance of STN is similar to TN but with computational costs closer to first-order methods for the selected convergence criteria.

4.2 Theory

Full-waveform inversion estimates subsurface parameters $\mathbf{m}$ by minimizing the least-squares waveform misfit functional over a set $\mathcal{D}$ of n sources:

$$\begin{aligned} J_{\mathcal{D}}(\mathbf{m}) &= \frac{1}{2|\mathcal{D}|}\sum_{i\in\mathcal{D}} \|\mathcal{P}_{\mathcal{D}}\mathbf{u}_i(\mathbf{m}) - \mathbf{d}_i\|^2 + R(\mathbf{m}) \\ &= \frac{1}{|\mathcal{D}|}\sum_{i\in\mathcal{D}} j_i(\mathbf{m}) + R(\mathbf{m}), \end{aligned} \tag{4.1}$$

where $\mathbf{u}_i(\mathbf{m})$ and $\mathbf{d}_i$ denote simulated and observed data for the i^{th} source, respectively. The sampling operator $\mathcal{P}_{\mathcal{D}}$ extracts the synthetic wavefield at the receiver positions, and $R(\mathbf{m})$ is a regularization term. The model vector for p independent parameters is $\mathbf{m} = [\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_p]^T$, where each $\mathbf{m}_i \in \mathbb{R}^M$ $i = \{1, \ldots, p\}$ represents a physical parameter discretized onto a grid of size M. Canonically, Equation 4.1 is minimized via gradient-based optimization algorithms that iteratively update estimates of the model via

$$\mathbf{m}^{k+1} = \mathbf{m}^k + \nu^k \delta\mathbf{m}^k, \tag{4.2}$$

where k denotes the iteration number, ν^k is a scalar step length, and $\delta\mathbf{m}^k$ is the model update (Tarantola, 1986; Virieux and Operto, 2009). First-order gradient methods construct $\delta\mathbf{m}$ using current, and potentially previous, iteration gradient $\mathbf{g}_{\mathcal{D}} = \nabla J_{\mathcal{D}}(\mathbf{m})$ information. The gradient associated with Equation 4.1 is

$$\mathbf{g}_{\mathcal{D}} = \frac{1}{|\mathcal{D}|}\sum_{i\in\mathcal{D}} \nabla j_i(\mathbf{m}) + \nabla R(\mathbf{m}). \tag{4.3}$$

Second-order gradient methods solve the Newton equation

$$\mathbf{H}_{\mathcal{D}}\delta\mathbf{m} = -\mathbf{g}_{\mathcal{D}}, \tag{4.4}$$

where $\mathbf{H}_{\mathcal{D}} = \nabla^2 J(\mathbf{m})$ is the Hessian of the objective function. Truncated Newton (TN) methods solve Equation 4.4 by solving

$$\arg\min_{\delta\mathbf{m}} \ \|\mathbf{H}_{\mathcal{D}}\delta\mathbf{m} + \mathbf{g}_{\mathcal{D}}\|^2, \tag{4.5}$$

using CG. Second-order adjoint-state methods can be used to compute Hessian-vector products $\mathbf{H}_{\mathcal{D}}\delta\mathbf{m}$ without explicitly forming the Hessian (Fichtner and Trampert, 2011a). The number of PDE solves required (per-iteration) to estimate $\delta\mathbf{m}$ are $2n$ and $2n + (N_{CG} \times 2n)$ for first-order gradient and TN methods, respectively; N_{CG} denotes the number of CG iterations performed in a single TN iteration. The costs depend linearly on the size of the data set $|\mathcal{D}| = n$, presenting a significant bottleneck for large data sets.

4.2.1 Stochastic optimization

Stochastic optimization exploits redundancy in the data to randomly omit samples (Bottou, 2010). Stochastic approximations to $J_{\mathcal{D}}$ and $\mathbf{g}_{\mathcal{D}}$ arise from substituting a random subset $\mathcal{X} \subseteq \mathcal{D}$ in place of $\mathcal{D}$ in equations 4.1 and 4.3. Estimating $\delta\mathbf{m}$ only requires $2|\mathcal{X}|$ PDE solves per iteration under the stochastic approximation. First-order stochastic gradient methods have been successfully adopted in FWI (van Leeuwen and Herrmann, 2013a).

Our contribution is to investigate the extension of stochastic optimization to second-order methods via the subsampled Truncated Newton method. The feasibility of the STN approach has been validated on machine learning applications (Byrd et al., 2011). The proposed method introduces a second level of subsampling such that Hessian-vector products are computed over a smaller subset $\mathcal{S} \subseteq \mathcal{X}$. The CG iterations of STN estimate $\delta\mathbf{m}$ by (approximately) solving

$$\arg\min_{\delta\mathbf{m}} \ \|\mathbf{H}_{\mathcal{S}}\delta\mathbf{m} + \mathbf{g}_{\mathcal{X}}\|^2, \tag{4.6}$$

where $\mathbf{H}_{\mathcal{S}}$ is the subsampled Hessian. The definition of $\mathbf{H}_{\mathcal{S}}$ is provided at a later stage. The subsets $\mathcal{X}$ and $\mathcal{S}$ are redrawn after every iteration. Fixing subsets $\mathcal{X}$ and $\mathcal{S}$ can bias the estimated search directions. With fixed subsets, acquisition related artefacts can stack as coherent noise over FWI iterations (van Leeuwen and Herrmann, 2013a). Dynamic subsets where $|\mathcal{X}|$ and $|\mathcal{S}|$ grow over iterations are possible (Byrd et al., 2012; Friedlander and Schmidt, 2012; van Leeuwen and Herrmann, 2013a; Bollapragada et al., 2016); however, we do not consider them here. For theoretical analyses on convergence and sufficient subset sizes, the reader is referred to the existing literature on the topic (Erdogdu and Montanari, 2015; Roosta-Khorasani and Mahoney, 2016a,b; Xu et al., 2017; Bollapragada et al., 2016). Assuming fixed subset sizes, the per-iteration cost of computing $\delta\mathbf{m}$ with STN is $2|\mathcal{X}| + (N_{CG} \times 2|\mathcal{S}|)$ PDE solves. A summary of the costs for various algorithms is presented in

Algorithm	Class	PDE solves per iter.
FG	First-order	$2n$
SG	First-order	$2\|\mathcal{X}\|$
TN	Second-order	$2n + (N_{CG} \times 2n)$
STN	Second-order	$2\|\mathcal{X}\| + (N_{CG} \times 2\|\mathcal{S}\|)$

Table 4.1: Computational cost associated with calculating $\delta\mathbf{m}^k$ for full (FG) and stochastic gradient (SG) methods, along with truncated Newton (TN) and subsampled truncated Newton (STN) methods. The costs assume that the wavefields required to construct the gradient (or Hessian-vector products) are stored.

Table 4.1.

We compute the gradient over all the sources ($\mathcal{X} = \mathcal{D}$). While gradient subsampling can readily be applied, it raises the question of whether STN should be benchmarked against stochastic or conventional gradient methods. To enable fair comparisons we do not consider gradient subsampling, but we acknowledge that additional efficiency gains may be attained by doing so. Bottou et al. (2018) caution against the use of small $|\mathcal{X}|$ ($|\mathcal{X}| << |\mathcal{D}|$) as noisy gradients may yield poor search directions from the truncated Newton iterations.

4.2.2 Sampling strategies

Given a discrete sampling probability distribution $r_i \;\; i = \{1, \dots, |\mathcal{X}|\}$, $\mathbf{H}_\mathcal{S}$ can be defined as

$$\mathbf{H}_\mathcal{S} = \frac{1}{|\mathcal{X}||\mathcal{S}|} \sum_{i \in \mathcal{S}} \frac{\nabla^2 j_i}{r_i} + \nabla^2 R(\mathbf{m}), \tag{4.7}$$

and is an unbiased estimator of $\mathbf{H}_\mathcal{X}$ ($\mathbb{E}[\mathbf{H}_\mathcal{S}] = \mathbf{H}_\mathcal{X}$) (Xu et al., 2016). The sub-Hessians $\nabla^2 j_i$ can refer to the full or Gauss-Newton form of the Hessian. For uniform sampling, $r_i = 1/|\mathcal{X}| \;\; i = \{1, \dots, |\mathcal{X}|\}$; however, this may be sub-optimal when the sources contribute non-uniformly to the Hessian. Xu et al. (2016) propose a non-uniform STN approach derived from matrix sketching techniques (Drineas et al., 2006). If $\mathbf{H}_\mathcal{X}$ refers to the Gauss-Newton Hessian, it can be decomposed in terms of the Jacobian matrices $\mathbf{G}_i = \frac{\partial \mathbf{u}_i(\mathbf{m})}{\partial \mathbf{m}}$ such that

$$\mathbf{H}_\mathcal{X} = \sum_{i \in \mathcal{X}} \mathbf{G}_i^T \mathbf{G}_i + \nabla^2 R(\mathbf{m}). \tag{4.8}$$

A non-uniform probability distribution can be generated from

$$r_i = \frac{\text{trace}(\mathbf{G}_i^T \mathbf{G}_i)}{\text{trace}(\mathbf{G}^T \mathbf{G})} \;\; i = \{1, \dots, \mathcal{X}\}, \tag{4.9}$$

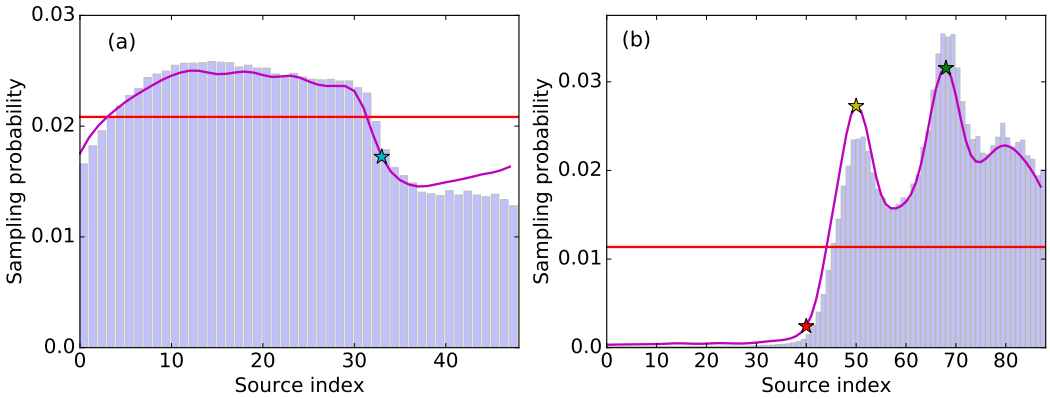

Figure 4.1: Probability distributions used for uniform (red line) and non-uniform (magenta line) sampling for (a) Marmousi II and (b) BP 2.5D models. The blue bars depict probability distributions estimated from stochastic trace estimation (200 random trials) (Hutchinson, 1990). Source indices marked by coloured stars are linked to their source positions in Figures 4.3-4.4 (a).

where $\mathbf{G} = \sum_{i \in \mathcal{X}} \mathbf{G}_i$ (Xu et al., 2016). The sub-Hessians $\mathbf{G}_i^T \mathbf{G}_i$, or their diagonals, are not typically constructed in FWI. In lieu of calculating the true diagonal, we employ the pseudo-Hessian, an approximation to the diagonal of the Gauss-Newton Hessian (Shin et al., 2001). The assumption herein is that the pseudo-Hessian captures the relative contribution of each source to the summed Hessian. To test our assumption, we estimate the true non-uniform sampling distribution (Equation 4.9) using stochastic trace estimates of the Gauss-Newton Hessian (Hutchinson, 1990). The sampling distributions for two test models are displayed in Figure 4.1. The pseudo-Hessian based sampling distribution is similar to the estimated true non-uniform distribution. Before proceeding, we analyse the errors in Hessian-vector products when the subsampled Hessian is used in place of the true one.

We compute the average error (over N_r random trials) $\frac{1}{N_r}\|\mathbf{H}\delta\mathbf{m} - \mathbf{H}_S\delta\mathbf{m}\|^2$ as a function of subset size; the results are displayed in Figure 4.2. While the non-uniform distributions (obtained from the pseudo-Hessian and stochastic trace estimation) are not exactly alike (Figure 4.1), the corresponding average errors are indistinguishable (not displayed). For the Marmousi II model, uniform and non-uniform sampling produce near identical average errors. With the BP 2.5D model, non-uniform sampling clearly provides more accurate estimates of $\mathbf{H}\delta\mathbf{m}$ for a given subset size. The differences between the two models is attributed to the extent of the non-uniformity in the underlying sampling distributions. For example, in Figure 4.1b approximately half of the sources have small, but non-zero, contributions to the Hessian. The distribution in Figure 4.1a shows a comparatively small deviation from the uniform distribution.

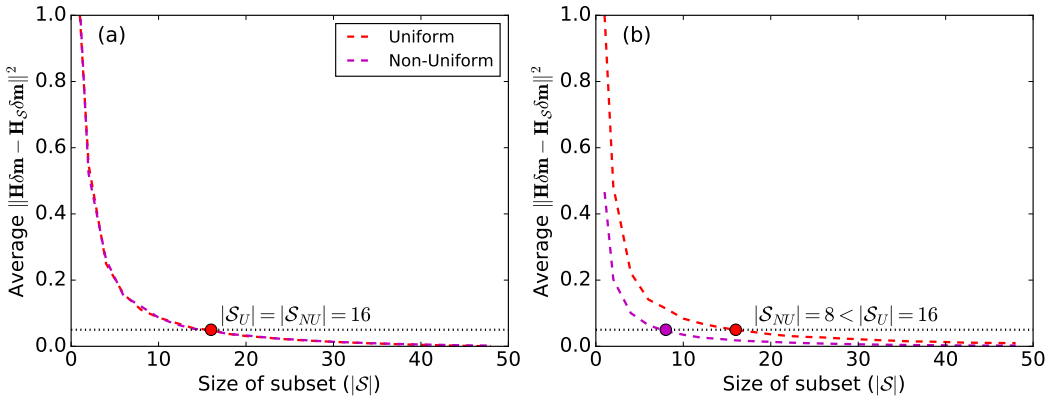

Figure 4.2: Normalized average error in subsampled Hessian-vector products (over 500 random trials) for (a) Marmousi II and (b) BP 2.5D models. Scales are normalized relative to the initial average error ($|\mathcal{S}| = 1$) for uniform sampling. The subset sizes required to achieve an average error of 0.05 (black dotted line) are listed and marked with dots. The Hessian-vector products are computed at the first iteration of either inversion.

The diagonal of the Hessian primarily describes geometrical spreading (Pratt et al., 1998), hence the probabilities can be loosely interpreted as the relative contribution of a source to the overall subsurface illumination. Assuming regularly distributed sources and consistent source signatures, non-uniformity primarily arises from the subsurface model. By correlating dips and peaks in the non-uniform sampling distributions to source positions in the model, we can gain further physical intuition on the influence of the model on non-uniformity in the Hessian. In the Marmousi II model, the marked source (cyan star in Figures 4.1-4.3a) lies in close proximity to two steeply dipping faults. The marked source, and those right of it, generate waves that are promptly scattered out of the model by the faults leading to lower illumination from these sources. The BP 2.D model consists of two distinct subsurface regions (Etgen and Regone, 2005). The left half ($\sim x < 5$ km) is composed of relatively low contrast sedimentary layers, whereas the right half ($\sim x > 5$ km) consists of high contrast layers with complex geometrical structures. Sources overlaying the low contrast region have relatively weak illumination. The high contrast layers induce strong scattering and internal reflections that trap energy and increase illumination. The coloured stars in Figures 4.1-4.3b identify regions of interest.

4.3 Method

The performance of STN, with uniform and non-uniform sampling, is benchmarked against LBFGS and TN algorithms under the framework of isotropic, elastic FWI in the time

domain. The observed and synthetic data are computed using 2D *P-SV* finite difference modelling (fourth order in space, second order in time) (Virieux, 1986; Levander, 1988). Absorbing boundary conditions are implemented in the form of convolutional perfectly matched layers (Komatitsch and Martin, 2007). The source wavelet is assumed to be known and is not estimated. Band-limited noise is added to the data to prevent the inverse crime and to test the performance of the proposed algorithms in the presence of noise. We generate Gaussian white noise arrays for each component of each shot record. The variance of the noise array is set by selecting a desired signal-to-noise ratio (S/N), defined as

$$\text{S/N (dB)} = 10 \log_{10} \left(\frac{a_{rms}^2}{\sigma^2} \right), \tag{4.10}$$

where a_{rms}^2 is the root mean square amplitude of the shot record. For any given shot, the noise arrays of both components (x, z) have equal variance resulting in variable S/N ratios for the two components. The z-component of the noisy data set used for inversion has S/N = 12 dB. Inversion workflows are managed with a customised version of the Seisflows framework (Modrak et al., 2018).

Our TN implementation follows that proposed by Métivier et al. (2013). For the sake of completeness, we review some key components of the algorithm. The CG iterations of TN methods terminate when a maximum number of iterations is reached, or when the Eisenstat-Walker stopping conditions are satisfied (Eisenstat and Walker, 1996; Métivier et al., 2013). The Eisenstat-Walker conditions set a forcing term ζ and terminate CG iterations if

$$\|\mathbf{H}_S \delta\mathbf{m} + \mathbf{g}_\chi\|^2 \leq \zeta \|\mathbf{g}_\chi\|^2. \tag{4.11}$$

The forcing term regulates the accuracy to which the linear system is solved. Its purpose is to prevent overfitting in the event that $J(\mathbf{m})$ is poorly approximated by a quadratic function. The scalar ζ is set dynamically at each FWI iteration with larger values yielding less accurate solutions to Equation 4.6. Crucially, the conditions allow for early termination preventing unnecessary PDE solves. The Gauss-Newton Hessian is symmetric positive semi-definite, violating the requirement of positive-definiteness for linear CG. The Marquadt-Levenberg algorithm guarantees the positive definiteness of the Gauss-Newton Hessian by adding a weighted identity matrix to it (Marquardt, 1963). The full Hessian can have positive or negative eigenvalues and offers no such guarantees even with damping. To overcome this, we test for negative curvature ($\mathbf{x}^T \mathbf{H}_S \mathbf{x} \leq 0$) , $\mathbf{x}$ is an arbitrary vector) at each CG iteration. If detected, CG iterations are terminated and the search direction is replaced by the most recent update direction before negative curvature was detected. If negative curvature is detected at the first CG iteration, a steepest descent direction is taken for the current FWI iteration.

Step lengths are computed with an Armijo backtracking line search for all optimization methods considered (Nocedal and Wright, 2006). Minmax normalization is applied to the inversion parameters to place them into the range [0, 1]. Preconditioning is not used although it can be readily accommodated (e.g. Métivier et al. (2015); Yang et al. (2018a)).

The performance of an algorithm is evaluated through comparisons of the relative cost $\eta = NS_{OPT}/NS_{LBFGS}$ and model error $p\ error^k = \frac{100}{M}\|\frac{\mathbf{m}_p^k - \mathbf{m}_p^*}{\mathbf{m}_p^*}\|_1$, where $\mathbf{m}_p^*$ is the true model for parameter p. The quantity η describes the ratio between the number of PDE solves required by a particular algorithm and LBFGS to reach a target misfit reduction J^*. In synthetic trials, the noise level J^{noise} (squared-norm of the noise) is known. If the data are fit exactly, the misfit converges to the noise level. Based upon this, the target misfit reduction is set as $J^* = J^{noise} + 0.1(J(\mathbf{m}^0) - J^{noise})$. Explicit regularization is omitted in favour of heuristic stabilization techniques; namely, gradient smoothing and limiting the number of iterations. We repeat STN trials 5 times to account for the variability introduced by random source subsampling.

4.4 Numerical experiments

Synthetic inversions are conducted on the Marmousi II (Martin et al., 2006) and BP 2.5D (Etgen and Regone, 2005) models. The Marmousi II model is fully elastic with spatially inconsistent v_p, v_s and ρ models i.e., the three parameters have independent structure and do not represent scaled versions of one another. The BP 2.5D model only provides a v_p model. A density model for the BP 2.5D model is obtained via Gardener's relation; an S-wave velocity is obtained via $v_s = v_p/\sqrt{3}$. Initial models are generated by smoothing the true models with Gaussian kernels with σ=100 m and σ=200 m for the Marmousi II and BP 2.5D model, respectively. The true and initial models are displayed in Figures 4.3-4.4a and b. Simulation and inversion parameters for each example are presented in Table 4.2. Subsets $\mathcal{S}$ are formed from twenty percent of the available sources ($|\mathcal{S}| = 0.2n$) for both trials. For the Marmousi II, $|\mathcal{S}| = 10$ corresponds to average errors of 0.09 for both sampling schemes (Figure 4.2a). In the BP 2.5D trial, $|\mathcal{S}| = 18$ produces average errors of 0.04 and 0.02 for uniform and non-uniform sampling, respectively (Figure 4.2b). Non-uniform probability distributions are kept fixed throughout the inversion. Updating the sampling distributions at each iteration did not alter the performance in preliminary trials. We suspect that this is due to the initial model being relatively close to the true model. In inversions where multi-scale strategies are necessary i.e. when the true and initial models are sufficiently different that cycle skipping occurs, updating the non-uniform probabilities will likely be required to improve performance.

	Marmousi II	BP
Numerical grid		
Dimensions (km)	4 x 1.4	11 x 2
Grid	400 x 140	551 x 101
Spacing ($\Delta x = \Delta z$) (m)	10	20
Acquisition		
No. Src/Rec	48/199	88/135
Src/Rec int. ($\Delta x_s/\Delta x_r$) (m)	80/20	120/80
Src/Rec depth (z_s/z_r) (m)	10/10	20/20
Max offset (km)	3.9	10.6
Simulation		
Time steps	2000	2000
Time interval (Δt) (s)	1.5e-3	2.4e-3
Source wavelet	8 Hz Ricker	5 Hz Ricker
Optimization		
Max iters.	50	100
Max CG iters.	8	8

Table 4.2: Inversion and simulation parameters.

Optimization	Iterations	No. simulations	η
Marmousi II			
LBFGS	21	3408	-
TN	13	7296	2.14
STN (U)	14.0 $\pm$ 1.5	3715 $\pm$ 161	1.09 $\pm$ 0.05
STN (NU)	14.0 $\pm$ 1.3	3760 $\pm$ 336	1.10 $\pm$ 0.10
BP 2.5D			
LBFGS	21	5984	-
TN	8	11000	1.80
STN (U)	14.0 $\pm$ 0.9	6478 $\pm$ 386	1.08 $\pm$ 0.06
STN (NU)	10.8 $\pm$ 1.2	5177 $\pm$ 425	0.86 $\pm$ 0.07

Table 4.3: Summary of inversion statistics evaluated at target misfit reduction J^*. STN trials display mean values and standard deviations computed over 5 independent trials. Uniform and non-uniform sampling trials are indicated by U and NU, respectively. Subset sizes $|\mathcal{S}|$ of 10 and 18 are used for the Marmousi and BP 2.5D trials, respectively.

Inverted models for the Marmousi II and BP 2.5D trials are displayed in Figures 4.3 and 4.4, respectively. Depth profiles of the true and inverted models are presented in Figures 4.5 and 4.6. The convergence properties of the different algorithms are summarized in Table 4.3 and Figure 4.7. Generally, TN trials provide the highest per-iteration convergence rates in data misfit and model error; however, they require around twice the number of PDE solves, compared to LBFGS, to reach J^*. The efficiency is improved for STN which demonstrates misfit and model error reductions similar to TN with a cost increase of less than 10% relative to LBFGS. In Figures 4.3c-e and 4.5, TN and STN inverted models exhibit better focusing at fault boundaries and improved amplitude recovery of model perturbations, most notably at deeper regions of the models. The performance difference between algorithms is more evident in the BP 2.5D inversion results (Figures 4.4 and 4.6). Of note, non-uniform STN appears

to yield improved inversion results when compared to its uniform sampling counterpart. Due to the highly non-uniform contribution of sources to the Hessian (Figure 4.2b), STN with uniform sampling only provides a minor improvement over the LBFGS inverted models. With non-uniform sampling, the inverted models are closer to those obtained with TN while requiring $\sim 15\%$ fewer PDE solves to reach J^*.

To gauge parameter trade-off we examine the inversion results of the Marmousi II model for which v_p, v_s and ρ are spatially inconsistent. For the optimization methods considered, the individual parameters are well resolved and exhibit monotonically decreasing model errors (Figure 4.7). Figures 4.3 and 4.5 do not indicate strong inter-parameter mappings where distinct features of one parameter have been discernibly mapped into another. Models inverted using TN/STN recover the amplitudes of model perturbations slightly better than LBFGS and achieve smaller model errors at any given iteration. The results imply that the parameter set is well decoupled for the chosen inversion configuration (acquisition geometry, offset ranges, frequency band etc.). The utility of the Hessian in mitigating parameter trade-off is more apparent in cases where the parameters are strongly coupled; one such example is presented by Yang et al. (2018a).

The performance of STN as a function of certain parameters (e.g., $N_{CG}, |\mathcal{S}|$) has not been tested, in part due to the computational expense of the endeavour. Furthermore, the implicit dependence of the optimization on the initial and true models means that the range of acceptable parameters may vary between test cases. A heuristic approach to determining sufficient subset sizes and data redundancy could be via the analysis presented in Figure 4.2. Limited tests using the full Hessian in uniform STN demonstrated similar performance gains as STN using the Gauss-Newton Hessian. Non-uniform STN with the full Hessian remains to be tested. Pseudo-Hessian based non-uniform sampling is only meaningful if the full Hessian is dominated by the Gauss-Newton component. If this is not the case, alternative strategies may be required to obtain representative trace estimates. Stochastic trace estimation could be used; however, it requires repeated applications of the Hessian to random vectors making it potentially expensive.

4.5 Conclusions

A subsampled truncated Newton has been proposed to ameliorate the computational demands of second-order optimization methods. Source subsampling employed in the computation of Hessian-vector products, reduces the number of PDE solves required at each iteration. The STN approach exhibits convergence rates comparable to TN methods while requiring a similar number of PDE solves as the LBFGS method to reach a target misfit. Non-uniform sampling outperformed uniform sampling in the BP 2.5D trials, but did not

produce a discernible difference in the Marmousi II trials. The results suggest that non-uniform sampling is beneficial when sources contribute in a highly non-uniform manner to the Hessian. The extent of non-uniformity required (in the Hessian) to make non-uniform sampling preferable, is unclear and a topic for future research. Our results demonstrate that second-order information can be accommodated into FWI with only minor algorithmic changes.

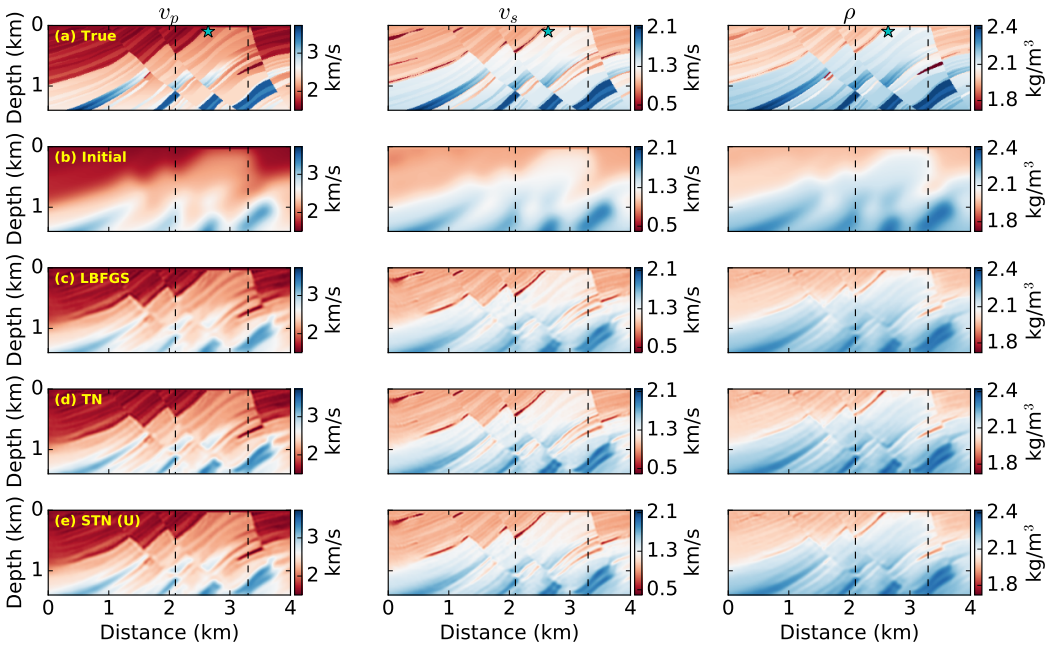

Figure 4.3: Marmousi II trials: (a) True model. (b) Initial model. (c-e) Models inverted after 21 non-linear iterations. Inverted models are compared at the iteration number where LBFGS reaches the target misfit J^*. Non-uniform STN results are not displayed due to their similarity with uniform STN. Dashed black lines depict the location of depth-profiles presented in Figure 4.5.

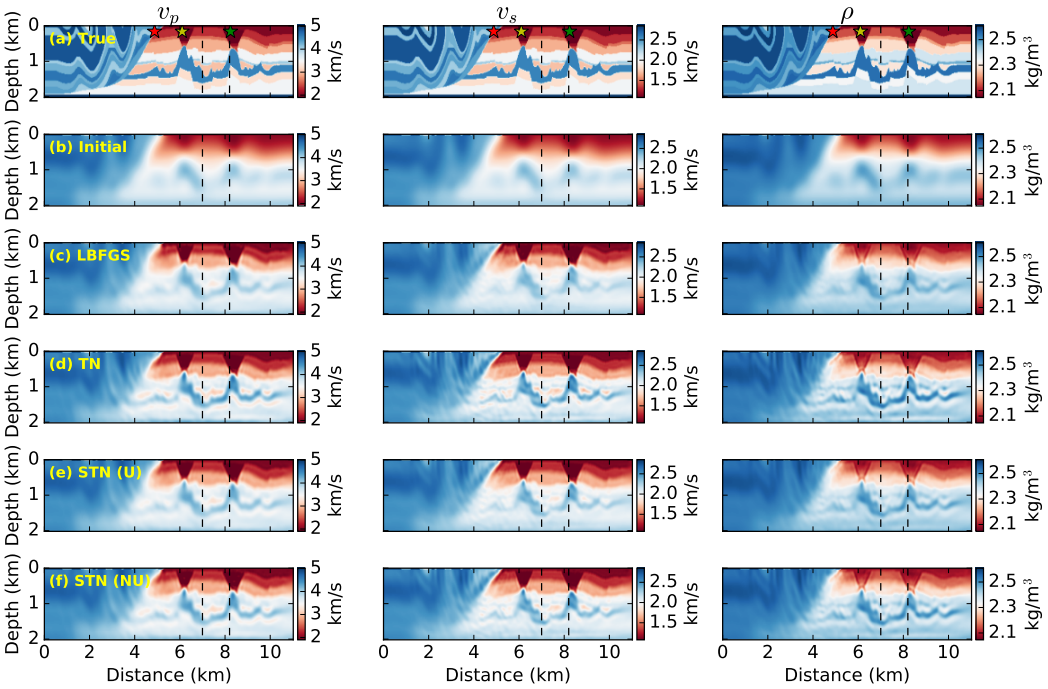

Figure 4.4: BP 2.5D trials: (a) True model. (b) Initial model. (c-f) Models inverted after 22 non-linear iterations. Inverted models are compared at the iteration number where LBFGS reaches the target misfit J^*. Dashed black lines depict the location of depth-profiles presented in Figure 4.6.

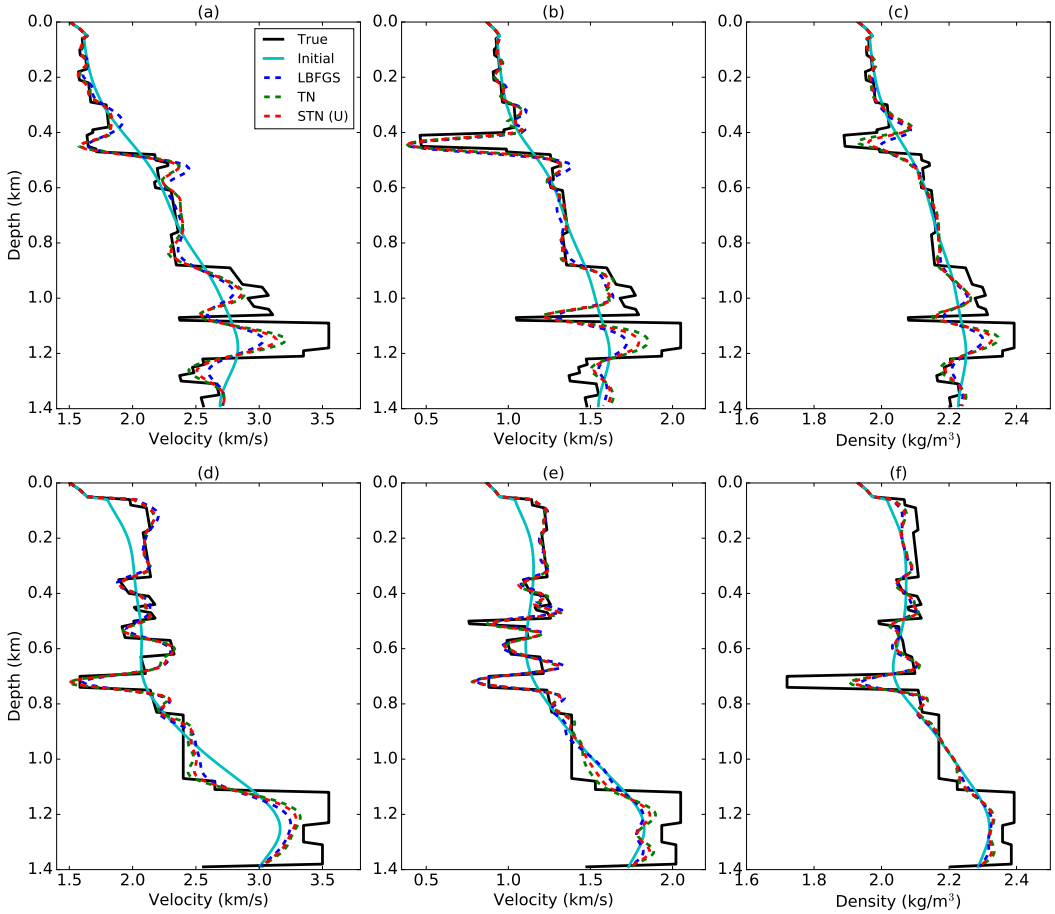

Figure 4.5: Marmousi II trials: Depth profiles for (a, d) v_p, (b, e) v_s and (c, f) ρ after 21 iterations. Profiles are taken at (a-c) $x = 2.1$ km and (d-f) $x = 3.3$ km.

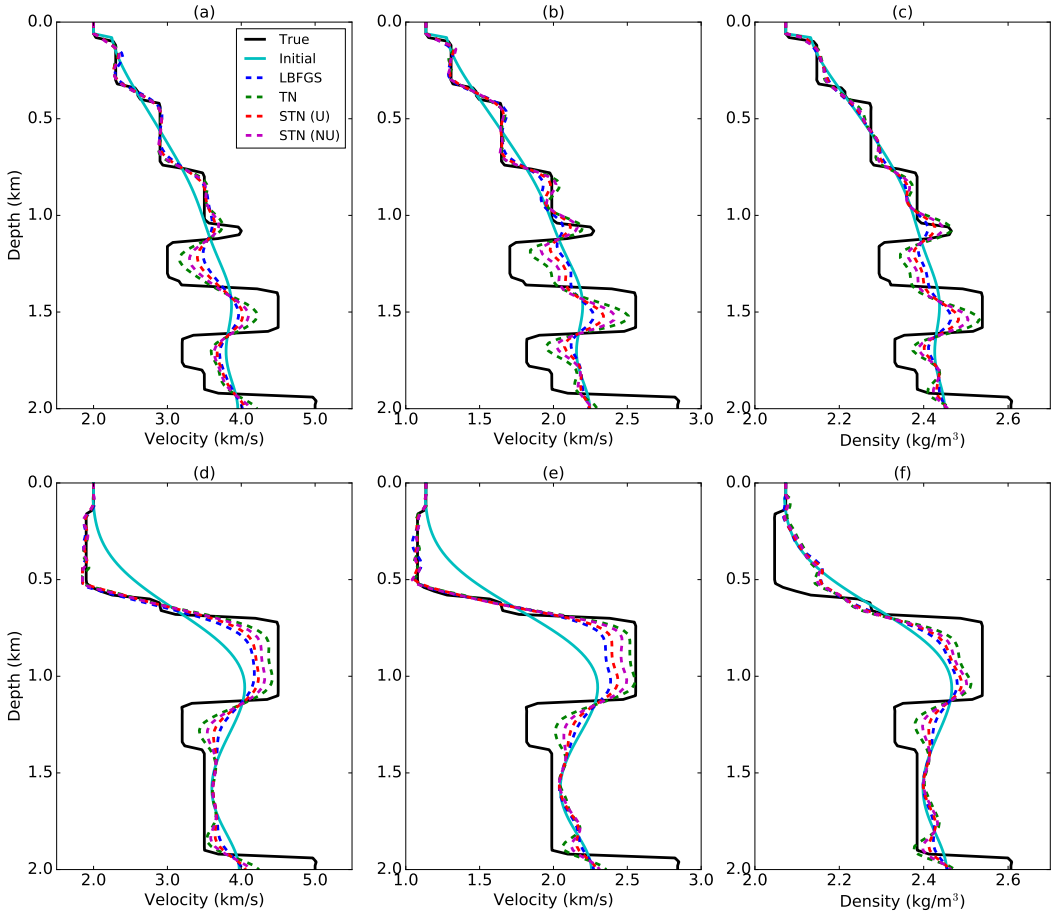

Figure 4.6: BP 2.5D trials: Depth profiles for (a, d) v_p, (b, e) v_s and (c, f) ρ after 22 iterations. Profiles are taken at (a-c) $x = 7.0$ km and (d-f) $x = 8.2$ km.

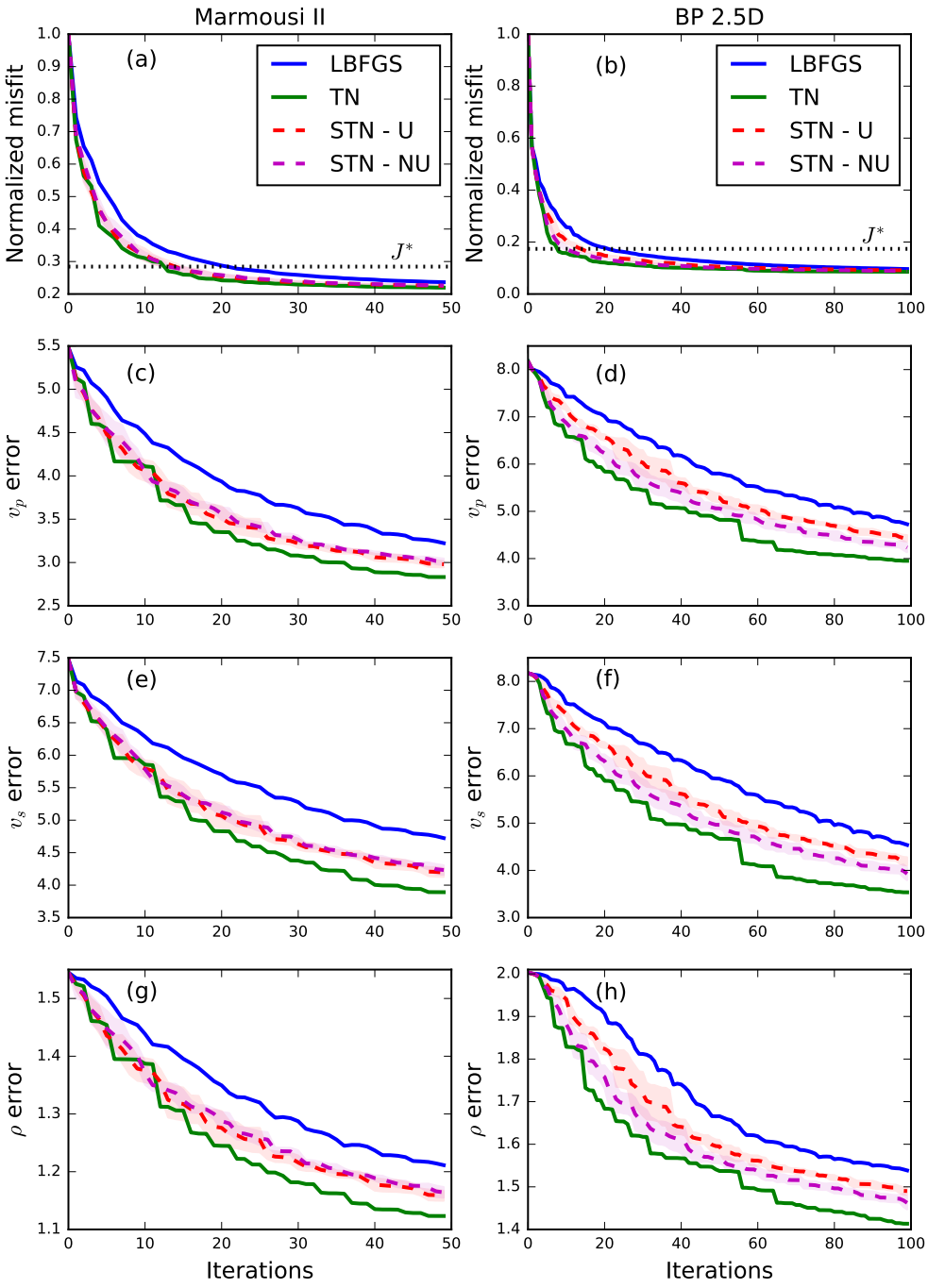

Figure 4.7: Convergence behaviour as a function of iteration number for (a, c, d, g) Marmousi II (b, d, f, h) BP 2.5D experiments. (a, b) Normalized misfit. The target misfit J^* is marked with a dotted black line. (c, d) v_p model error. (e, f) v_s model error. (g, h) ρ model error. Dashed lines for the subsampled trials represent mean values computed over 5 independent trials; error bands represent one standard deviation.

CHAPTER 5

Resolution analysis in FWI using the Kronecker-factored Hessian[1]

5.1 Introduction

Advanced techniques such as reverse-time migration and full waveform inversion (FWI) have led to a step change in the resolving capability of seismic imaging. Much attention has been given to the development of these algorithms to produce high-resolution models of the Earths subsurface; however, essential tools for rigorous resolution and uncertainty analysis remain underdeveloped. Backus and Gilbert (1968); Backus et al. (1970) pioneered early concepts of resolution for linear and non-linear geophysics inverse problems. Probabilistic formulations of the non-linear inverse problem seek posterior model distributions that are highly complex and multi-modal. Global search algorithms offer a natural approach to exploring complex distributions as they sample different regions of the posterior in search of a maximum likelihood model. Early studies using global search methods include Monte Carlo search using the Metropolis algorithm (Mosegaard and Tarantola, 1995; Sambridge and Mosegaard, 2002; Tarantola, 2005), neighbourhood search (Sambridge, 1999a,b), genetic algorithms (Gallagher et al., 1991; Stoffa and Sen, 1991; Sambridge and Drijkoningen, 1992), and simulated annealing (Mosegaard and Vestergaard, 1991; Sen and Stoffa, 1991). The immense size of the model space in 2D/3D problems coupled with the prohibitive computational cost of repeatedly evaluating the FWI objective function precludes FWI from global optimization methods using current compute systems.

Subsequent studies in FWI have simplified the formulation of the inverse problem to make

[1] A version of this chapter is being considered for a journal submission. The material has been presented at conferences.

resolution analysis more feasible. Tarantola (2005) formulates a Bayesian inverse problem applicable to the regime where the forward modelling operator can be linearized. Rawlinson et al. (2014) reviews popular techniques suitable for seismic tomography, including the use of resolution matrices for spike tests. In a similar vein, Hessian probing has been used to extract local resolution information in global tomography (Fichtner and Trampert, 2011b; Trampert et al., 2013; Fichtner and Leeuwen, 2015; Rawlinson and Spakman, 2016). Bui-Thanh et al. (2013) perform linearized Bayesian inversion for 3D global tomography using hundreds of thousands of model parameters. They utilize a low-rank approximation of the Hessian that facilitates efficient sampling of a Gaussian posterior distribution. Zhu et al. (2015) apply the same computational framework to waveform inversion in exploration seismology. Fang et al. (2018) apply Bayesian inversion to a penalty formulation of FWI. Ely et al. (2018) use a Metropolis-Hastings algorithm and a fast local solver to sample the posterior distribution. Thurin et al. (2019) use an ensemble data assimilation technique based on ensemble Kalman filters to quantify uncertainty in frequency-domain FWI.

Resolution analyses in FWI, either involving resolution operators or posterior model covariances (in Bayesian formulations), are challenging as they involve the Hessian. The computational cost associated with computing and storing the Hessian has prevented extensive research on the topic. We employ a computationally efficient approximation of the Hessian as a superposition of Kronecker products (Gao et al., 2020). The factorization yields a Hessian approximation that honours its block, banded-diagonal structure. The factor matrices are small relative to the size of the Hessian. Hessian-vector products are approximated through a series of computationally inexpensive matrix multiplications involving relatively small matrices. In this chapter we exploit the Kronecker-based factorization of the Hessian to perform local resolution analyses and linearized Bayesian inversion. We first formulate the Bayesian inverse problem before reviewing the Kronecker-based factorization of the Hessian. We subsequently explore properties of the Kronecker factors as directional blurring operators. We present a numerical example for the acoustic Marmousi model. In the local resolution analysis, we probe the Hessian to extract horizontal and vertical resolution lengths at various points in the subsurface. A low-rank approximation of the Hessian is used to sample the posterior model distribution and compute sample standard deviations. We observe that resolution deteriorates at deeper regions of the model where the illumination quality from the surface acquisition decreases.

5.2 Theory

5.2.1 Resolution analysis in linear problems

We initially explore resolution analysis for linear inverse problems as a prelude to 'local resolution analysis', a simplified non-linear analogue. Consider the linear system

$$\mathbf{Gm} = \mathbf{d}, \tag{5.1}$$

where $\mathbf{G} \in \mathbb{R}^{N \times M}$ is a modelling kernel, $\mathbf{m} \in \mathbb{R}^{M \times 1}$ are the model parameters, or unknowns, and $\mathbf{d} \in \mathbb{R}^{N \times 1}$ are the observations; N and M are the number of observations and model parameters, respectively. For observations $\mathbf{d}^{obs} \in \mathbb{R}^{N \times 1}$, the estimated model parameters can be computed via

$$\mathbf{m}^{est} = \mathbf{G}^{-g}\mathbf{d}^{obs}, \tag{5.2}$$

where $\mathbf{G}^{-g}$ represents the generalized inverse. The generalized inverse is associated with the minimum norm solution for underdetermined problems, or the least-squares solution for overdetermined problems e.g. $\mathbf{G}^{-g} = (\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T$. Substituting Equation 5.1 into Equation 5.2 yields

$$\mathbf{m}^{est} = \mathbf{G}^{-g}\mathbf{G}\mathbf{m}^{true} = \mathbf{R}\mathbf{m}^{true}, \tag{5.3}$$

where $\mathbf{R} \in \mathbb{R}^{M \times M}$ is the resolution matrix (Backus et al., 1970; Menke, 1984). Each row of $\mathbf{R}$ corresponds to a filtering kernel for a discrete location in the model space. When the resolution matrix is $\mathbf{R} = \mathbf{I}$, where $\mathbf{I}$ is the identity, the exact model can be recovered.

For non-linear inverse problems, an analogous interpretation of resolution matrices can be derived by first assuming convergence to the global minimum. The objective function $J(\mathbf{m})$ can be Taylor expanded by applying a small perturbation $\delta\mathbf{m}$ to the optimal model $\mathbf{m}^*$,

$$J(\mathbf{m}^* + \delta\mathbf{m}) = J(\mathbf{m}^*) + \delta\mathbf{m}^T\mathbf{g}(\mathbf{m}^*) + \frac{1}{2}\delta\mathbf{m}^T\mathbf{H}(\mathbf{m}^*)\delta\mathbf{m} + \mathcal{O}(\delta\mathbf{m}^3), \tag{5.4}$$

where $\mathbf{g}(\mathbf{m}^*) = \frac{\partial J}{\partial \mathbf{m}}\big|_{\mathbf{m}=\mathbf{m}^*}$ and $\mathbf{H}(\mathbf{m}^*) = \frac{\partial^2 J}{\partial \mathbf{m}^2}\Big|_{\mathbf{m}=\mathbf{m}^*}$ are the gradient and the Hessian of the objective function, respectively. Given that the gradient at the minimum vanishes ($\mathbf{g}(\mathbf{m}^*) = 0$), a steepest descent step taken from the perturbed state $\mathbf{m}^* + \delta\mathbf{m}$ results in the update formula

$$\tilde{\mathbf{m}} = (\mathbf{m}^* + \delta\mathbf{m}) - \nu\mathbf{H}\delta\mathbf{m}, \tag{5.5}$$

given a scalar step length ν. The gradient update is the true model perturbation scaled by the Hessian matrix. In FWI, the Hessian has been demonstrated to act as a low-pass filter (Pratt, 1999; Fichtner and Leeuwen, 2015). Fichtner and Leeuwen (2015) estimate vertical

and horizontal resolution lengths by repeatedly applying the Hessian to random perturbations. This form of local resolution analysis can also be obtained within the framework of linearized Bayesian inversion which is explored in the next section.

5.2.2 Linearized Bayesian inversion

The ill-posedness of the FWI inverse problem means that an infinite number of models can fit the data. In spite of this, it remains common practice for researchers to present a single velocity model as the inversion result. Contrary to this practice, Tarantola (2005) advocates for inversions results in the form of probability distributions or model ensembles. With a probability distribution, numerous models can be sampled and used to characterize uncertainty. We outline the Bayesian formulation of FWI made more tractable through a series of simplifying assumptions.

The posterior probability distribution is a conditional distribution that characterizes the probability of a model $\mathbf{m}$ given observed data $\mathbf{d}$. Neglecting constants, the posterior is proportional to the product of the model-prior and the likelihood

$$\rho(\mathbf{m}|\mathbf{d}) \propto \rho(\mathbf{d}|\mathbf{m})\rho(\mathbf{m}), \tag{5.6}$$

where $\rho(\mathbf{m})$ and $\rho(\mathbf{d}|\mathbf{m})$ denote the prior probability and likelihood, respectively. Assuming Gaussian priors, the model prior and likelihood are expressed as

$$\rho(\mathbf{m}) \propto \exp\left[-(\mathbf{m}-\mathbf{m}_0)\mathbf{C}_\mathbf{m}^{-1}(\mathbf{m}-\mathbf{m}_0)\right], \tag{5.7}$$

and

$$\rho(\mathbf{d}|\mathbf{m}) \propto \exp\left[-(\mathbf{u}(\mathbf{m})-\mathbf{d})^T\mathbf{C}_\mathbf{d}^{-1}(\mathbf{u}(\mathbf{m})-\mathbf{d})\right], \tag{5.8}$$

where $\mathbf{C}_\mathbf{m}^{-1} \in \mathbb{R}^{M\times M}$ and $\mathbf{C}_\mathbf{d}^{-1} \in \mathbb{R}^{N\times N}$ are the inverse model and data covariances, respectively. The use of Gaussian priors imposes certain assumptions on the data residuals and model perturbations. In the presence of non-Gaussian noise or large modelling errors, the data residuals are unlikely to be Gaussian and zero-mean. Likewise, the model prior assumes the Earth model can be approximated as Gaussian perturbations imposed on an initial background model. If the subsurface exhibits large velocity contrasts, as is the case when salt bodies are present, this assumption becomes invalid. For this study, we construct scenarios where these assumptions remain reasonable. However, when these assumptions are violated the prior distributions are no longer reasonably approximated by a Gaussian distribution and the formulation will not be suitable. Substituting Equations 5.7 and 5.8

into Equation 5.6 results in the following expression for the posterior distribution

$$\rho(\mathbf{m}|\mathbf{d}) \propto \exp[-(\mathbf{u}(\mathbf{m})-\mathbf{d})^T\mathbf{C}_\mathbf{d}^{-1}(\mathbf{u}(\mathbf{m})-\mathbf{d}) - (\mathbf{m}-\mathbf{m}_0)^T\mathbf{C}_\mathbf{m}^{-1}(\mathbf{m}-\mathbf{m}_0)]. \tag{5.9}$$

In the probabilistic interpretation, the objective of an inverse problem is to find the most likely model given some observed data. Mathematically, this equates to maximizing the log-posterior distribution

$$\arg\max_{\mathbf{m}} \ \log \rho(\mathbf{m}|\mathbf{d}). \tag{5.10}$$

Applying the logarithm to Equation 5.9 converts the maximization problem to a minimization of a regularized least-squares objective:

$$J(\mathbf{m}) = \frac{1}{2}\|\mathbf{u}(\mathbf{m})-\mathbf{d}\|^2_{\mathbf{C}_\mathbf{d}^{-1}} + \frac{1}{2}\|\mathbf{m}-\mathbf{m}_0\|^2_{\mathbf{C}_\mathbf{m}^{-1}}, \tag{5.11}$$

where the norms are defined as $\|\mathbf{w}\|^2_{\mathbf{C}_\mathbf{m}^{-1}} = \mathbf{w}^T\mathbf{C}_\mathbf{m}^{-1}\mathbf{w}$ for an arbitrary vector $\mathbf{w}$. The solution of Equation 5.11 can be obtained using iterative gradient-based optimization techniques. In the context of FWI, gradients can be calculated using the adjoint-state method (Plessix, 2006). The gradient of Equation 5.11 is

$$\frac{\partial J}{\partial \mathbf{m}} = \mathbf{G}^T\mathbf{C}_\mathbf{d}^{-1}(\mathbf{u}(\mathbf{m})-\mathbf{d}) + \mathbf{C}_\mathbf{m}^{-1}(\mathbf{m}-\mathbf{m}_0), \tag{5.12}$$

where $\mathbf{G} = \frac{\partial \mathbf{u}(\mathbf{m})}{\partial \mathbf{m}} \in \mathbb{R}^{N\times M}$ denotes the Fréchet derivative matrix. The corresponding Hessian, after neglecting terms involving second order derivatives in $\mathbf{u}(\mathbf{m})$, is defined as

$$\mathbf{H} = \mathbf{G}^T\mathbf{C}_\mathbf{d}^{-1}\mathbf{G} + \mathbf{C}_\mathbf{m}^{-1}. \tag{5.13}$$

Equation 5.13 is the Gauss-Newton form of the Hessian. For the remainder of this chapter, we use Hessian to exclusively refer to the Gauss-Newton approximation. Following conventional gradient-based optimization schemes, the model update can be obtained from the solution of the Newton system

$$(\mathbf{G}^T\mathbf{C}_\mathbf{d}^{-1}\mathbf{G} + \mathbf{C}_\mathbf{m}^{-1})\delta\mathbf{m} = -\mathbf{G}^T\mathbf{C}_\mathbf{d}^{-1}(\mathbf{u}(\mathbf{m})-\mathbf{d}) - \mathbf{C}_\mathbf{m}^{-1}(\mathbf{m}-\mathbf{m}_0). \tag{5.14}$$

Equation 5.14 can be used to derive a local resolution analysis for the Bayesian formulation (Bosch et al., 2005). First, we assume that the observed data $\mathbf{d} = \mathbf{u}(\mathbf{m}^{true})$. From there, we consider a perturbation to the true model such that $\mathbf{m}^{true} \rightarrow \mathbf{m}^{true} + \delta\mathbf{m}^{true}$ and $\mathbf{d}^{obs} \rightarrow \mathbf{d}^{obs} + \delta\mathbf{d}^{true} \approx \mathbf{d}^{obs} + \mathbf{G}\delta\mathbf{m}^{true}$. Here we use the Born approximation and assume that the data perturbation $\delta\mathbf{d}^{true}$ is linearly related to the model perturbation. Inserting

the perturbed forms into Equation 5.14 yields

$$(\mathbf{G}^T\mathbf{C}_\mathbf{d}^{-1}\mathbf{G}+\mathbf{C}_\mathbf{m}^{-1})(\delta\mathbf{m}+\delta\mathbf{m}') = -\mathbf{G}^T\mathbf{C}_\mathbf{d}^{-1}(\mathbf{u}(\mathbf{m})-\mathbf{d}-\mathbf{G}\delta\mathbf{m}^{true})-\mathbf{C}_\mathbf{m}^{-1}(\mathbf{m}-\mathbf{m}_0), \quad (5.15)$$

where $(\delta\mathbf{m}+\delta\mathbf{m}')$ represents the perturbed second-order model update. Subtracting Equation 5.14 from Equation 5.15 results in

$$(\mathbf{G}^T\mathbf{C}_\mathbf{d}^{-1}\mathbf{G}+\mathbf{C}_\mathbf{m}^{-1})\delta\mathbf{m}' = \mathbf{G}^T\mathbf{C}_\mathbf{d}^{-1}\mathbf{G}\delta\mathbf{m}^{true}. \quad (5.16)$$

Equation 5.16 states that an estimate $\delta\mathbf{m}'$ of the true perturbation $\delta\mathbf{m}^{true}$ is obtained through the solution of a linear system. For illustration, setting $\mathbf{C}_\mathbf{m} = \mathbf{C}_\mathbf{d} = \mathbf{I}$ reduces the previous equation to

$$\delta\mathbf{m}' = (\mathbf{G}^T\mathbf{G}+\mathbf{I})^{-1}\mathbf{G}^T\mathbf{G}\delta\mathbf{m}^{true}, \quad (5.17)$$

$$(5.18)$$

The term $\mathbf{G}^T\mathbf{G}$ is precisely the Gauss-Newton Hessian in conventional FWI. The damped inverse Hessian acts as a focusing operator (Pratt et al., 1998). Fichtner and Leeuwen (2015) suggest that a conservative estimate of resolution is obtained by computing the action of the Hessian on a model perturbation. In this case the (damped) inverse Hessian is replaced with an identity and $\delta\mathbf{m}' = \mathbf{H}\delta\mathbf{m}^{true}$.

Sampling the posterior

Local resolution analysis does not permit a probabilistic interpretation of the subsurface. To enable this, we estimate, then sample from, the posterior distribution. We initially assume that the observed data can be represented as a small perturbation from the modelled data at the maximum a posteriori (MAP) model $\mathbf{m}^*$ and that the perturbation is linearly related to a model perturbation such that

$$\mathbf{d}^{obs} \approx \mathbf{u}(\mathbf{m}^*) + \mathbf{G}(\mathbf{m}-\mathbf{m}^*). \quad (5.19)$$

Following the linearization, Tarantola (2005) demonstrates that the posterior distribution is also a Gaussian distribution of the form

$$\rho(\mathbf{m}|\mathbf{d}) \propto \exp\left[-(\mathbf{m}-\mathbf{m}^*)\tilde{\mathbf{C}}_\mathbf{m}^{-1}(\mathbf{m}-\mathbf{m}^*)\right], \quad (5.20)$$

where $\tilde{\mathbf{C}}_\mathbf{m}$ is the posterior covariance. For Gaussian priors, the posterior is given by

$$\tilde{\mathbf{C}}_\mathbf{m} = (\mathbf{G}^T\mathbf{C}_\mathbf{d}^{-1}\mathbf{G}+\mathbf{C}_\mathbf{m}^{-1})^{-1}. \quad (5.21)$$

Referring to Equation 5.13, it is apparent that the posterior covariance is equivalent to the inverse Hessian. Before further discussing the posterior distribution, it warrants mentioning that the goal is to produce model ensembles and quantify uncertainties. A sample from the posterior distribution can be computed via

$$\mathbf{m}_s = \mathbf{m}^* + \tilde{C}_{\mathbf{m}}^{\frac{1}{2}}\mathbf{n}, \quad \mathbf{n} \sim \mathcal{N}(0,1), \tag{5.22}$$

where $\mathbf{n} \in \mathbb{R}^{N\times 1}$ is a random vector drawn from a zero-mean Gaussian with unit variance.

Assuming we obtain the MAP model, sampling and characterizing the posterior remains a challenge. The most apparent issue is that the Fréchet derivative matrix (or operator) $\mathbf{G}$ is not typically computed during FWI. By not computing the Fréchet derivatives, direct computation of the Hessian is also avoided. For inverse problems with large model spaces, representing covariances as matrices is also no longer practical. It is more common to consider covariance operators that act on vectors. For these reasons, computing the posterior covariance as defined in Equation 5.21 is not feasible. By extension, the square root (posterior) operator required to sample from the posterior is also not readily available.

Low-rank approximation

Here we review a low-rank approximation of the prior-preconditioned Hessian that permits efficient sampling of the posterior distribution (Bui-Thanh et al., 2013). For a thorough description, readers are referred to the original manuscript. The first step utilizes an alternative expression for the posterior covariance (Equation 5.21),

$$\tilde{\mathbf{C}}_{\mathbf{m}} = \mathbf{C}_{\mathbf{m}}^{\frac{1}{2}}(\mathbf{C}_{\mathbf{m}}^{\frac{1}{2}}\mathbf{G}^T\mathbf{C}_{\mathbf{d}}^{-1}\mathbf{G}\mathbf{C}_{\mathbf{m}}^{\frac{1}{2}} + \mathbf{I})^{-1}\mathbf{C}_{\mathbf{m}}^{\frac{1}{2}}. \tag{5.23}$$

The first term within the parentheses is related to the data misfit and is termed the prior-preconditioned Hessian by Bui-Thanh et al. (2013). Bui-Thanh et al. (2013) approximate the prior-preconditioned Hessian with a truncated singular value decomposition:

$$\mathbf{C}_{\mathbf{m}}^{\frac{1}{2}}\mathbf{G}^T\mathbf{C}_{\mathbf{d}}^{-1}\mathbf{G}\mathbf{C}_{\mathbf{m}}^{\frac{1}{2}} \approx \mathbf{V}_r\mathbf{\Gamma}_r\mathbf{V}_r^T, \tag{5.24}$$

where $\mathbf{\Gamma}_r \in \mathbb{R}^{r\times r}$ is a diagonal matrix of the r largest singular values and $\mathbf{V}_r \in \mathbb{R}^{M\times r}$ is a matrix whose columns contain the r^{th} largest singular vectors. Inserting the low-rank approximation into Equation 5.23 and applying the Sherman-Morrison-Woodbury formula results in

$$(\mathbf{V}_r\mathbf{\Gamma}_r\mathbf{V}_r^T + \mathbf{I})^{-1} = \mathbf{I} - \mathbf{V}_r\mathbf{D}_r\mathbf{V}_r^T, \tag{5.25}$$

where $\mathbf{D}_r = diag(\frac{\lambda_i}{\lambda_i+1}), i = 1, \cdots, r \in \mathbb{R}^{r\times r}$. An approximation of the posterior covariance may then be expressed as

$$\tilde{\mathbf{C}}_{\mathbf{m}} = \mathbf{C}_{\mathbf{m}} - \mathbf{C}_{\mathbf{m}}^{\frac{1}{2}}(\mathbf{V}_r\mathbf{D}_r\mathbf{V}_r^T)\mathbf{C}_{\mathbf{m}}^{\frac{1}{2}} = \mathbf{L}\mathbf{L}^T \tag{5.26}$$

for a factor matrix $\mathbf{L}$. Bui-Thanh et al. (2013) demonstrate that the factor $\mathbf{L}$ of the posterior covariance is given by

$$\mathbf{L} = \mathbf{C}_{\mathbf{m}}^{\frac{1}{2}}(\mathbf{V}_r\mathbf{P}_r\mathbf{V}_r^T + \mathbf{I}), \tag{5.27}$$

where $\mathbf{P}_r = diag(\frac{1}{\sqrt{\lambda_i+1}} - 1), i = 1, \cdots, r \in \mathbb{R}^{r\times r}$.

Computing the low-rank approximation requires calculating the r^{th} largest singular vectors of the prior-preconditioned Hessian. This is achieved using Lanczos iterations where only the action of the (prior-preconditioned) Hessian on a vector is required. In FWI, this is conventionally done using the second-order adjoint state method to compute Hessian-vector products (Fichtner and Trampert, 2011a). We explore an alternative approach that first approximates the Hessian as a superposition of Kronecker products. The factorization permits fast Hessian-vector products through a sequence of inexpensive matrix multiplications.

5.3 Implementation details

In this section, we address additional theory related to the specific implementation of linearized Bayesian inversion presented in this study. Firstly, we discuss a preconditioned formulation of the iterative gradient-based minimization. Secondly, we establish a prior model covariance that allows for the inclusion of structural constraints. Finally, we introduce a novel factorization of the Hessian that facilitates fast-Hessian vector products required for the Lanczos iterations.

5.3.1 Preconditioned formulation

The regularized least-squares objective Equation 5.11 is impractical within the framework of FWI for a number of reasons. Regularized inverse problems require tunable hyperparameters to balance gradient contributions that come from the data misfit and model regularization terms. The scaling is implicit in Equation 5.11 and is embedded within the model and data covariance matrices. For demonstration, consider model and data covariances that are uncorrelated such that $\mathbf{C}_{\mathbf{d}}^{-1} = \frac{1}{\sigma_{\mathbf{d}}^2}\mathbf{I}$ and $\mathbf{C}_{\mathbf{m}}^{-1} = \frac{1}{\sigma_{\mathbf{m}}^2}\mathbf{I}$, where $\sigma_{\mathbf{d}}^2$ and $\sigma_{\mathbf{m}}^2$ are the data and model variances, respectively. The original objective function can be simplified to

$$J(\mathbf{m}) = \frac{1}{2}\|\mathbf{u}(\mathbf{m}) - \mathbf{d}\|_2^2 + \frac{\mu}{2}\|\mathbf{m} - \mathbf{m}_0\|_2^2, \tag{5.28}$$

where $\mu = \frac{\sigma_{\mathbf{d}}^2}{\sigma_{\mathbf{m}}^2}$. In this scenario, the misfit terms are balanced by a ratio of the variances; however, the variances are typically not known. In practice, the value of μ is selected heuristically by balancing the data and model misfits. The objective is to ensure that data misfit is the primary contributor to the gradient while the regularization term provides an ancillary contribution. Testing different values of hyperparameters is expensive due to the high computational cost of a single FWI iteration. Furthermore, as an inversion progresses the data misfit will decrease while the model misfit will likely increase due to the perturbations added to the initial model. The balance between the two terms changes throughout the inversion suggesting that a constant hyperparameter may be suboptimal.

A more practical approach reformulates the regularized form into a preconditioned one. The first step involves a change of variables where $\hat{\mathbf{m}} = \mathbf{C}_{\mathbf{m}}^{-\frac{1}{2}}(\mathbf{m} - \mathbf{m_0})$ (Guitton et al., 2012). The regularizing term in Equation 5.11 then becomes

$$\frac{1}{2}\|(\mathbf{m} - \mathbf{m}_0)\|^2_{\mathbf{C}_{\mathbf{m}}^{-1}} = \frac{1}{2}\|\hat{\mathbf{m}}\|_2^2. \tag{5.29}$$

The new variable $\hat{\mathbf{m}}$ is updated using a gradient step,

$$\hat{\mathbf{m}}^{k+1} = \hat{\mathbf{m}}^k + \nu^k \frac{\partial J}{\partial \hat{\mathbf{m}}}. \tag{5.30}$$

Expanding out the variables and applying the chain rule leads to a preconditioned update for the original model vector $\mathbf{m}$

$$\mathbf{m}^{k+1} = \mathbf{m}^k + \nu^k \mathbf{C}_{\mathbf{m}} \frac{\partial J}{\partial \mathbf{m}}. \tag{5.31}$$

In the preconditioning formulation, we neglect the formal gradient that would arise from the $\frac{1}{2}\|\hat{\mathbf{m}}\|_2^2$ term as its effect can be mimicked by truncating the number of iterations in an inversion.

In the context of conventional regularized inverse problems, $\mathbf{C}_{\mathbf{m}}^{-\frac{1}{2}}$ represents an operator $\mathbf{W}$ that penalizes undesirable features in the model. For example, a spatial derivative operator promotes smooth models by penalizing discontinuities. The inverse of the prior covariance can be expressed as $\mathbf{C}_{\mathbf{m}}^{-1} = \mathbf{W}^T\mathbf{W}$. Conversely, the prior covariance $\mathbf{C}_{\mathbf{m}} = \mathbf{L}\mathbf{L}^T$ where $\mathbf{L} = \mathbf{W}^{-1}$. A preconditioning operator has the interpretation of promoting desirable features in a model.

5.3.2 Prior covariance

The prior covariance operator embeds prior knowledge about the subsurface model into the inversion. In the simplest case, $\mathbf{C}_{\mathbf{m}} = \mathbf{I}$ and the model perturbations are uncorrelated.

This assumption states that any given point in the subsurface exhibits no correlation with other points in the subsurface. Such an assumption is unrealistic due to the prevalence of structured media in the subsurface e.g., in the form of layered media with a preferred orientation. Subsurface points within some vicinity or within a single geological unit (under similar physical conditions) likely possess similar rock properties, thus are correlated. We enforce this interpretation with preconditioning operators that promote structural coherence. Specifically, the preconditioner is designed to smooth the model in directions consistent with the geology. Structural orientation can be inferred by computing structure tensors from seismic images. Structure-oriented preconditioners have been used in a number of studies in the past (Guitton et al., 2012; Bui-Thanh et al., 2013; Li et al., 2016; Trinh et al., 2017). For this study, we employ an adapted form of a Matérn covariance (or shaping covariance) proposed by Hale (2014). Applying the prior covariance operator to a vector requires solving an anisotropic PDE, the cost of which is small compared to a solution of the wave equation. The operator is characterized by a user-defined scale length (r_0) that defines the spatial extent of correlations/smoothing. Larger scale lengths correspond to increased correlations/smoothing. A more detailed description of the shaping covariance operator is presented in Appendix A.

5.3.3 The Kronecker-factored Hessian

The local resolution analysis and linearized Bayesian inversion require repeated applications of the Hessian to a vector. Hessian-vector products can be computed using second-order adjoint-state methods (Fichtner and Trampert, 2011a). The computational cost of a single Hessian-vector products is $4N_s$ PDE solves, twice that of the gradient. This section provides a brief review of theory developed in our earlier study Gao et al. (2020); interested readers may refer to the published article for extensive details on the approximation. The Kronecker-based factorization permits fast Hessian-vector products through operations involving (relatively) small matrices. At present, the factorization has only been tested for 2D problems.

As a prelude to the Kronecker-based approximation of the Hessian, we provide some brief conceptual motivation. The Hessian in FWI possesses a block band-diagonal structure (Pratt et al., 1998; Operto et al., 2013). The structure is a consequence of the model discretization and the band-limited nature of seismic waves. An example of this structure is displayed in Figure 5.1 for an explicitly computed Hessian matrix. The reason for pursuing a Kronecker-based factorization of the Hessian becomes more apparent by examining the

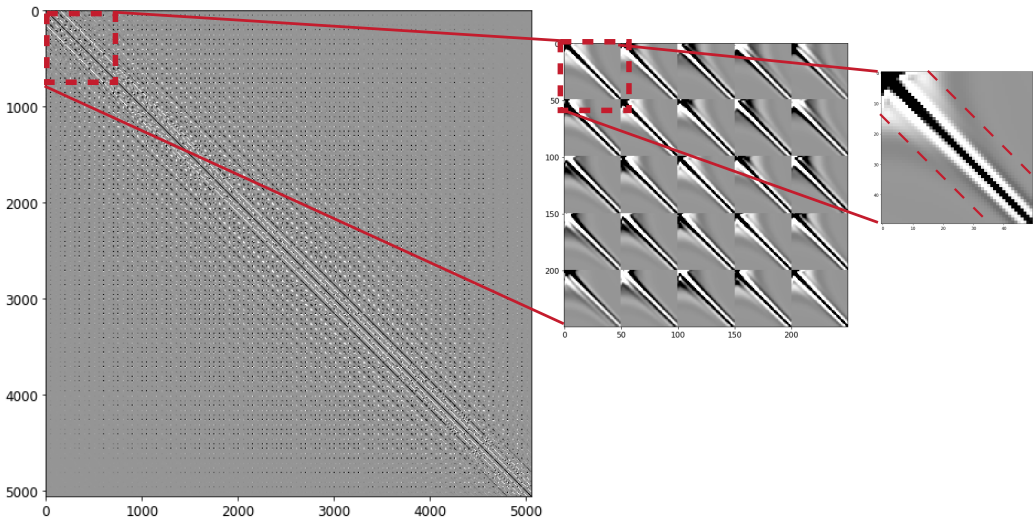

Figure 5.1: Illustration of the block banded-diagonal structure of the Hessian matrix. The Hessian was computed explicitly for a homogeneous acoustic model discretized on a 100×50 finite-difference grid. Zooming into the large matrix reveals its block banded-diagonal structure. Within the matrix, each block has dimensions of $n_z \times n_z$; a total of n_x such blocks exist in the $n_z n_x \times n_z n_x$ Hessian matrix.

definition of a Kronecker product. Consider the 2×2 matrices $\mathbf{A}$ and $\mathbf{B}$ defined as

$$\mathbf{A} = \begin{bmatrix} a_1 & a_3 \\ a_2 & a_4 \end{bmatrix}, \ \mathbf{B} = \begin{bmatrix} b_1 & b_3 \\ b_2 & b_4 \end{bmatrix}. \tag{5.32}$$

The Kronecker product $(\mathbf{A} \otimes \mathbf{B})$, by definition is

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_1 & a_3 \\ a_2 & a_4 \end{bmatrix} \otimes \begin{bmatrix} b_1 & b_3 \\ b_2 & b_4 \end{bmatrix} = \begin{bmatrix} a_1 \begin{pmatrix} b_1 & b_3 \\ b_2 & b_4 \end{pmatrix} & a_3 \begin{pmatrix} b_1 & b_3 \\ b_2 & b_4 \end{pmatrix} \\ a_2 \begin{pmatrix} b_1 & b_3 \\ b_2 & b_4 \end{pmatrix} & a_4 \begin{pmatrix} b_1 & b_3 \\ b_2 & b_4 \end{pmatrix} \end{bmatrix} = \begin{bmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{bmatrix}, \tag{5.33}$$

where $\otimes$ denotes the Kronecker product. The block elements $\mathbf{C}_{ij}$ are 2×2 matrices obtained by multiplying the complete matrix $\mathbf{B}$, by a single element of matrix $\mathbf{A}$. From this simple example, it is evident that the Kronecker product of two matrices yields a matrix with block structure. This property motivates us to pursue an approximation of the Hessian in terms of Kronecker products.

For a discrete 2D grid of size $n_z \times n_x$, the Hessian $\mathbf{H} \in \mathbb{R}^{n_z n_x \times n_z n_x}$ is a large square matrix. The proposed method approximates the Hessian matrix as a superposition of Kronecker products

$$\mathbf{H} \approx \sum_{i=1}^{k} \mathbf{A}_i \otimes \mathbf{B}_i. \tag{5.34}$$

The matrices $\mathbf{A}_i \in \mathbb{R}^{n_x \times n_x}$ and $\mathbf{B}_i \in \mathbb{R}^{n_z \times n_z}$ are referred to as Kronecker factors; k denotes the number of factors used in the approximation. The Hessian in FWI exhibits block-band structure owing to the finite-frequency nature of the seismic wavefield and the discretization of the subsurface (Pratt et al., 1998; Operto et al., 2013).

Assuming the Hessian can be decomposed into a superposition of Kronecker products, the identity

$$(\mathbf{A} \otimes \mathbf{B})\mathbf{m} = \text{vec}(\mathbf{B}\mathbf{M}\mathbf{A}^T), \tag{5.35}$$

is useful. The operator vec denotes the vectorization of a matrix i.e. appending the columns of a matrix into a long vector. The matrix $\mathbf{M}$ is defined such that $\mathbf{m} = \text{vec}(\mathbf{M})$. In 2D, $\mathbf{M}$ is simply the discretized model in physical dimensions. Using equations Equation 5.34 and Equation 5.35, approximate Hessian-vector products can be computed using

$$\mathbf{H}\mathbf{m} \approx (\sum_{i=1}^{k} \mathbf{A}_i \otimes \mathbf{B}_i)\mathbf{m} = \sum_{i}^{k} \text{vec}(\mathbf{B}_i \mathbf{M} \mathbf{A}_i^T). \tag{5.36}$$

Equation 5.36 states that, assuming the Kronecker factors are known, a Hessian-vector

product can be approximated by a superposition of matrix multiplications involving small matrices. The significance of this result is that it allows for very fast computation of Hessian-vector products compared to conventional second-order adjoint-state methods.

Estimating the Kronecker factors is a nuanced process and not the focus of this study. An extensive outline describing the procedure for estimating the Kronecker factors is provided by Gao et al. (2020). To briefly summarize, the Kronecker factors are estimated by solving a low-rank matrix completion problem for a rearranged form of the Hessian. The low-rank completion problem requires samples of the Hessian which are obtained using receiver Green's functions. Computation of the receiver Green's functions marks the most computationally intensive part of the algorithm as it requires N_r PDE solves.

5.4 Numerical experiments

Synthetic inversions are conducted on an acoustic version of the Marmousi II model (Martin et al., 2006). The inversion parameter is the P-wave velocity v_p. A heterogeneous density model is used; however, we do not update density and assume that it is known. The 9.0 x 3.0 km model is discretized on a 900 x 300 regular grid with a spacing of 10 m. An initial v_p model is obtained by smoothing the true model with a Gaussian kernel (σ=200 m). The true and initial v_p models are displayed in Figure 5.2. The synthesized acquisition consists of 22 sources placed at 400 m intervals one grid point beneath the surface ($z_s = 20\ m$). 225 receivers are positioned at the surface at 40 m intervals. To simplify the problem, we assume that the source wavelet is known and model it with a 10 Hz Ricker. A 3 Hz lowcut is applied to the source to remove low frequencies that are typically unavailable in real data. The 'observed' data are pressure component seismograms generated with the true source wavelet and model. Data are generated using an acoustic modelling engine and absorbing boundary layers at each boundary; absorbing boundaries are implemented via convolutional perfectly matched layers (Komatitsch and Martin, 2007). A multi-scale inversion, progressing from low-to-high frequency bands, is used to ensure proper convergence (Bunks et al., 1995). The frequency bands used for this study are 3-5 Hz, -7 Hz and -9 Hz. 30 preconditioned NLCG iterations are used at each inversion stage. All resolution analyses are performed for the final inversion results from the 3-9 Hz inversion.

The prior covariance operator is parametrized using structure tensors computed from a reverse-time migration (RTM) image computed in the initial model (see Appendix A). The RTM image, overlain with representations of computed structure tensors, is displayed in Figure 5.3. To illustrate the effect of the prior covariance operator, we apply it to a random vector in Figure 5.4 and to a raw FWI gradient in Figure 5.5. Figure 5.4a depicts a random vector sampled from a zero-mean, unit-variance normal distribution. Figures 5.4b and c are

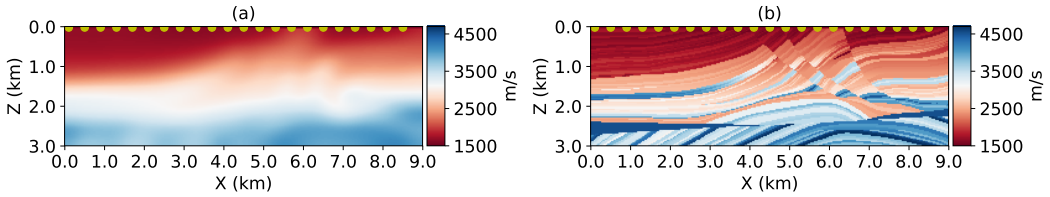

Figure 5.2: Marmousi model. (a) Initial v_p. (b) True v_p. Yellow dots indicate 22 source locations.

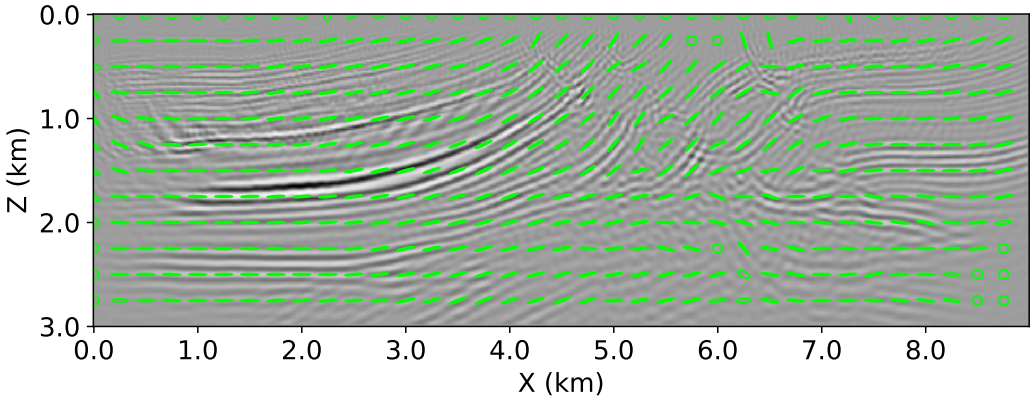

Figure 5.3: RTM image computed using the initial v_p model. Green ellipses represent a subset of structure tensors. In regions where coherent structure is detected, the structure tensors are almost linear. In the absence of structure, the structure tensors appear circular. A linear depth scaling is applied to the RTM image for display purposes. Structure tensors are used to characterize the prior covariance operator.

the action of different prior covariance operators on the random vector. Figures 5.4b and c differ in the characteristic scale length (r_0) of the prior covariance operator. The scale length controls the effective correlation length of the prior covariance operator. Larger scale lengths (Figure 5.4b) produce correlations between a larger neighbourhood of points than short scale lengths (Figure 5.4a). Applied to the gradient (Figure 5.5), the prior covariance operator smooths the gradient in directions of coherent structure. Larger r_0 results in increased smoothing of the gradient. The choice of r_0 is subjective and would ideally be informed by prior knowledge about the subsurface. For this study, we opt for the smaller scale length ($r_0 = 15$) to prevent excessive smoothing to the FWI gradient.

5.4.1 Kronecker factors

Figure 5.6 displays the initial and inverted (or MAP) v_p models for the multi-scale inversion. A good initial model coupled with the multi-scale scheme results in a well recovered model. We observe a slight degradation in the recovery below 2 km depth that is attributed to inadequate illumination.

The horizontal and vertical factors, computed in the MAP model at the 3-9 Hz frequency band, are displayed in Figures 5.7 and 5.8, respectively. The Kronecker factors are diagonally dominant matrices that become increasingly complex with increasing factor number. Figure 5.9 displays the log-diagonal of the Kronecker approximation to the Hessian (reshaped to model dimensions). The Hessian diagonal is often used as a proxy for subsurface illumination (Shin et al., 2001); illumination appears to drop considerably below 1.5 km. Figure 5.10 displays the action of the Kronecker-factored Hessian on a sequence of spike perturbations. In Figure 5.10b, the result has been preconditioned by the inverse of the diagonal Hessian to balance amplitudes at different depths. The Hessian defocuses the spikes to pulses of varying size reflecting the variable resolution in the subsurface. As described in the theory, the Kronecker decomposition approximates the action of the Hessian (on a vector) with two directional operators; this behaviour is demonstrated in Figure 5.11. The horizontal factors $\mathbf{A}_i^T$ act to smear the spikes horizontally whereas the vertical factors $\mathbf{B}_i$ smear input in the vertical direction. By repeating this procedure for multiple factors and summing the output, we approximate the more complex blurring pattern of the Hessian (Figure 5.10). The columns of $\mathbf{A}_i$ and $\mathbf{B}_i$ are point-spread functions (PSFs), which when convolved with some input produce a blurred result. Examples for a spike at $x = 4.5$ km and $z = 1.5$ km are presented in Figure 5.12. Sharp boundaries in the PSFs stem from the band constraints imposed during the estimation of the Kronecker products (Gao et al., 2020). The PSFs for small factor numbers resemble finite-width spikes. As the factor number increases, the PSFs become more oscillatory.

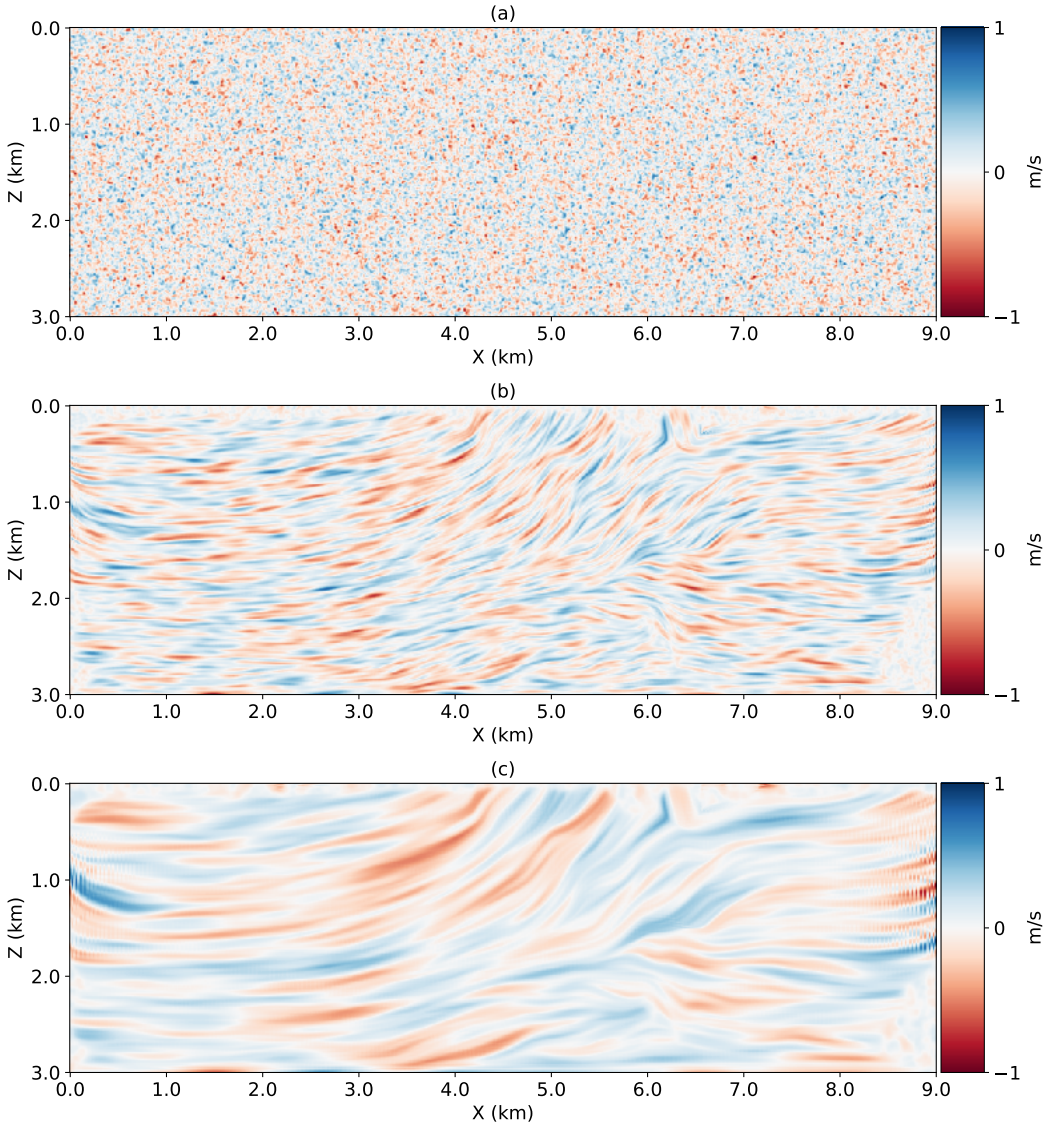

Figure 5.4: Action of the prior covariance operator on a random vector. (a) Random noise vector. (b) Prior covariance operator (short-scale length, $r_0 = 15$) applied to a random noise vector. (c) Prior covariance operator (intermediate-scale length, $r_0 = 50$) applied to a random noise vector.

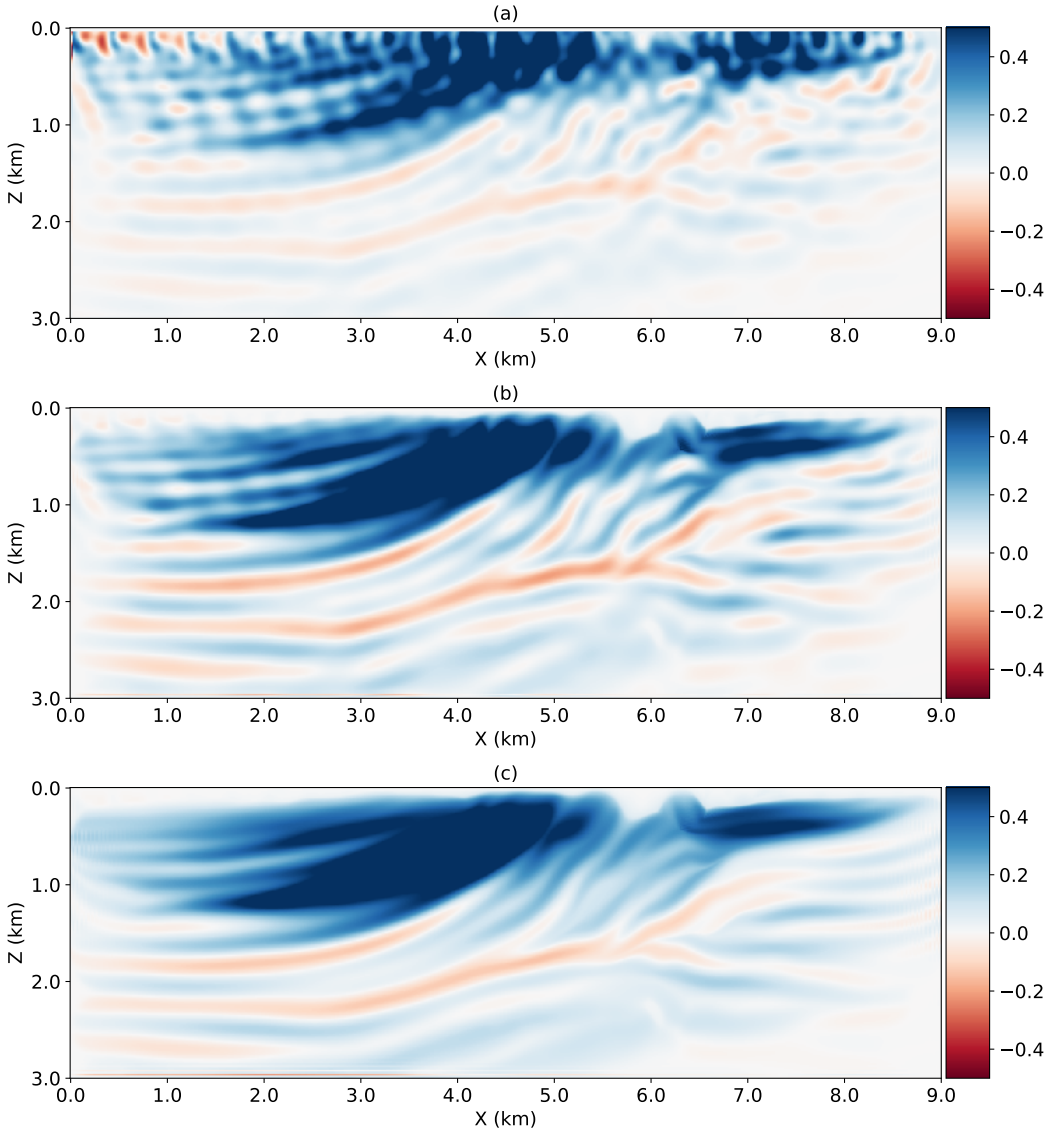

Figure 5.5: Action of the prior covariance operator as a preconditioner to the FWI gradient. (a) Initial v_p gradient in Marmousi model at 3-5 Hz. (b) Preconditioned gradient (short-scale length, $r_0 = 15$). (c) Preconditioned gradient (intermediate-scale length, $r_0 = 50$). The covariance operator has smoothed the gradient along the coherent directions in the seismic image.

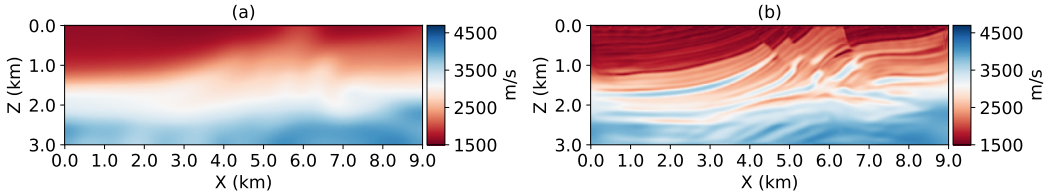

Figure 5.6: Marmousi v_p inversion results (a) Prior mean, the initial model. (b) MAP model (posterior mean), the inverted model.

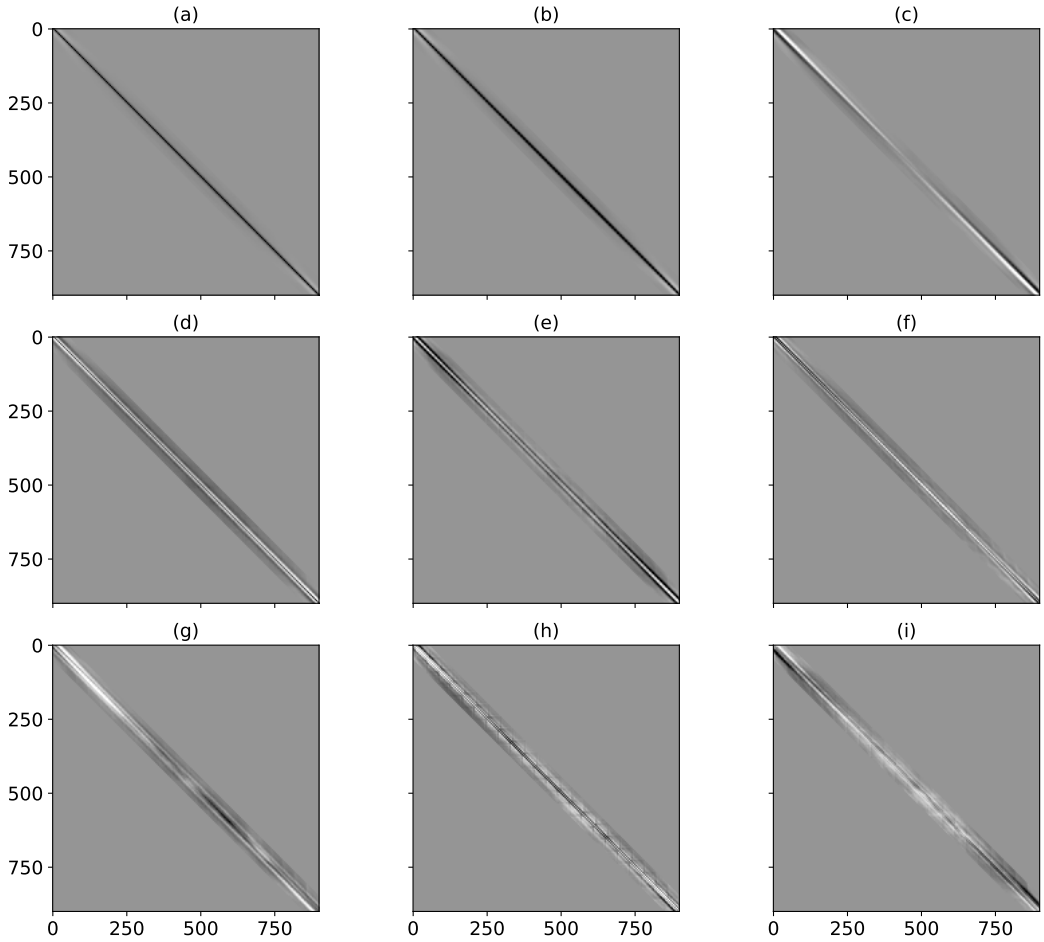

Figure 5.7: Estimated 'horizontal' Kronecker factors ($\mathbf{A}_i \in \mathbb{R}^{900\times900}$) used to approximate the Hessian computed at the MAP model. (a-f) The 9 factors arranged from largest to smallest; all factor matrices exhibit banded diagonal structure. The complexity of the diagonal structure generally increases with factor number.

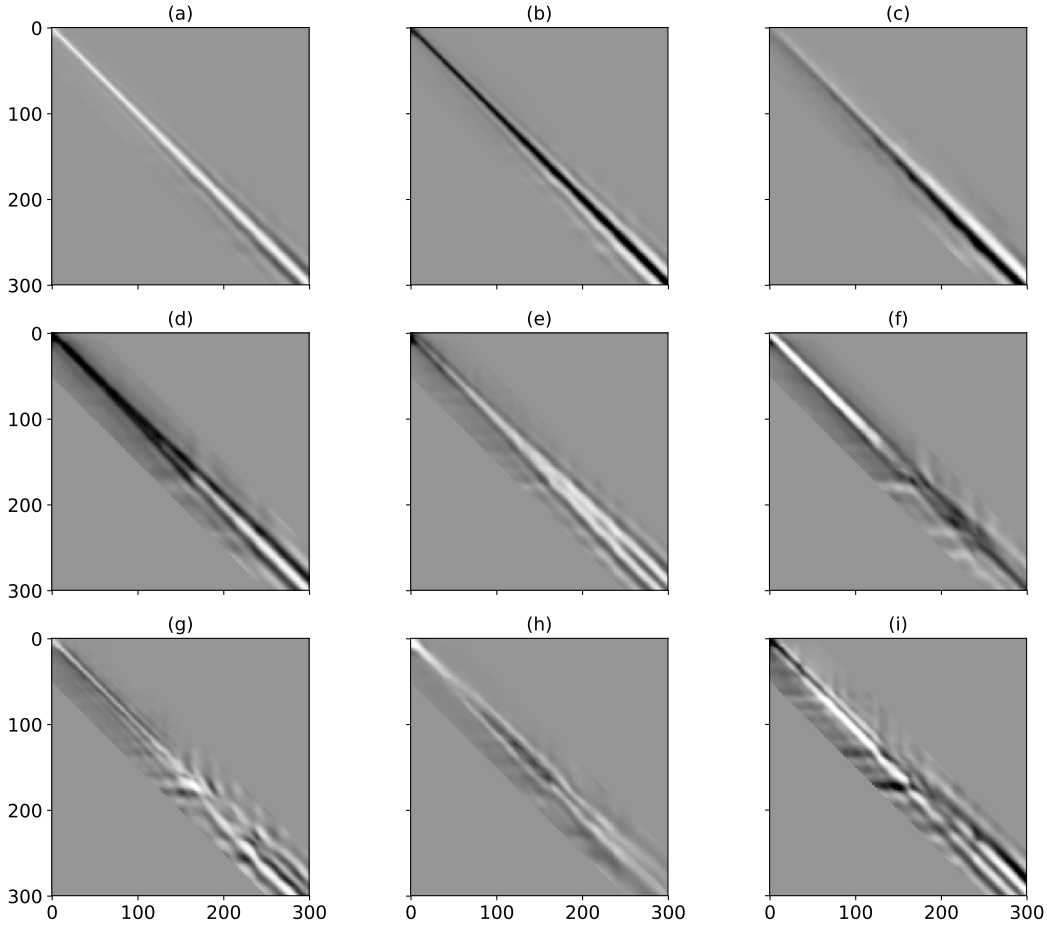

Figure 5.8: Estimated 'vertical' Kronecker factors ($\mathbf{B}_i \in \mathbb{R}^{300\times300}$) used to approximate the Hessian computed at the MAP model. (a-f) The 9 factors are arranged from largest to smallest; all factor matrices exhibit banded diagonal structure.

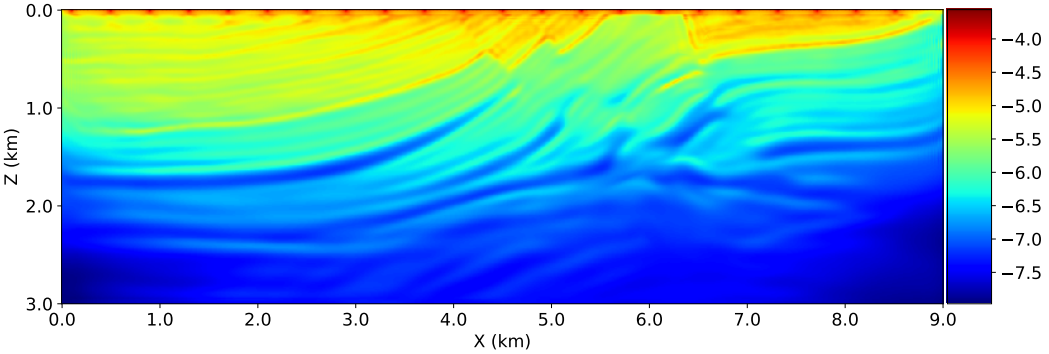

Figure 5.9: Log-diagonal of the approximated Hessian (reshaped to model dimensions). The inverse of the Hessian diagonal is a useful preconditioner as it balances amplitudes in regions where illumination is otherwise inadequate (e.g., in deeper regions).

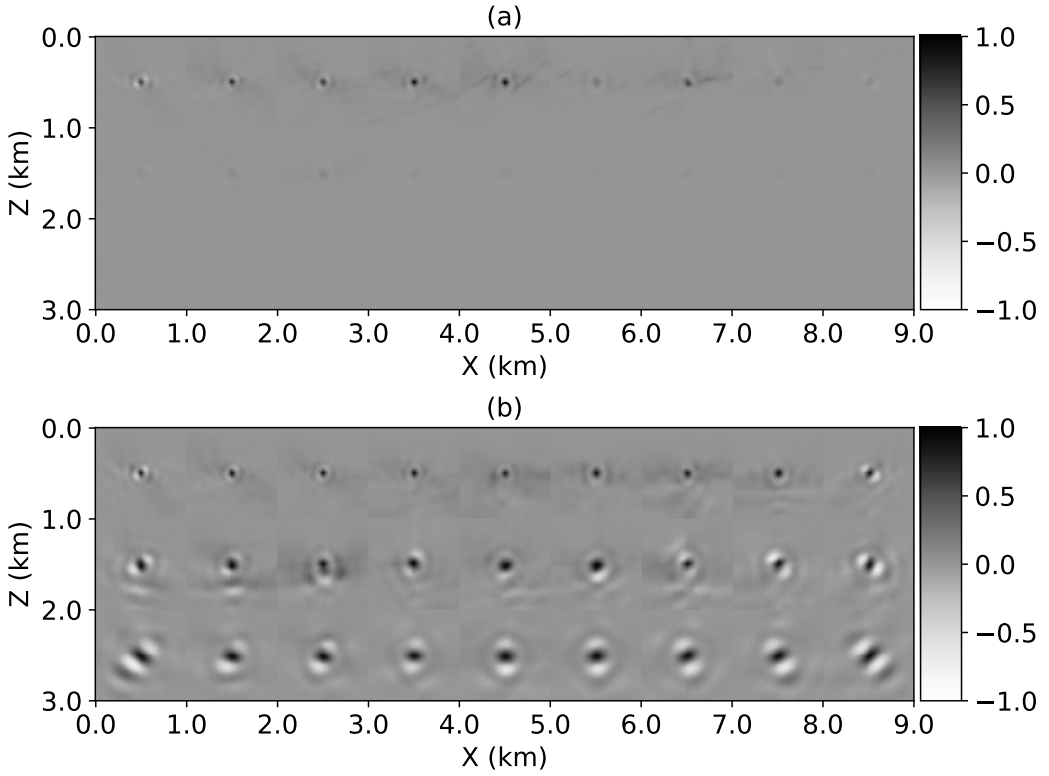

Figure 5.10: Application of the Hessian to an array of spike perturbations. The spike array is composed of unit perturbations at 1 km $\times$ 1 km intervals. The results are presented (a) without and (b) with inverse-diagonal Hessian preconditioning. The smearing effect of the Hessian is apparent. In deeper regions, spikes are less focused and occasionally appear with prominent side lobes.

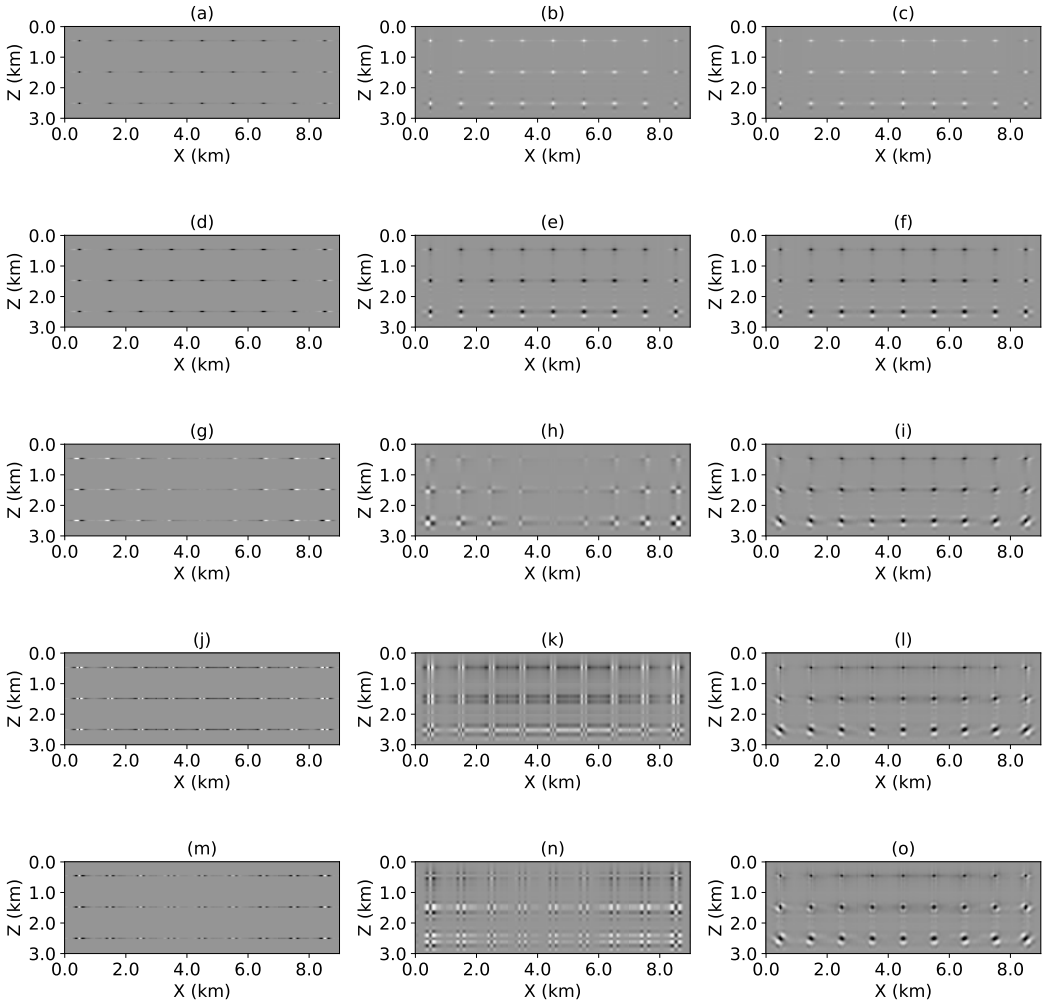

Figure 5.11: Various stages of the Kronecker approximation of $\mathbf{H}\delta\mathbf{m}$ for the 5 largest Kronecker factors: (a-c) $k = 1$, (d-f) $k = 2$, (g-i) $k = 3$, (j-l) $k = 4$, (m-o) $k = 5$. The input perturbation $\delta\mathbf{m} = \text{vec}(\delta\mathbf{M})$, is a spike array with 1 km $\times$ 1 km spacing. For each k, the columns display (left) $\delta\mathbf{M}\mathbf{A}_k^T$, (middle) $\mathbf{B}_k\delta\mathbf{M}\mathbf{A}_k^T$ and (right) $\sum_{i=1}^{i=k}\mathbf{B}_i\delta\mathbf{M}\mathbf{A}_i^T$. The horizontal factors $\mathbf{A}_i$ smear the perturbations horizontally, whereas the vertical factors $\mathbf{B}_i$ smear perturbations vertically. The superposition (right column), gradually improves the approximation of the Hessian-vector product, adding more nuanced details to the image as more factors are included in the approximation.

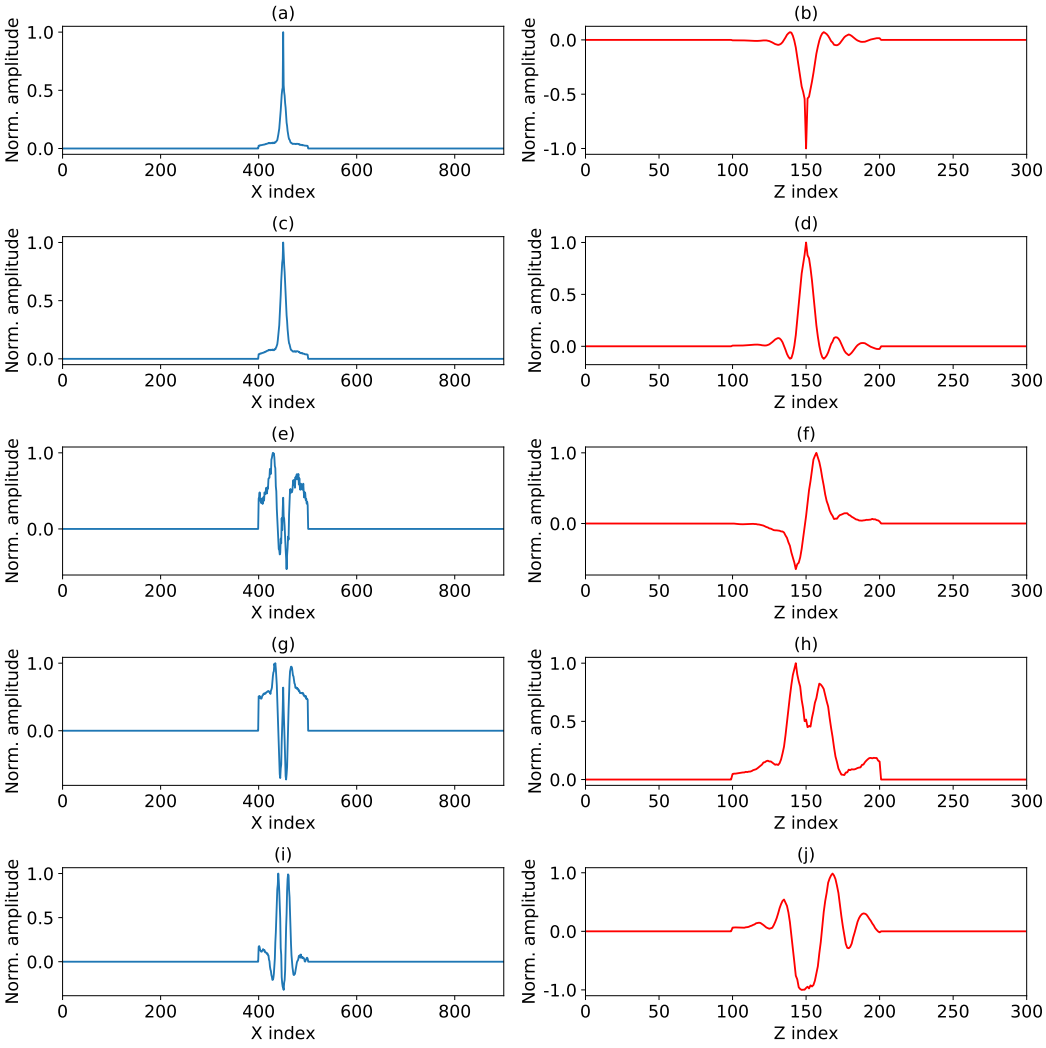

Figure 5.12: Point spread functions extracted from the 5 largest Kronecker factors: (a, b) $k = 1$, (c, d) $k = 2$, (e, f) $k = 3$, (g, h) $k = 4$, (i, j) $k = 5$. (left column) 450th column from horizontal factors $\mathbf{A}_i$. (right column) 150th column from vertical factors $\mathbf{B}_i$.

5.4.2 Local resolution analysis

Applying the Hessian to a unit-spike perturbation equates to sampling a column of the Hessian. Similar to the Kronecker factors, the columns of the Hessian can be considered as PSFs (Fichtner and Leeuwen, 2015). To extract resolution information, we pick vertical and horizontal widths of Hessian PSFs. Examples demonstrating the procedure are depicted in Figures 5.13-5.15. In Figure 5.13, slices through the PSF illustrate how the resolution lengths are selected (black bars slice plots). To expedite the process, only 8 Kronecker factors are used to compute approximate PSFs; 8 was the minimum number of factors that produced PSFs with similar characteristics to the true PSFs. For a complete characterization of the subsurface, we compute resolution lengths at 100 m intervals (every 10 grid points). The resultant horizontal and vertical resolution length maps are interpolated to fill gaps; the maps are displayed in Figure 5.16. A gradual deterioration of resolution is apparent from increasing vertical and horizontal resolution lengths with depth. In addition, a comparison of resolution lengths with the v_p structure suggests a potential correlation between increased resolution lengths and more complex subsurface structure. Towards boundaries of the model, PSFs become less circular and resemble ellipses with some orientation (e.g., Figure 5.15). In such cases, resolution lengths as they are measured in this study may not be entirely representative of resolution. For the final stage of our local analysis, we generate a bank of non-stationary, 1D Gaussian filters parametrized by the measured resolution lengths (appear as dashed blue lines in Figures 5.13-5.15). We perform a 1D non-stationary convolution of the true v_p model with the Gaussian filters. The reasoning for this test is as follows, if the measured resolution lengths are indicative of the smallest resolvable features in the inverted model, then the true model filtered to these scale lengths should resemble the inversion result. Given the simplicity of this interpretation, this test serves as more of a quality control measure. Ideally, this test would be performed using a 2D non-stationary convolution with normalized Hessian PSFs, but we forego that here. A more comprehensive analysis would integrate the local resolution analysis into the framework of homogenization theory. Homogenized models a smooth version of the true Earth model that is "visible" to the finite-frequency data (Capdeville and Mtivier, 2018). Vertical and horizontal velocity profiles are presented in Figures 5.17 and 5.18, respectively. The filtered profiles demonstrate a reasonable correspondence, in terms of the resolved features, with the inverted model. The filtered profiles do not account for variable subsurface illumination; therefore, better represent the true amplitudes of perturbations. The horizontal profile (Figure 5.18) exhibits instances where the filtered profile resolves smaller scale features than the inversion. This may suggest that the horizontal resolution lengths are underestimated in these areas.

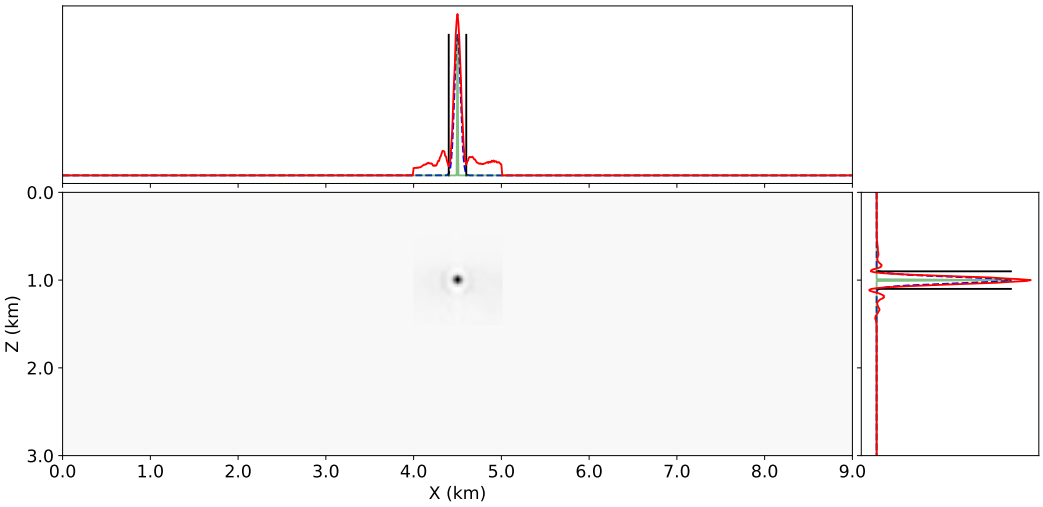

Figure 5.13: Kronecker-factored Hessian applied to a spike perturbation $\delta\mathbf{m}$ at $x = 4.5$ km, $z = 1.0$ km. The panels above and to the right display horizontal and vertical slices through the image, respectively. The green line is the spike perturbation. The black bars mark the picked resolution length at this point. The dashed blue line is Gaussian parametrized with the selected horizontal and vertical resolution lengths.

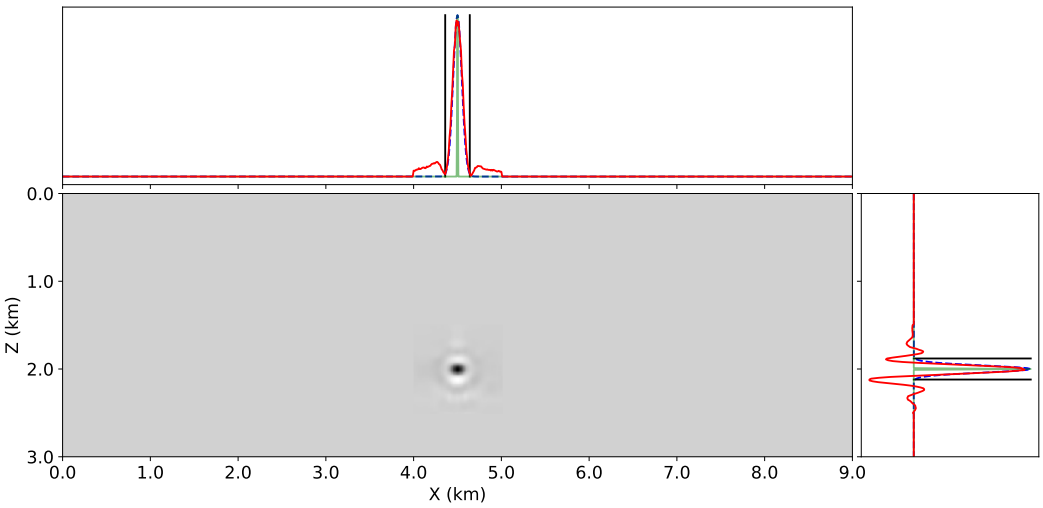

Figure 5.14: Same as Figure 5.13 for a spike perturbation $\delta\mathbf{m}$ at $x = 4.5$ km, $z = 2.0$ km. The vertical and horizontal resolution lengths are notably larger than in Figure 5.13

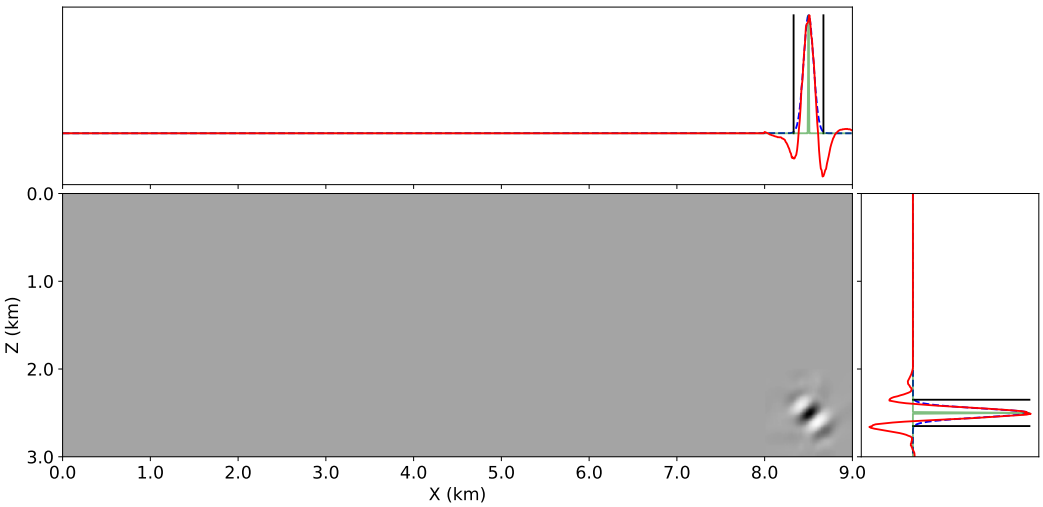

Figure 5.15: Same as Figure 5.13 for a spike perturbation $\delta\mathbf{m}$ at $x = 8.5$ km, $z = 2.5$ km. Towards the limits of the model, unbalanced illumination results in less-circular spikes with preferred orientations.

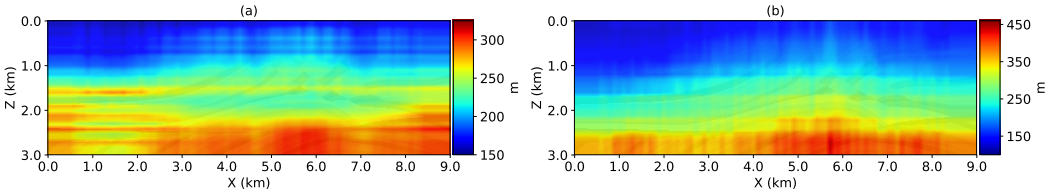

Figure 5.16: Interpolated (a) vertical and (b) horizontal resolution lengths. The resolution lengths were picked from Hessian PSFs computed at 100 m intervals; gaps were filled via interpolation. A greyscale v_p model overlay is included to demonstrate correlations between structure and resolution length.

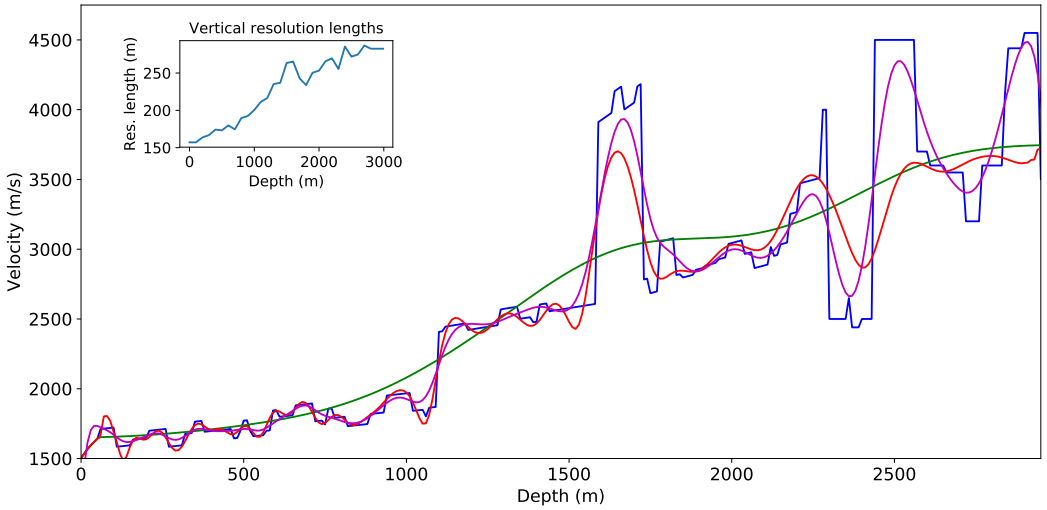

Figure 5.17: Vertical pseudo well log at $x = 2.0$ km. Comparison of the true (blue line), initial (green line), inverted (red line) and smoothed true model (magenta line). The smoothed log is obtained by performing a 1D non-stationary convolution of the velocity profile with a bank of non-stationary Gaussian filters whose widths are parametrized by the resolution lengths. Inset displays the vertical resolution lengths used to design the non-stationary Gaussian filters.

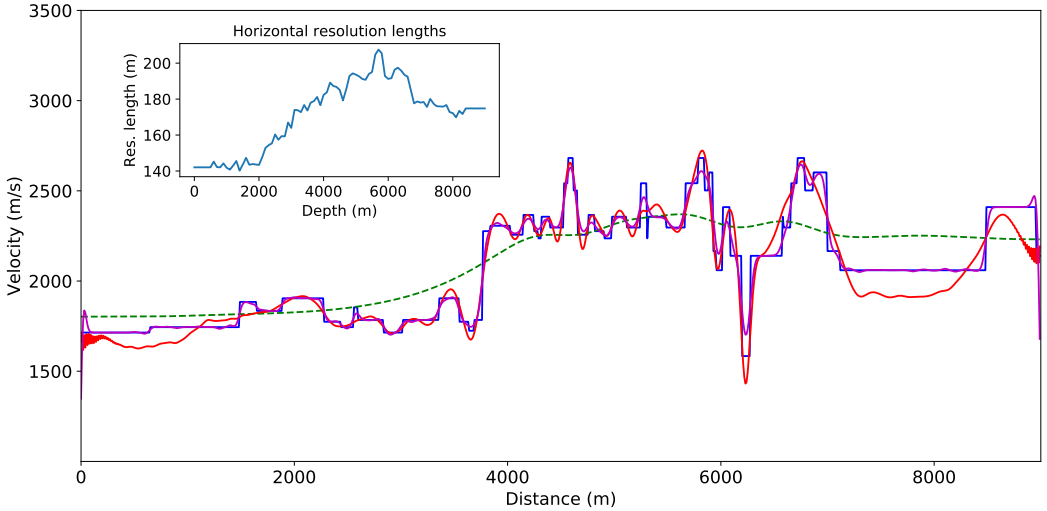

Figure 5.18: Similar to Figure 5.17 but for a horizontal log at $z = 0.75$ km.

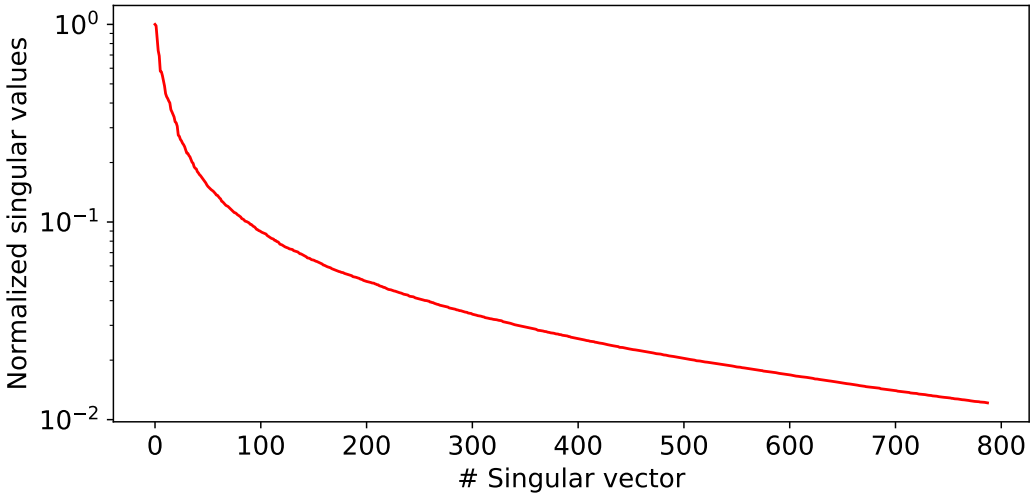

Figure 5.19: Normalized singular values of prior-preconditioned Hessian.

5.4.3 Linearized Bayesian inversion

Figure 5.19 displays the 800 largest singular values of the prior-preconditioned Hessian computed from Lanczos iterations. Each Lanczos iteration is accelerated by the use of the Kronecker-based Hessian approximation (which uses 100 Kronecker factors). The singular values, normalized by the largest singular value, fall below 0.1 after approximately 100 singular vectors. The 12 largest singular vectors of the prior-preconditioned Hessian appear in Figure 5.20. The largest singular vectors are smooth directions concentrated towards the surface. For smaller singular values, the corresponding singular vectors become increasingly oscillatory with shorter wavelength patterns that extend deeper into the subsurface. Figure 5.21 compares the action of the prior-preconditioned Hessian on a vector, with it's low rank approximation (r=800). Qualitatively, the low rank approximation yields similar output although some localized differences are evident (Figure 5.21d).

Figure 5.22 displays the prior and posterior means and standard deviations computed from 500 random samples drawn from each distribution. By construction, the prior and posterior means are the initial and inverted models, respectively. The standard deviation of the prior covariance shows limited spatial variations. By contrast, the standard deviation of the posterior covariance exhibits lower values for $z < 1.5$ km where illumination is greatest. The patches of low deviation, apparent with both the prior and posterior covariances, are a consequence of the diffusion-tensor determinant related scaling in the prior. Regions of the RTM image lacking structure were downweighted and essentially neglected by the covariance operators. Below 1.5 km depth, the STD of the posterior covariance appears

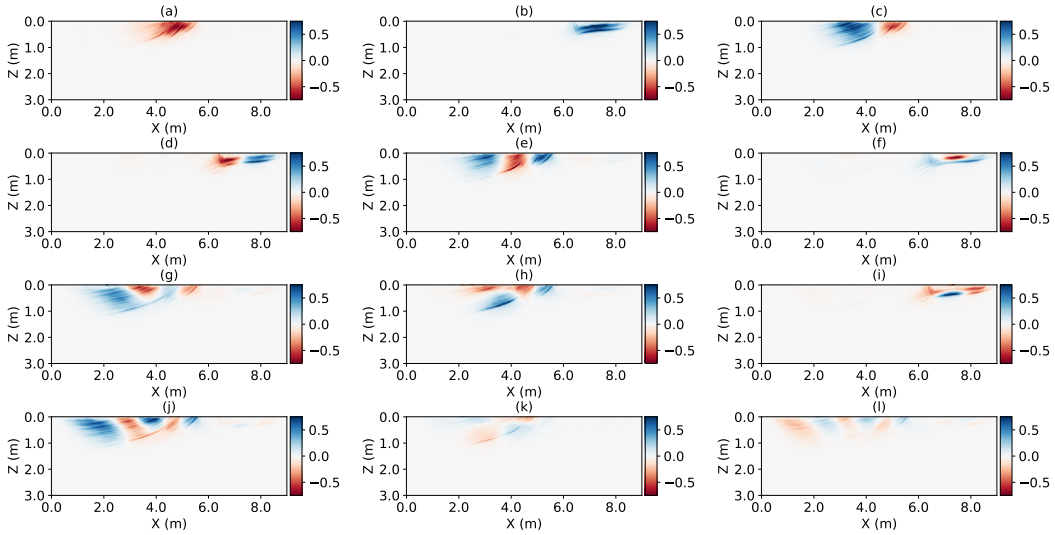

Figure 5.20: Estimated singular vectors of prior-preconditioned Hessian. (a-l) 12 largest estimated singular vectors ordered left-to-right, top-to-bottom.

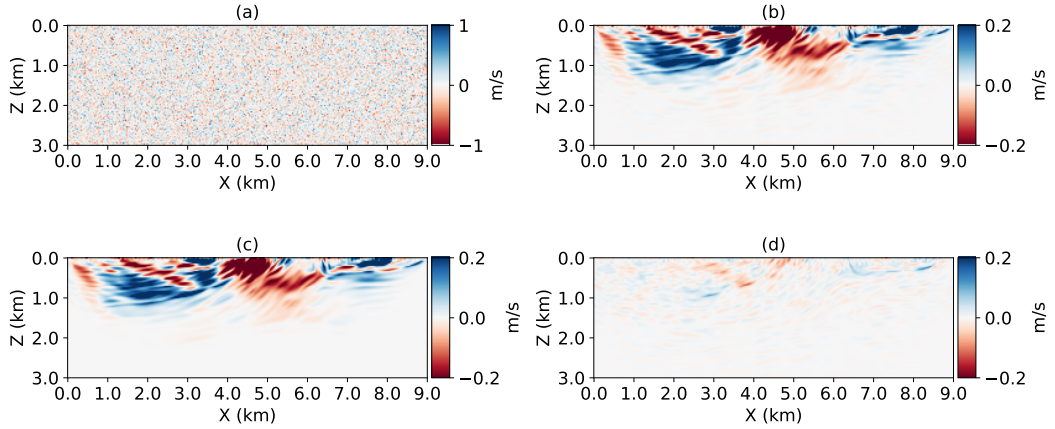

Figure 5.21: Comparison of the prior-preconditioned Hessian and its low-rank approximation. (a) Random noise vector. (b) Action of prior-preconditioned Hessian applied on random vector. (c) Action of low-rank approximation of the prior-preconditioned Hessian on a random vector. (d) Difference between true product (b) and low-rank approximation (a).

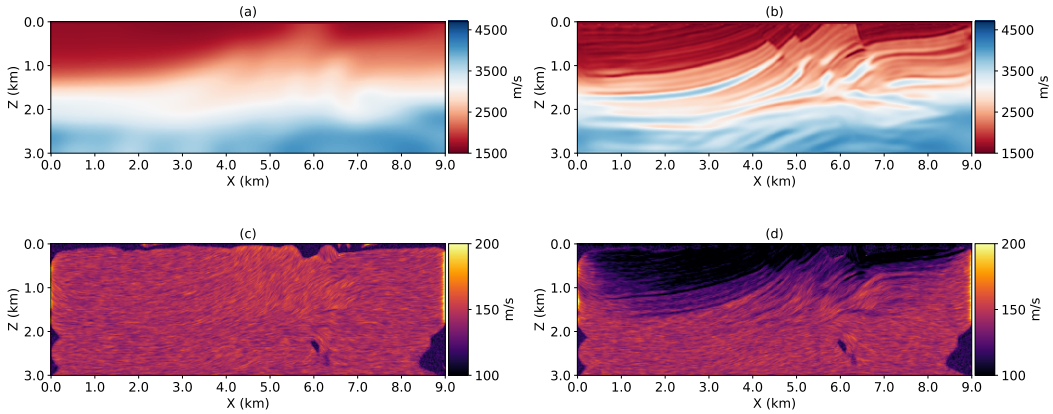

Figure 5.22: Prior and posterior mean and standard deviations for Marmousi inversion. (a) Prior mean, the initial model. (b) Posterior mean, the inverted model. (c) Prior standard deviation (diagonal of prior covariance). (d) Posterior standard deviation (diagonal of posterior covariance). The introduction of the data constraints reduces the uncertainty in the shallow regions where illumination is greatest. Due to a lack of illumination, deeper regions are not better constrained than by the prior distribution.

largely unchanged suggesting that the data has not helped to further constrain this region of the subsurface. There are further nuances to this statement that we address in the discussion. The difference between the standard deviation of prior and posterior covariances is displayed in Figure 5.23. Figure 5.24 presents random samples drawn from the prior and posterior distributions. Figures 5.25 and 5.26 display depth profiles of velocity before and after inversion. Figure 5.25 is an example where the standard deviation of the posterior covariance is reduced relative to that of the prior covariance. This results in narrower confidence intervals for $z < 1.5$ km, indicating better constrained subsurface structure. For $z >$1.5 km, the confidence intervals have similar widths for both the prior and posterior covariances. Figures 5.27 and 5.28 display horizontal velocity profiles taken at different depths. Again, in the shallow areas the posterior exhibits tighter confidence intervals than the prior covariance. At depth, the uncertainties remain similar.

5.5 Discussion

Our development relies on an informative initial subsurface image to characterize the prior covariance operator. If this is not available, for example, due to poor data quality or a poor initial velocity model, a more conservative prior can be selected. Conservative priors may include isotropic covariance operators or operators that capture depth-dependent velocity trends. In our example, the small scale length of the prior covariance operator results in

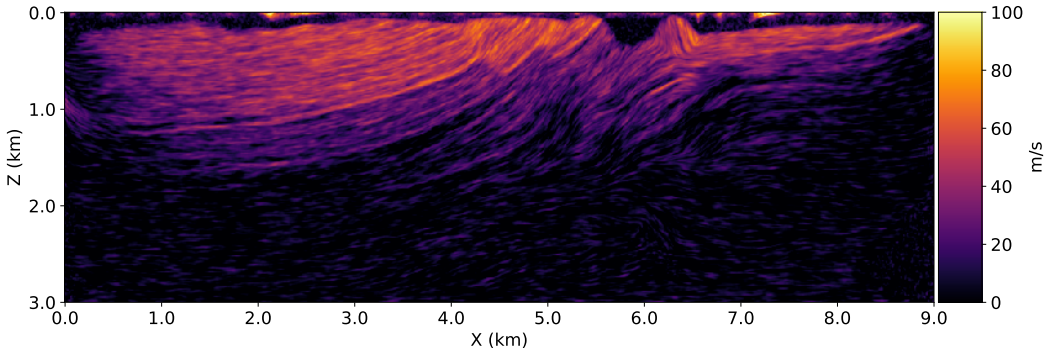

Figure 5.23: Difference between the prior and posterior standard deviations. Similar to Figures 5.22c, d. The standard deviation of the posterior distribution is primarily reduced in the shallower regions ($z < 1.5$ km) where data illumination is largest.

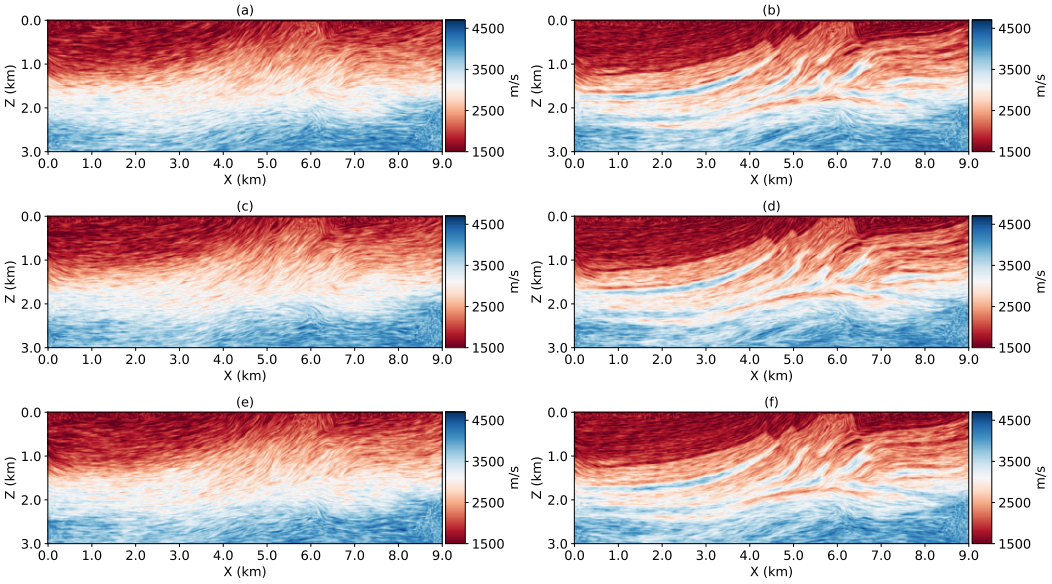

Figure 5.24: Random samples from the (a, c, f) prior and (b, d, f) posterior distribution.

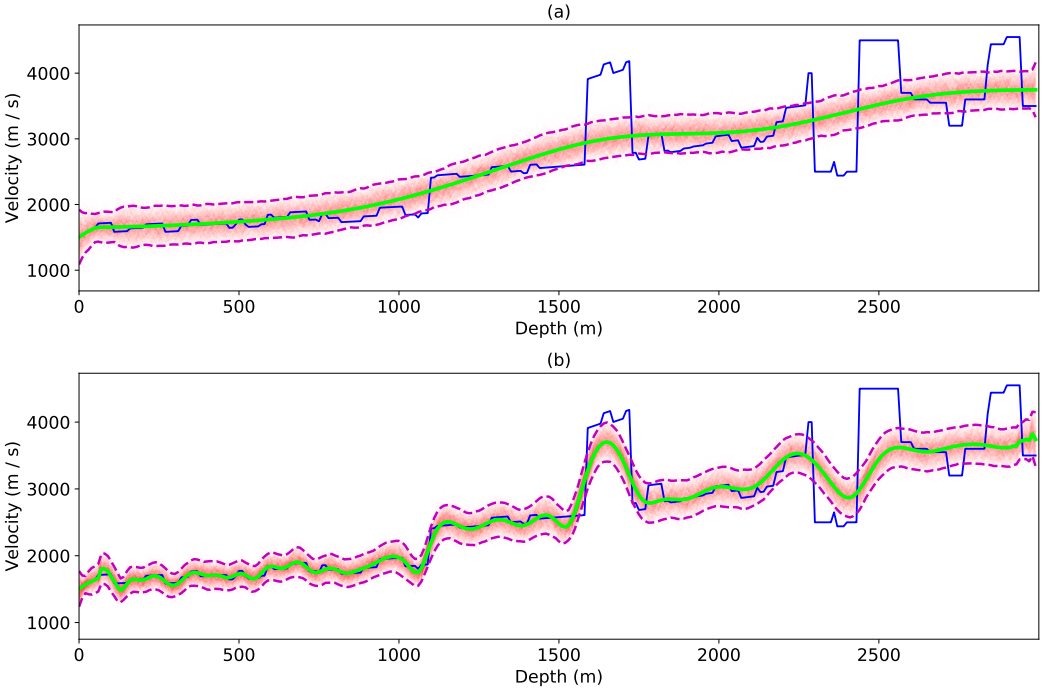

Figure 5.25: Vertical pseudo well log at $x = 2.0$ km. (a) Prior distribution. (b) Posterior distribution. The prior and posterior means (green line) are bounded by the 95% confidence intervals (dashed magenta lines). 500 random samples, drawn from the prior and posterior distributions, are plotted on top of one another (red lines). The true model is displayed as a blue line.

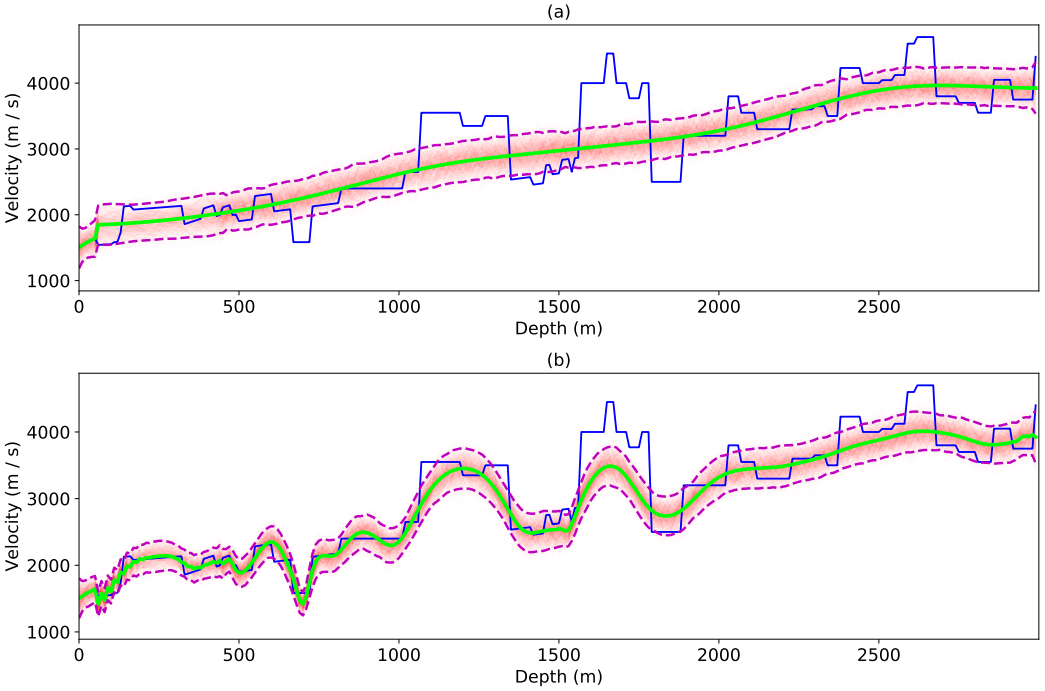

Figure 5.26: Vertical pseudo well log at $x = 6.35$ km. (a) Prior distribution. (b) Posterior distribution. The prior and posterior means (green line) are bounded by the 95% confidence intervals (dashed magenta lines). 500 random samples, drawn from the prior and posterior distributions, are plotted on top of one another (red lines). The true model is displayed as a blue line.

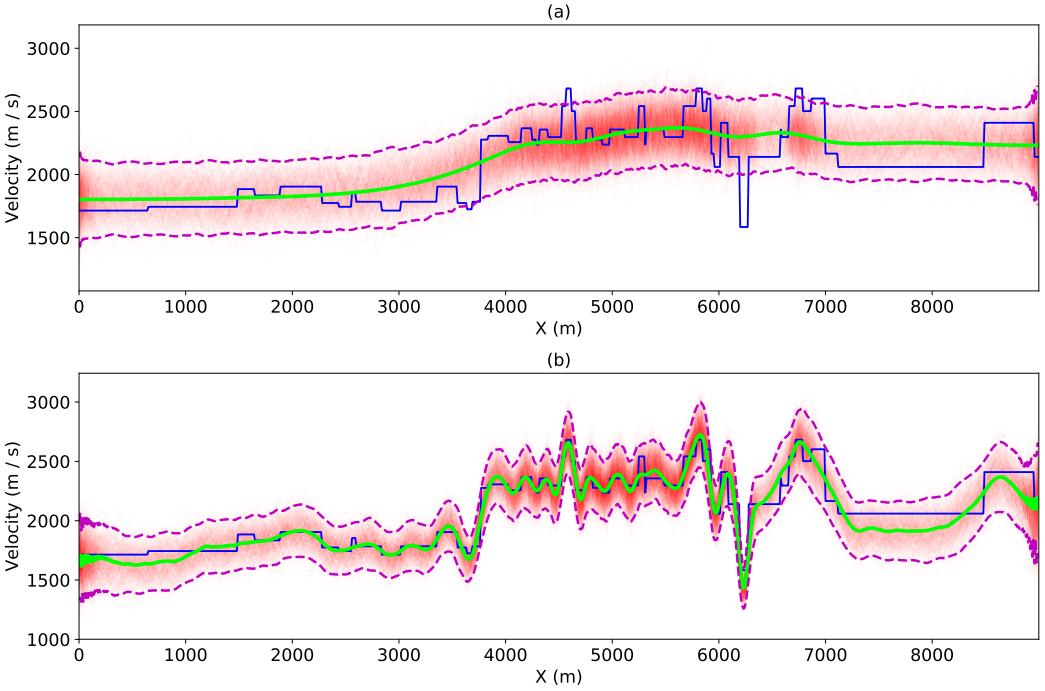

Figure 5.27: Horizontal pseudo well log at $z = 0.75$ km. (a) Prior distribution. (b) Posterior distribution. The prior and posterior means (green line) are bounded by the 95% confidence intervals (dashed magenta lines). 500 random samples, drawn from the prior and posterior distributions, are plotted on top of one another (red lines). The true model is displayed as a blue line.

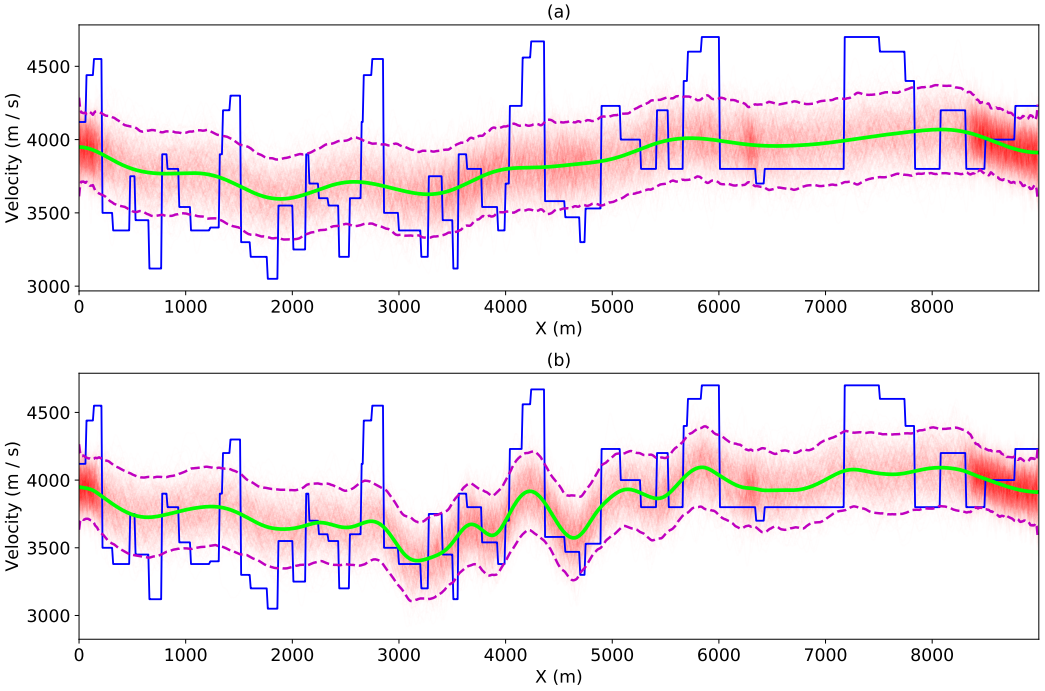

Figure 5.28: Horizontal pseudo well log at $z = 2.5$ km. (a) Prior distribution. (b) Posterior distribution. The prior and posterior means (green line) are bounded by the 95% confidence intervals (dashed magenta lines). 500 random samples, drawn from the prior and posterior distributions, are plotted on top of one another (red lines). The true model is displayed as a blue line.

prior/posterior samples with potentially unrealistic small scale perturbations. For more realistic applications, r_0 should be informed by prior geological knowledge.

The posterior STD suggests that the data constraints provide a minimal reduction in the uncertainty at depths below 1.5 km. However, if we examine the MAP model we observe weakly resolved but well defined structure throughout the complete depth extent of the model. A potential explanation for this comes from the use of a multi-scale inversion approach. Lower frequency data produce gradients with larger Fresnel zones, thus produce deeper updates than higher frequency data. Deeper structure is inserted during the lower frequency updates. The uncertainty analysis utilizes only the 3-9 Hz data making it somewhat incomplete. How to properly integrate a multi-scale approach into a Bayesian inversion warrants further consideration but is beyond the scope of this study. One of the target hydrocarbon reservoirs in the Marmousi model appears at approximately $x = 6.5$ km and $z = 2.5$ km. At this depth, it appears that the data do not help to further constrain the velocity structure (when compared to the prior); however, this example is also representative of the issue with multi-scale uncertainty estimation. Assuming that our data do not help to further constrain the velocity model, it indicates potential deficiencies in the acquisition where the target area is not being adequately illuminated. Such information could be potentially useful when considering improved survey design. Imaging at depth could be improved by wider offset acquisition or improved low frequency information in the data. Shallower gas traps in the Marmousi model are located in regions where the data constraints help to reduce the posterior STD (relative to the prior STD). The reduction in uncertainty could be useful to interpreters seeking to identify hydrocarbon reserves through quantitative interpretation of inverted models.

For simplicity, our example was noise free and assumed that the data covariances were uncorrelated. Because the inverse of the data covariance is embedded between the Fréchet derivative matrices/operators (Equation 5.13), accommodation of a data covariance operator would have to be done prior to the estimation of the Kronecker factors. Specifically, its inclusion would feature during the computation of Hessian samples required to estimate the Kronecker factors.

Extensions of the Kronecker-based factorization of the Hessian are possible for the multi-parameter Hessian. The bulk of the computational cost still lies in computing samples of the Hessian from receiver Green's functions. Since the multi-parameter Hessian is a block matrix, it may be possible to decompose the matrix into block elements and estimate Kronecker factors for each block element individually (as opposed to for the full matrix). Further investigation is required to assess the merits of each approach. A multi-parameter Bayesian inversion would follow a similar formulation as used in this study. Additional consideration would need to be given to the design of a suitable model prior. A multi-parameter prior

model covariance operator should consider both correlations between different subsurface points as well as correlations between independent physical parameters. The latter requires a deeper understanding of rock physics in the survey area. Li et al. (2016) use a stochastic rock physics modelling approach to design multi-parameter prior covariance operators.

5.6 Conclusions

We have presented two forms of resolution analysis for FWI. The first performs a local resolution analysis assuming convergence to a model in the vicinity of the global minima. This approach extracts horizontal and vertical resolution lengths by examining the action of the Hessian, a blurring operator, on spike perturbations at various points in the subsurface. The validity of the resolution lengths is further investigated by performing non-stationary convolution of the true model with Gaussian filters parametrized by the measured resolution lengths. The second approach to resolution analysis formulates FWI as a Bayesian inverse problem following a linearization of the modelled data about the MAP model. Gaussian priors are assumed for the data and model. The prior model covariance operator is designed using an anisotropic shaping covariance that is parametrized by coherent structures in an image of the subsurface. We employ a low-rank approximation of the prior-preconditioned Hessian to facilitate sampling of the posterior distribution. Random samples drawn from the prior and posterior distributions are used to compute standard deviations for the subsurface model. A numerical test is performed on the Marmousi model assuming the acoustic approximation. A multi-scale inversion is performed up to a frequency band of 3-9 Hz. We approximate the Gauss-Newton Hessian as a superposition of Kronecker products, which themselves are relatively small matrices. The Kronecker factorization permits an accurate approximation combined with a compact representation of the Hessian. Fast Hessian-vector products, required to probe the subsurface and for Lanczos iterations, can be computed via matrix multiplications involving small matrices. The local resolution analysis suggests that resolution lengths i.e. smallest resolvable features, increase with depth. This is consistent with the fact that illumination at deeper parts of the model are more limited. Similar effects are observed in the STD of the posterior covariance. With the inclusion of data constraints, the posterior distribution exhibits smaller deviations below 1.5 km depth where illumination is greatest. During this process we identify some shortcomings of the current implementation. The prior covariance could be better designed. Furthermore, the Hessian was computed at the highest frequency band used for inversion. As a result, the results suggest limited data-constraints at depths where the inverted model appear to be reasonably resolved. We interpret this as relating to the fact that higher bandwidth data have lower penetration depths.

CHAPTER 6

Acoustic and elastic FWI in the western Canadian sedimentary basin[1]

6.1 Introduction

Full waveform inversion (FWI) has matured considerably since its conception in the 1980's (Lailly, 1983; Tarantola, 1984b). Advances in computational hardware, data acquisition, and algorithmic developments have resulted in FWI being integrated into standard velocity model building workflows in exploration seismology. The majority of case studies explore marine streamer or ocean-bottom node datasets (e.g., Shipp and Singh (2002); Ravaut et al. (2004); Sears et al. (2008); Warner et al. (2013); Prieux et al. (2013a,b)). While relatively robust workflows exists for FWI in marine settings, the same is less true for land data. Performing FWI on land is considerably more challenging due to, for example, degraded data quality, strong elastic effects, and surface topography (Stopin et al., 2014). Comparatively poor signal-to-noise ratios (SNR) can compromise low-frequency information in the data that is vital for FWI. Furthermore, ground roll generated by the elastic free surface obscures reflected arrivals at near offsets in the data. In general, applications of land FWI require more design to ensure successful inversions. The adoption of elastic FWI has been precluded by its high computational cost meaning that the vast majority of FWI studies have been performed using the acoustic or viscoacoustic approximation. Improved understanding of the FWI algorithm has led to some initial studies being conducted for elastic (Vigh et al., 2014; Raknes et al., 2015) and land (Plessix et al., 2012; Stopin et al., 2014; Vigh et al., 2018; Sedova et al., 2019; Solano and Plessix, 2019; Trinh et al., 2019) applications of FWI.

[1] A version of this chapter is being prepared for a manuscript submission to *Geophysics*.

In this chapter, we apply acoustic and elastic FWI to a 2D land dataset from the western Canadian sedimentary basin (WCSB). The study aims to establish the feasibility and utility of FWI for land data in this region. We seek to develop FWI workflows that may be readily transferred to similar datasets in the region. Finally, we also explore the limitations of the dataset, the differences between acoustic and elastic inversion, and a strategy to incorporate reflection data into the elastic inversion. The theory used in this chapter follows that established in Chapter 1. The first section details the geological setting and the Cynthia dataset. The second and third sections describe the data preprocessing and initial model building steps that precede FWI. The fourth section outlines the generic elements of the FWI workflow utilized for acoustic and elastic inversion. The subsequent three sections present results for the acoustic, elastic and reflection-based elastic inversions. The final section focuses on validation of the FWI results.

6.2 Geological background and dataset

6.2.1 Western Canadian sedimentary basin

The study region resides in a conventional oil field in the western Canadian sedimentary basin (WCSB). The geology of the WCSB has been studied extensively owing to an abundance of outcrops that enable identification of various geological strata. Fundamentally, the WCSB is a wedge of sedimentary strata deposited on the stable North American craton (Mossop and Shetsen, 1994). The basin can be separated into two parts: a predominantly carbonate base formed between the Paleozoic and Jurassic period and a foreland basin formed primarily during the Cretaceous period. The Cretaceous foreland basin consists predominantly of marine shales interlaced with thin sandstone layers. The sandstone layers within these sequences are of great economic importance as they host the majority of oil and gas reservoirs in the region. More recent formations are composed of mostly sand and gravel. The uppermost layers are primarily glacial till and unconsolidated sediments (Mossop and Shetsen, 1994). Notable shale formations identified in the survey area include the Lea park, Colorado, and Fernie groups. The major sandstone formations identified are the Belly river, Cardium, Viking, and Mannville groups; the Cardium and Viking formations are prominent oil reserves. The transition to the carbonate base is marked by the Nordegg formation. The tops of select formations are marked in Figure 6.1.

Figure 6.1 displays a schematic interpretation of the subsurface geology (overlain on a PSTM image). The various colour overlays indicate the dominant rock types associated with major geological formations identified in well logs from the survey area. The depths of the first two layers are not accurate as information was not available for these layers. The near-surface

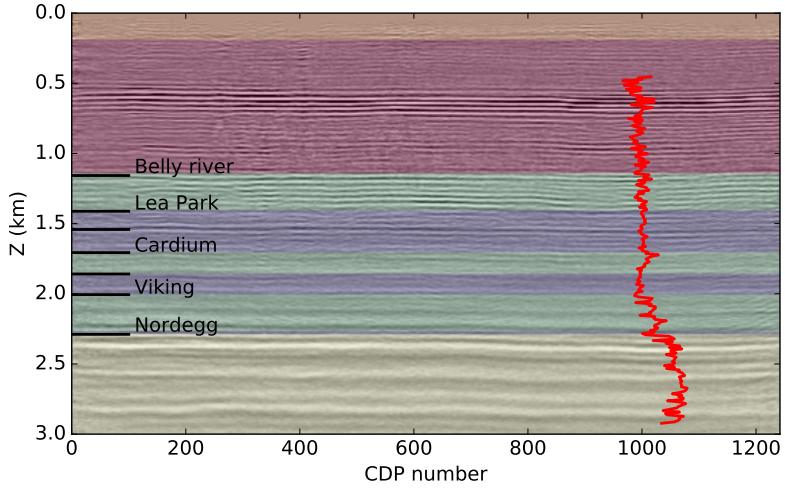

Figure 6.1: Schematic geological interpretation of the survey area. The coloured overlay indicates the dominant rock types associated with various geological formations identified in the region. The depths of the first two layers are not accurate and possibly exaggerated as constraints were not available. The first layer (orange) is composed primarily of glacial till and unconsolidated sediments. The second layer (red) is a mixture of sand and gravel which transitions to primarily sandstone layers at some unidentified depth. Hydrocarbon reservoirs are mostly found in the sequence of interlaced sandstone (green) and shale (blue) layers. A transition to carbonate layers (yellow) conincides with a sharp increase in P-wave velocities. Black lines mark geological formation tops identified in a nearby log; notable formation tops are labelled. The position of the sonic log (red line) does not coincide with its true x location and is only included for display purposes.

is composed of glacial till and unconsolidated sediments, with aquifers also documented within these layers. Loose sediment layers are a likely source of attenuation which would require visco-acoustic/elastic inversion to be properly accounted for. Anisotropy is also not accounted for due to a lack of constraints. Anisotropy is likely present either due to inherent rock properties or effective anisotropy from a horizontally layered subsurface (Backus, 1962).

6.2.2 Cynthia 2D land dataset

The Cynthia dataset is a 2D seismic survey composed of 148 dynamite sources and 639 fixed-spread 3C receivers. Sources and receivers are regularly spaced at 30 m and 10 m intervals, respectively. A map of the source and receiver positions is displayed in Figure 6.2a. It should be noted that the x-axis represents the inline direction (North-South). Source and receivers exhibit crossline (y direction) deviations of less than 70 m from a line of best fit for the acquisition. In addition, the variation in source and receiver elevations is less than 40 m

(Figure 6.2b). Given that the variations in y and z source and receiver positions are small relative to the length of the line (6.4 km), we neglect them, thus approximating the survey as a flat 2D line. We do not account for topography in our numerical wave propagation, nor do we apply elevation statics to the data during preprocessing.

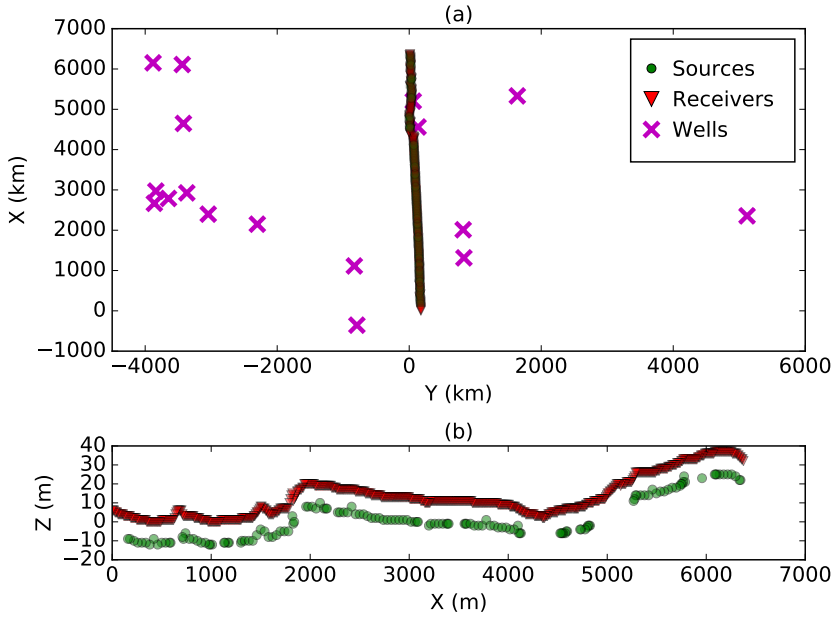

Figure 6.2: Source and receiver distributions after transformation to a local coordinate system. (a) Aerial map of source, receiver and well (sonic log) locations. (b) Inline elevation profile (vertical exaggeration $\sim 20:1$).

TGS supplied raw and preprocessed horizontal and vertical component data along with corresponding PP and PS pre-stack time migrated (PSTM) images. The TGS preprocessing included coherent noise attenuation, ground roll removal, and surface-consistent deconvolution. Representative raw shot records are displayed in Figure 6.3. Average amplitude spectra for the data peak at around 50 Hz for most sources. Ground roll is more prominent in the horizontal component (Figures 6.3b, d). The PSTM images displayed in Figure 6.4 provide initial indications of subsurface structure. Both PP and PS images indicate a finely layered subsurface with multiple strong, flat reflectors. While TGS provided processed data, we opt to conduct our own processing, catering it to the needs of various forms of FWI.

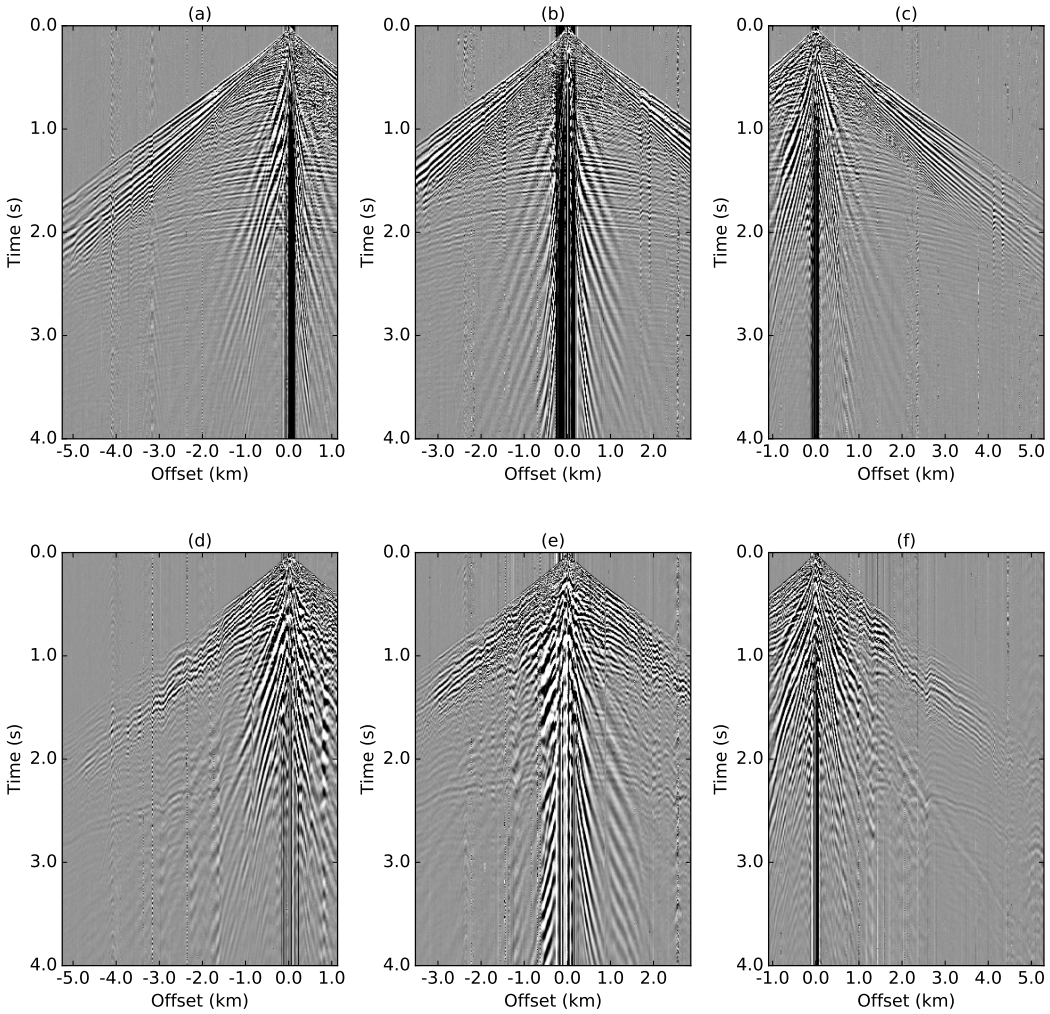

Figure 6.3: Raw (a, b, c) vertical and (d, e, f) horizontal component data. (a, d) Shot #10 ($x = 1.1$ km), (b, e) Shot #94 ($x = 2.86$ km) and (c, f) Shot #100 ($x = 5.3$ km) in sequence.

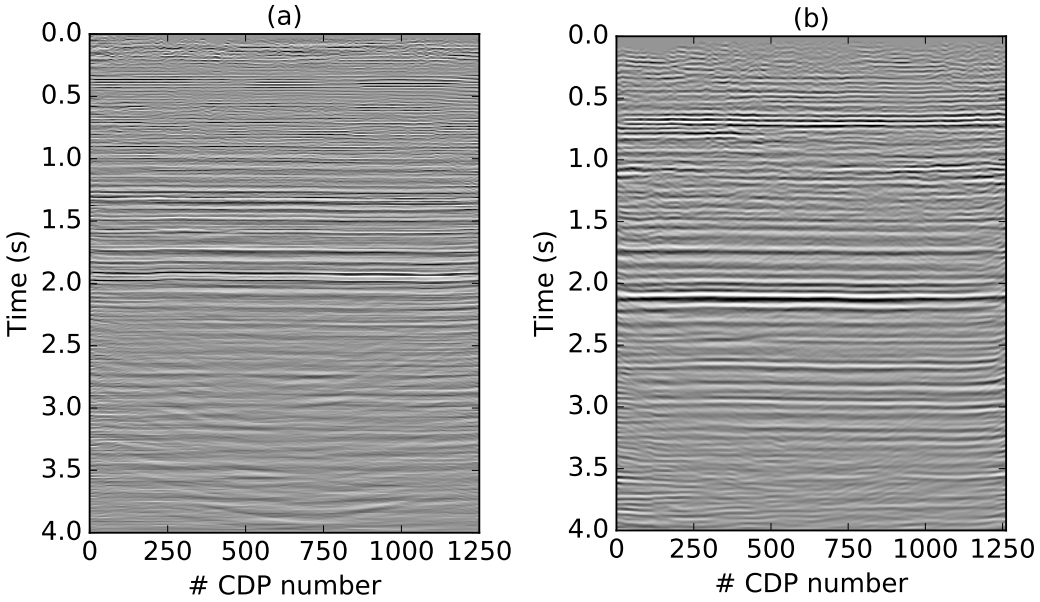

Figure 6.4: Pre-stack time migrated (PSTM) images supplied by TGS. (a) PP image (b) PS image.

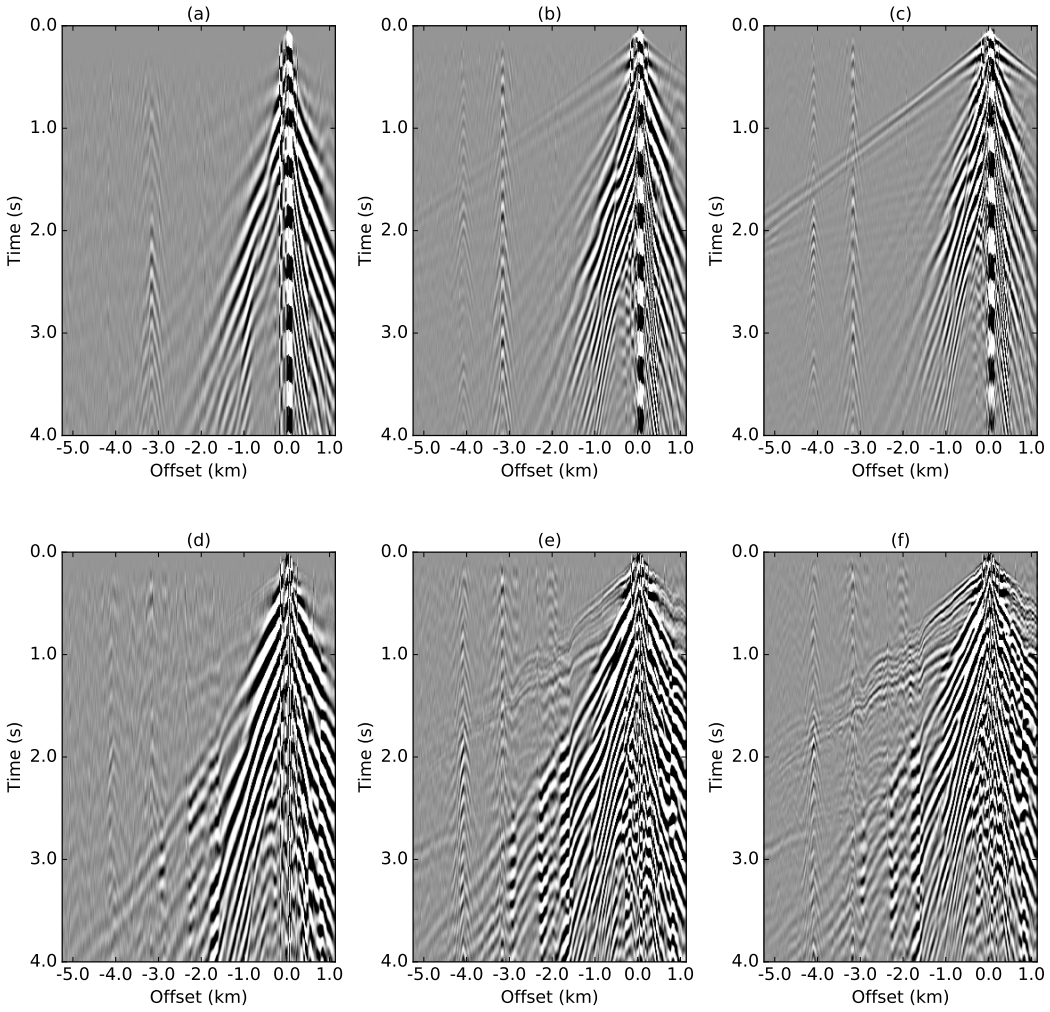

Figure 6.5: Bandpass filtered raw data for shot #10. (a, b, c) Vertical component data. (d, e, f) Horizontal component data. (a) 2-4 Hz vertical. (b) 2-6 Hz vertical. (c) 2-8 Hz vertical. (d) 2-4 Hz horizontal. (e) 2-6 Hz horizontal. (f) 2-8 Hz horizontal.

Low frequencies in the data are essential to FWI as they constrain the long-wavelength components of the velocity model (Virieux and Operto, 2009). We examine the data in various frequency bands to assess the usability of the lower frequencies; the filtered data are presented in Figure 6.5. Ground roll and noise dominate the 2-4 Hz bandpass (Figure 6.5a, d). No significant P-wave energy is apparent except at very short offsets. At 2-6 Hz, usable P-wave signal and weak reflections are evident on the vertical component (Figure 6.5b); the SNR on the horizontal component remains poor. Clear signal is observed in the 2-8 Hz bandpass for both horizontal and vertical component data; however, the noise levels

in the horizontal component are still problematic. Horizontal component data exhibit rapid time shifts within certain offsets (e.g., between -2 and -3 km offsets in Figure 6.5f). This behaviour is also apparent in the fullband raw data between -3–2 km and 2-3 km offsets in Figures 6.3d and f, respectively. Upon further examination, we observe that the rapid variations occur within a particular range of offsets that coincide with a fixed location in the survey, specifically between $x = 3 - 4$ km. Limited elevation variations in this region make elevation an unlikely source of the shifts. In addition, we do not observe similar time-shifts in the vertical components which would be expected if the shifts were attributed to elevation variations. In fact, within the range of interest, there are no discernible variations in the vertical component data. While we cannot confirm it, we speculate that this effect is a consequence of a localized, shallow shear-wave velocity anomaly. The low-frequencies (2-4 Hz) and horizontal component data are too noisy to utilize for FWI. The complex wave phenomena in the horizontal component also makes it challenging to fit with FWI. Based on these factors, we focus our attention on fitting vertical component data for frequencies above 4 Hz.

6.3 Preprocessing

While preprocessing is a necessary step for FWI, it's importance is arguably reduced compared to more conventional seismic imaging procedures. There are a number of factors that contribute to this. First, the importance of low frequencies in FWI means that practitioners emphasize preserving as much signal as possible. As such, unless noise attenuation algorithms can avoid harming lower frequency signal, they may be avoided. In relation to this, the influence of noise or undesirable features in the data can be suppressed by the FWI algorithm itself through methods including selective windowing of the data, the use of robust objective functions and gradient preconditioning. In practice, it is not uncommon to utilize all these approaches to some degree. Finally, since FWI involves simulating the wave equation, it can account for multiples, elevation variations, source and receiver ghost effects, and other effects that are commonly removed for conventional imaging.

The primary objective of preprocessing in FWI is to remove, or normalize, features in the data that cannot be modelled by the wave propagation engine. For example, in a land setting acoustic modelling will not generate shear waves or surface waves; therefore, they should either be removed from the data or excluded from the inversion to prevent erroneous fitting of the data. Our processing sequence emphasizes the removal of coherent noise, amplitude corrections and ground roll attenuation for the acoustic inversion. Ground roll removal is done via FK (frequency-wavenumber domain) filtering. The processing flows are displayed in Figure 6.6. Since ground roll is not removed in the elastic inversion, we are able to use

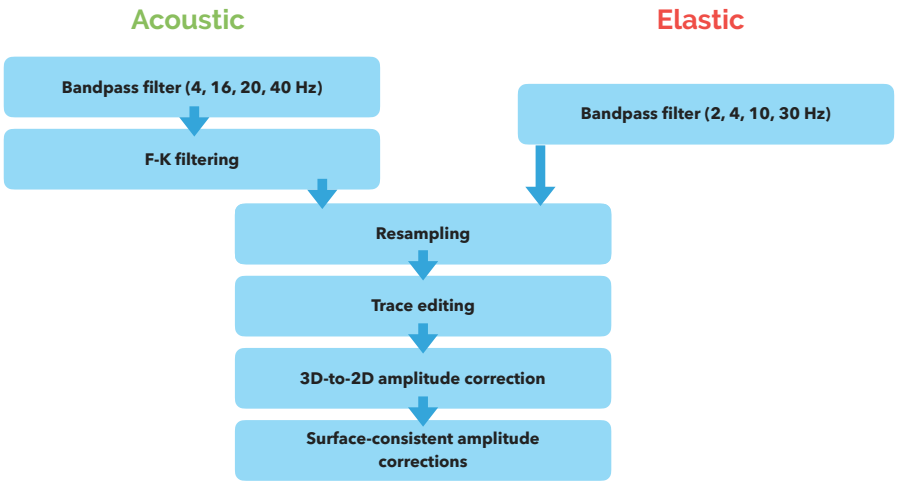

Figure 6.6: Processing sequences for acoustic and elastic FWI.

a shorter taper at the lower frequencies. The 3D-to-2D amplitude corrections are a crude geometrical spreading correction that correct for the fact that we are using a 2D propagator to simulate 3D field data (Crase et al., 1990).

Variable source and receiver coupling along with complex near surface structure can contribute to inconsistent signal amplitudes between sources and receivers. To mitigate this effect, we apply surface-consistent amplitude corrections to the data. The original surface-consistent hypothesis uses a convolutional model to decompose the data into source, receiver and average amplitude terms (Taner and Koehler, 1981). Different studies have proposed alternative decompositions, for example, Cary and Lorentz (1993) use a four component convolutional model that separates the data into source, receiver, offset and common-depth point terms. van Vossen et al. (2006) replace the offset and common-depth point terms with Green's functions computed for the subsurface. Our approach is similar to that of Kamei et al. (2015) and can be viewed as an extension, in principle, to the study of van Vossen et al. (2006). We focus solely on amplitude corrections, thus separate the root-mean square amplitudes of the data into three terms:

$$D_{ij} = U_{ij} S_i R_j, \tag{6.1}$$

where $D_{ij} = \sqrt{\int_T \mathbf{d}_i(\mathbf{x}_j, t)^2 \, \mathrm{d}t}$ and $U_{ij} = \sqrt{\int_T \mathbf{u}_i(\mathbf{x}_j, t)^2 \, \mathrm{d}t}$. The decomposition is in terms of source (S_i), receiver (R_j) scalars and the RMS amplitudes of the synthetic data. To solve

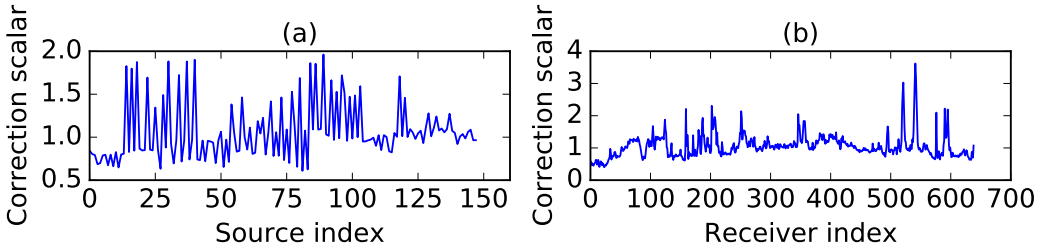

Figure 6.7: Surface-consistent scalars for acoustic inversion. (a) Source scalars. (b) Receiver scalars.

for S_i and R_j, Equation 6.1 is linearized by taking a log-transform such that

$$\tilde{D}_{ij} = \tilde{U}_{ij} + \tilde{S}_i + \tilde{R}_j, \tag{6.2}$$

where the tilde denotes the log-transformed variable (e.g., $\tilde{D}_{ij} = \log D_{ij}$). The contribution from the synthetic data is known and can be subtracted from the left hand side of Equation 6.2 to give

$$\tilde{\Delta}_{ij} = \tilde{S}_i + \tilde{R}_j, \tag{6.3}$$

where $\tilde{\Delta}_{ij} = \tilde{D}_{ij} - \tilde{U}_{ij}$. The linear problem is solved with Gauss-Seidel iterations that update S_i and R_j in an alternating manner (Cary and Lorentz, 1993). The decomposition we propose assumes that the data and modelled data are similar down to a source and receiver amplitude scaling. This is unlikely to be true in practice since errors in the velocity model, source wavelet and modelling physics result in synthetics that do not match the data. In our case, this issue is partially alleviated by source inversion and an initial model that provides a reasonable initial fit to the data. In addition, we use time windows to limit the data and synthetics to early arriving phases where the similarity is highest. Windowed versions of the data and synthetics are used to solve the problem in Equation 6.3.

The estimated source and receiver amplitude scalars are presented in Figure 6.7. To assess the correction, Figure 6.8 plots the RMS amplitudes for every trace in the data before and after corrections. Columns or rows with low/high RMS values, relative to their neighbours, correspond to inconsistent amplitudes in receivers or sources, respectively. Figure 6.8 depicts a number of receivers with relatively high RMS values prior to surface consistent corrections. After the corrections, the source and receiver amplitudes are more balanced. The effect of these corrections can be observed in the shot gathers before and after surface consistent corrections in Figure 6.9. Traces with high RMS amplitudes, relative to neighbouring traces, have been scaled down after surface consistent corrections. We observe a weak offset depending scaling (after processing) that arises from geometrical-spreading related amplitude

differences between the data and synthetics that persist even after the 3D-to-2D correction. We address concerns with the current amplitude scaling scheme in later discussion.

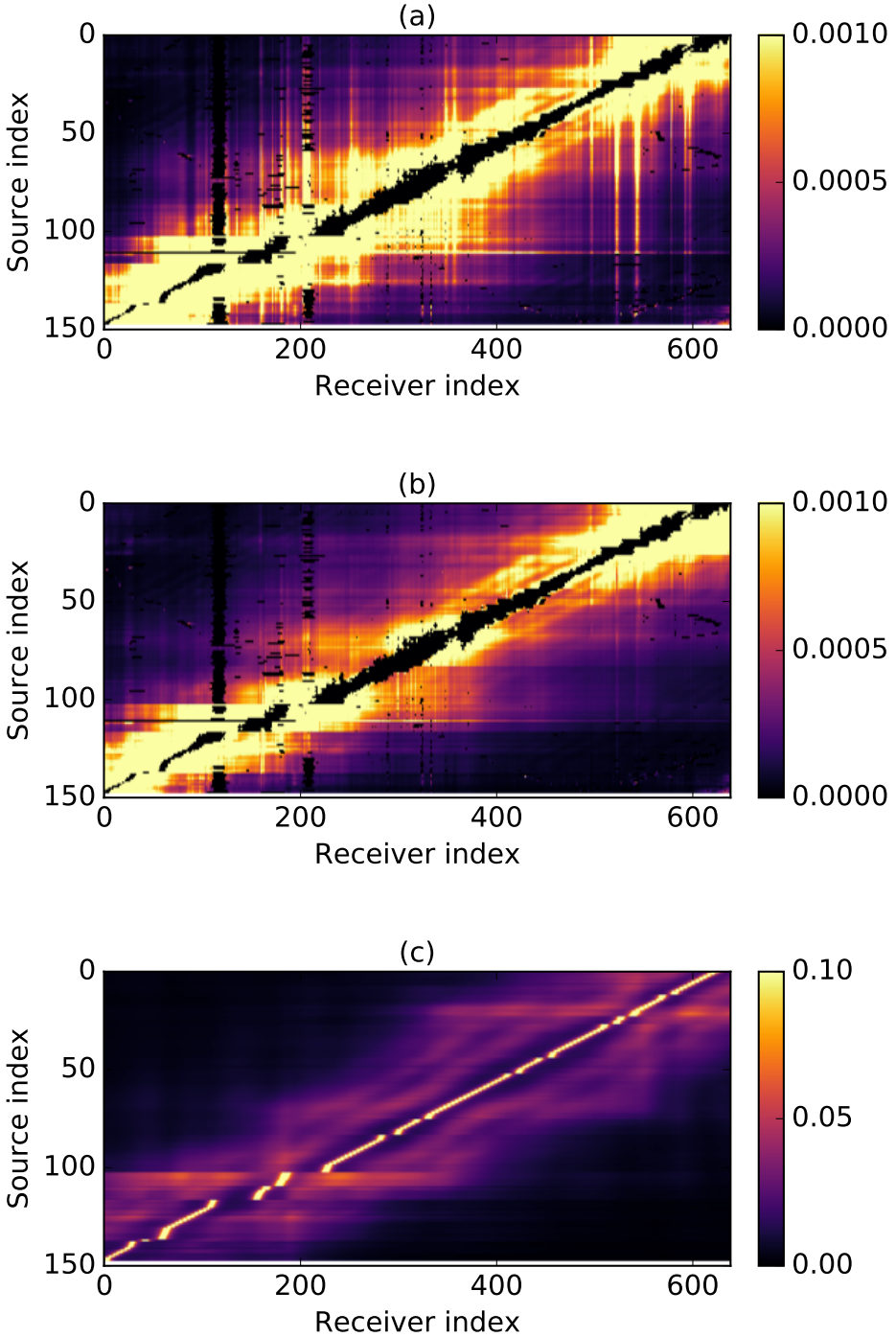

Figure 6.8: Tracewise RMS amplitudes of data. (a) Processed acoustic data before surface-consistent corrections. (b) Processed acoustic data after surface-consistent corrections. (c) Acoustic synthetics. Gaps in the amplitude maps occur due to mutes applied to the data.

Examples of the final processed data for acoustic and elastic inversion are presented in Figures 6.9d-f and 6.10. Figure 6.11 displays the average amplitude spectra for 3 shots before and after processing. The data processed for elastic inversion possess more low frequency content since the ground roll has not been removed.

6.4 Initial model building

The optimization landscape in FWI is highly non-convex and contains numerous local minima (Virieux and Operto, 2009). The formulation of FWI as a linearized inverse problem assumes that the initial model is in the vicinity of the true model. When this assumption is violated, the local optimization will likely converge to a local minima that is not representative of the true subsurface Earth model. A 'good' initial model is characterized as one that does not produce synthetics that are cycle skipped. Cycle skipping is the phenomenon whereby distinct phases in the simulated and observed data are shifted in time by more than half a period of one another. Such behaviour is symptomatic of bulk errors in the velocity model. Despite algorithmic advances in recent years, a good initial velocity model remains the most essential ingredient to successful applications of FWI.

Our first iteration velocity model is an interval velocity model obtained through a Dix conversion of the migration velocities provided by TGS (Dix, 1955). The synthetics generated with this velocity model are cycle skipped by 2-3 periods. Bandpass filtered and fullband data comparisons for select shots are presented in Figures 6.12 and 6.13. The synthetics are delayed at all offsets indicating that the current velocity model is too slow. Given how prominent the cycle skipping is, even in the lower bandpass, we do not anticipate being able to invert for such significant velocity errors. To develop a more reliable initial model, we combine information from sonic logs and first-break traveltime tomography.

6.4.1 Near-surface tomography

First-break traveltime tomography is performed using the TomoPlus software developed by GeoTomo. First-breaks in the data are picked manually and exhibit good coherence except at higher offsets ($5-6$ km). Ray tracing indicates that the penetration depth of refracted arrivals is severely limited. The highest ray densities are observed between 300-400 m depth, with only a few rays penetrating to the maximum penetration depth of 600 m. We expect diving-wave FWI to allow for slightly deeper updates than this since FWI accounts for the finite-frequency nature of wave propagation. Diving waves describes a set of wave modes that are redirected to the Earth's surface via refraction. Preliminary tests indicate that FWI yields reliable updates down to depths of 750 m. The final tomography model is displayed

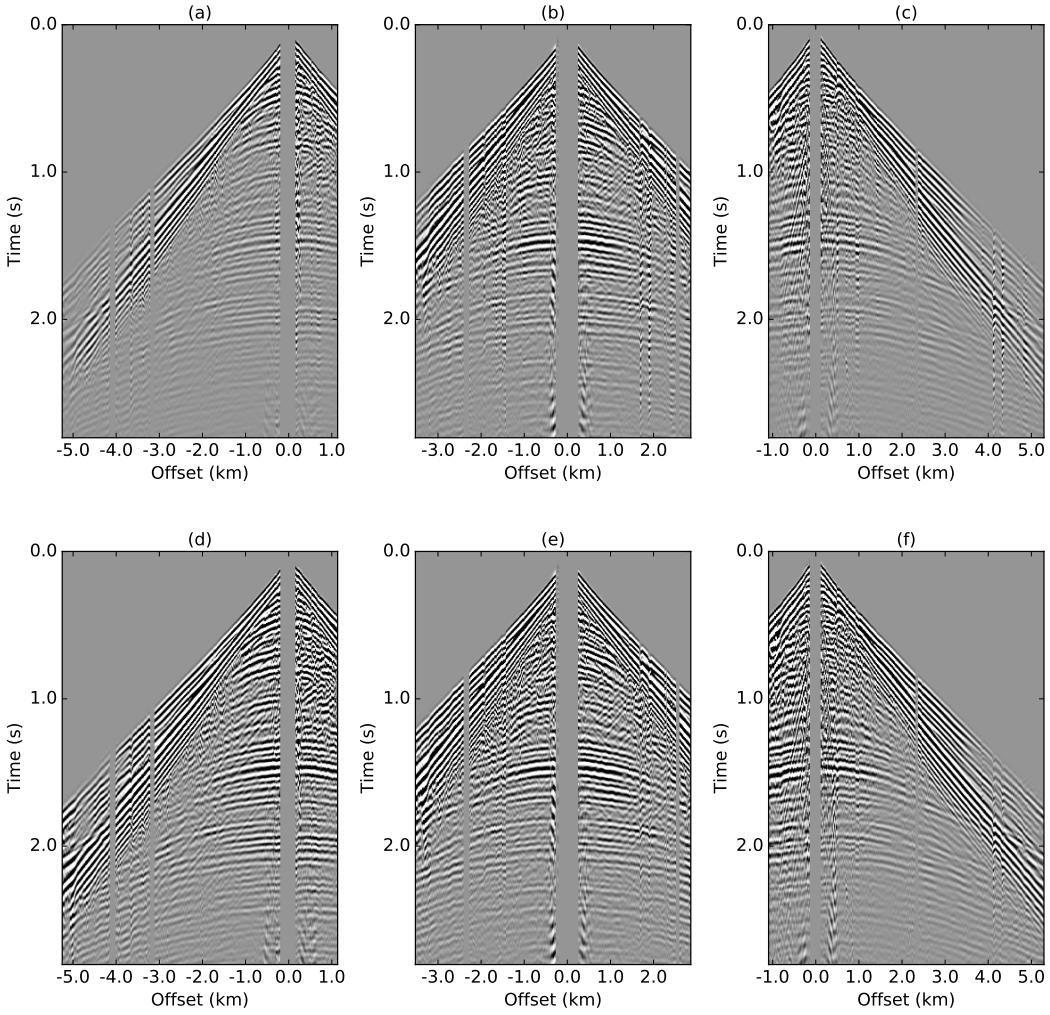

Figure 6.9: Comparison of processed data (acoustic inversion) (a-c) before and (d-f) after surface-consistent corrections. (a, d) Shot #10 ($x = 1.1$ km), (b, e) Shot #94 ($x = 2.86$ km) and (c, f) Shot #100 ($x = 5.3$ km). Processed data after surface consistent corrections are the final data used for FWI.

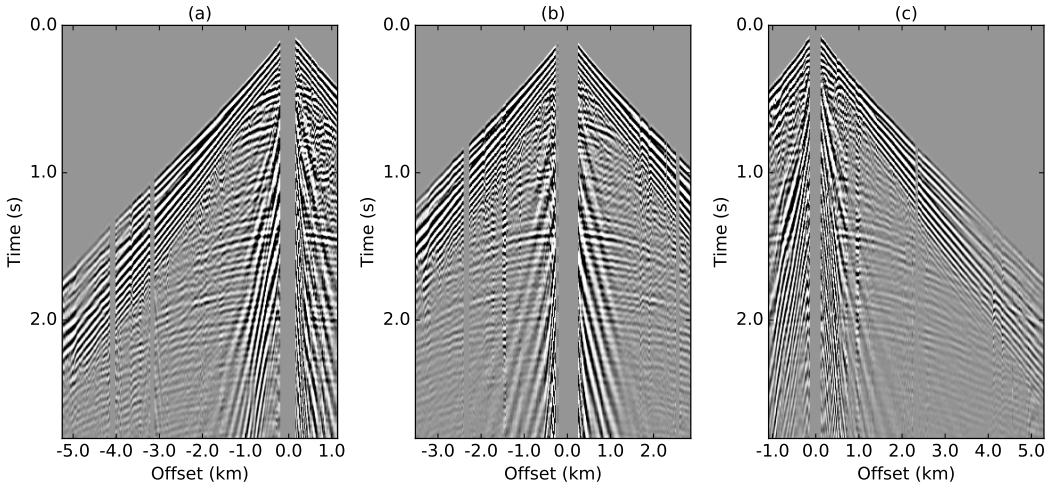

Figure 6.10: Processed data for elastic inversion. (a) Shot #10 (x = 1.1 km), (b) Shot #94 (x = 2.86 km) and (c) Shot #100 (x = 5.3 km). FK filtering is not applied to the elastic data.

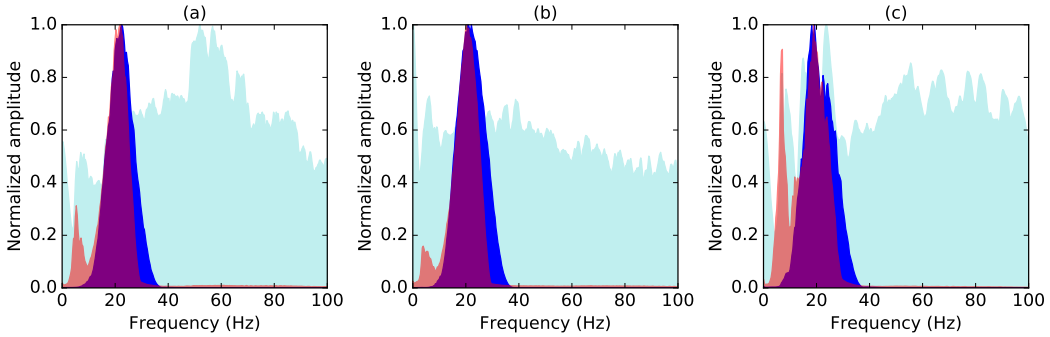

Figure 6.11: Comparison of amplitude spectra before and after processing. Spectra for the raw (cyan), processed acoustic (blue) and processed elastic (red) data are displayed. (a) Shot #10 (x = 1.1 km), (b) shot #94 (x = 2.86 km) and (c) shot #100 (x = 5.3 km).

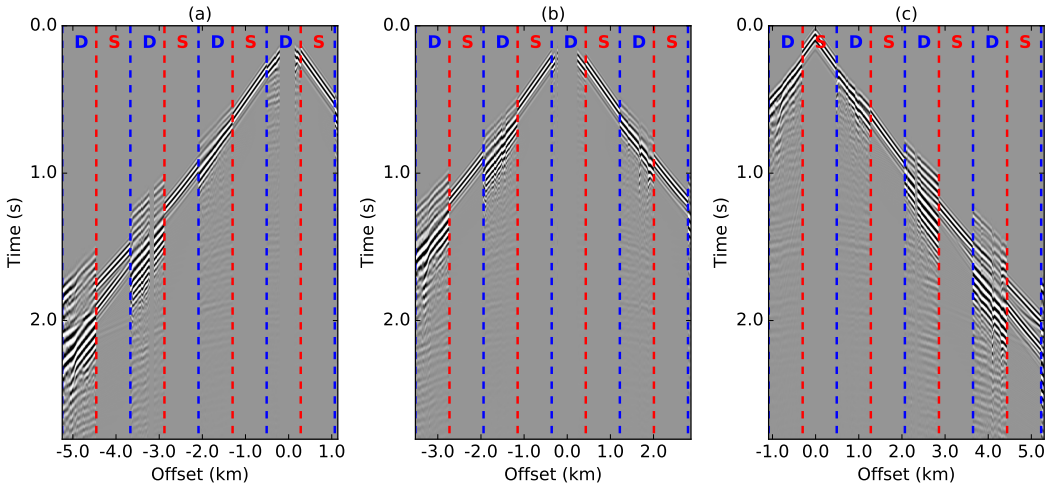

Figure 6.12: Data comparison for acoustic observations and synthetics generated in the interval velocity model. (a) Shot #10 ($x = 1.1$ km), (b) Shot #94 ($x = 2.86$ km) and (c) Shot #100 ($x = 5.3$ km). The sections display observed and synthetic traces in interlaced blocks. Viewed from left-to-right, data traces are bound by dashed blue $\rightarrow$ red lines, whereas synthetic traces are bound by dashed red $\rightarrow$ blue lines. The interval velocity model yields synthetics that are significantly cycle skipped. Sections are trace normalized for display purposes.

in Figure 6.14. Figure 6.15 displays a comparison of bandpass-filtered data and synthetics computed in the initial tomography model. In the 4-8 Hz band, the data and synthetics exhibit good agreement, particularly in the first breaks. For the elastic inversion, an S-wave velocity model is also required; however, we lack reliable constraints on S-wave velocities. Shallow v_p/v_s ratios are only available in some distant logs (approximately 5 km crossline distances). The shallow logs exhibit $v_p/v_s \approx 2$ between 100-600 m depths. Low v_s in the near surface is consistent with the slow moveout of surface waves in the data which are known to propagate at approximately $0.9v_s$ and be sensitive to structure within 1-2 shear wavelengths of the surface. Based on these observations, we assume a constant v_p/v_s ratio of 2.0 below 100 m, and a linearly increasing v_p/v_s ratio to a maximum of 2.5 just beneath the surface. Some shallow sonic logs, while removed from the survey area, demonstrate an improved agreement with the tomographic initial model (Figure 6.16).

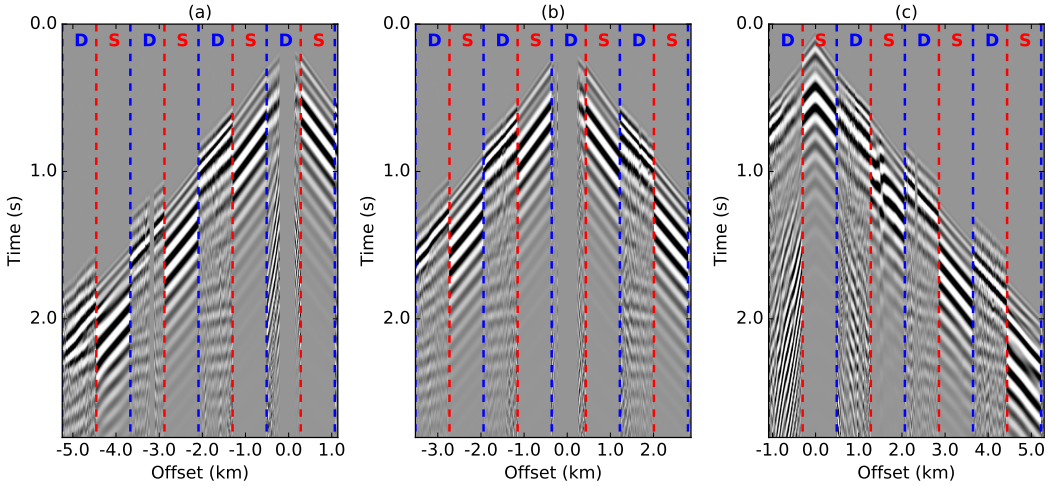

Figure 6.13: Data comparison for bandpass filtered (4-8 Hz) acoustic observations and synthetics generated in the interval velocity model. (a) Shot #10, (b) Shot #94 and (c) Shot #100.

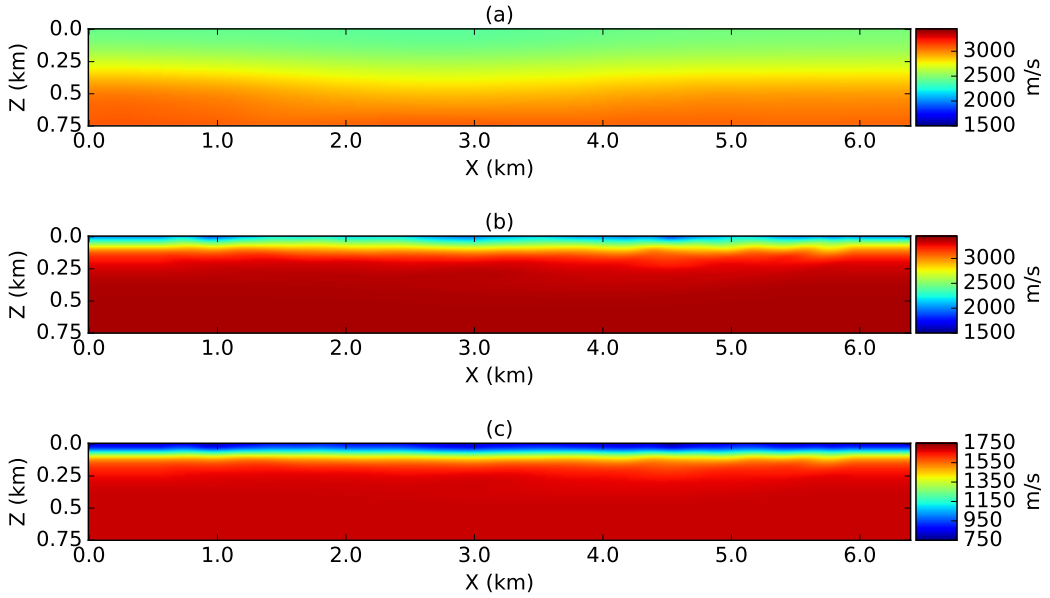

Figure 6.14: Initial velocity models. (a) P-wave velocity model obtained via Dix conversion of migration (PSTM) velocities. (b) P-wave velocity model obtained from near-surface traveltime tomography. (c) S-wave velocity model.

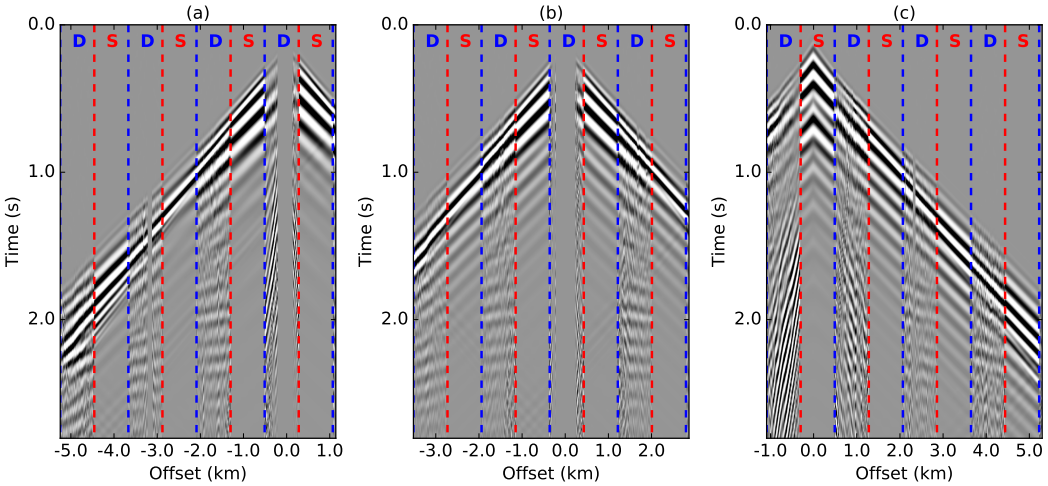

Figure 6.15: Data comparison for bandpass filtered (4-8 Hz) acoustic observations and synthetics generated in the initial velocity model (after tomography). (a) Shot #10, (b) Shot #94 and (c) Shot #100.

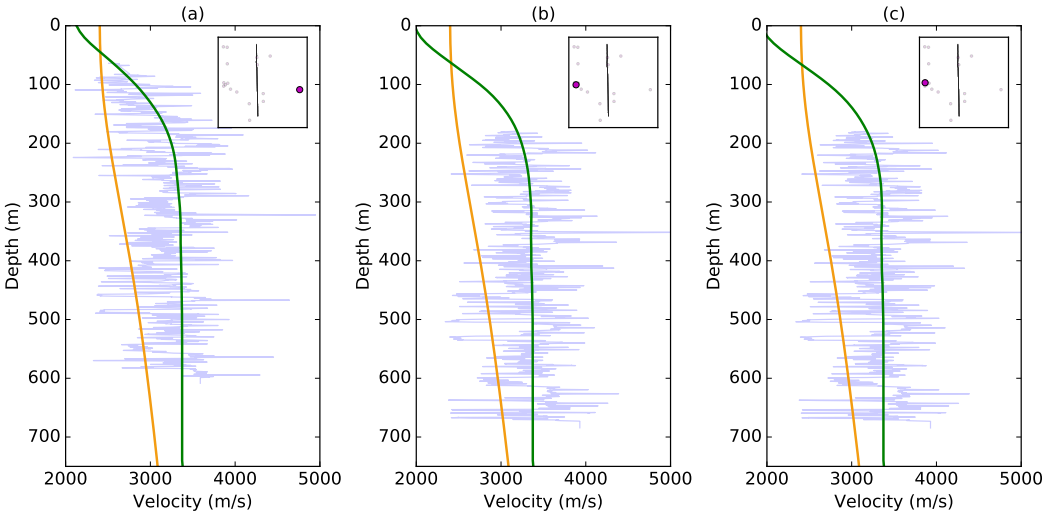

Figure 6.16: Sonic logs for shallow structure. The plots display depth profiles for the sonic logs (light blue), interval velocity (orange) and initial (green) P-wave velocity models. Inset maps display the locations of the logs relative to the survey line. The interval velocity is significantly slower than the tomography model and sonic logs.

A laterally homogeneous velocity model, derived from a smoothed sonic log, is inserted below 750 m depth. The PSTM images suggest a flat subsurface with no significant dips. This interpretation of the subsurface is consistent with the sonic longs which exhibit very similar depth profiles. A range of sonic logs are plotted with the updated initial model in Figure 6.17. A density model is generated using Gardner's relation (Gardner et al., 1974); the resultant model shows a reasonable agreement with available log constraints (6.17). Unlike the sonic logs, the density logs do not display a slight decrease in density at 2 km depth. We acknowledge that we have introduced an apparent error into the density model; however, due to the limited effect that density has on the kinematics of wave propagation, we proceed without further corrections.

6.5 FWI workflow

While FWI is a highly non-linear inverse problem, a number of successful strategies have been established to assist with convergence to meaningful subsurface models. Developing successful FWI strategies requires careful consideration of the dataset, acquisition, and the parameters that are to be estimated. Limitations of the dataset often restrict the scope of FWI. The properties of the dataset are used to customize an FWI workflow that promotes robust convergence. We use a similar workflow for both acoustic and elastic inversions, with the main differences coming in the choice of parameters in the optimization scheme.

Synthetic data are generated with a 2D time-domain, staggered-grid finite-difference solver (Virieux, 1986; Levander, 1988). The solver simulates isotropic $P-SV$ wave propagation in the elastic case; the variable-density acoustic solver is a modified version of the elastic code. During inversion, we mute the near offsets (< 300 m) to prevent contamination from strong noise in the data. Time windowing is utilized to focus the inversions on diving P-waves and other early arriving events. For the acoustic inversion, since elastic effects are neglected, it is not meaningful to fit amplitudes in the data; therefore, we replace the waveform objective function with the global correlation objective function (Routh et al., 2011; Choi and Alkhalifah, 2012). The global correlation objective is defined in the time-domain as

$$J(\mathbf{m}) = -\sum_{s=1}^{N_s}\sum_{r=1}^{N_r} \frac{\int_T \mathbf{u}_s(\mathbf{x}_r,t;\mathbf{m})\cdot\mathbf{d}_s(\mathbf{x}_r,t)\ \mathrm{d}t}{\sqrt{\int_T \mathbf{u}_s(\mathbf{x}_r,t;\mathbf{m})^2\ \mathrm{d}t}\sqrt{\int_T \mathbf{d}_s(\mathbf{x}_r,t)^2\ \mathrm{d}t}}, \tag{6.4}$$

and is analogous to phase-only objective functions used in frequency-domain FWI (Shin and Min, 2006). Equation 6.4 is equivalent to minimizing a normalized waveform difference objective function (Choi and Alkhalifah, 2012). By matching phases in the data and synthetics, we focus on fitting the kinematics of the data. Altering the objective function in FWI only changes the adjoint source used during gradient computations (Plessix, 2006). The adjoint source for the global correlation objective function is defined by Routh et al. (2011); Choi and Alkhalifah (2012). While elastic inversion allows us to account for elastic effects, amplitude information is still unreliable due to errors from 2D modelling and the neglecting of attenuation. For this reason, we continue to emphasize fitting phase-information with the global correlation norm in the elastic inversions.

Multi-scale strategies are effective at mitigating non-linearities in the FWI objective function. The principle of multi-scale strategies in FWI is to perform a series of inversions in a hierarchical manner, fitting large scale features of the data before progressing to smaller scale features. Conventionally, scale separation in the data is achieved in the Fourier domain, with large scales corresponding to low frequencies in the data and vice versa. For time-domain FWI, we adopt the multi-scale strategy of Bunks et al. (1995). The method operates in successive frequency stages that increase the upper cutoff frequency of a band-pass filter (applied to the data and synthetics). The initial model for each stage (except for stage one) is taken as the final inverted model from the preceding stage. The frequency bands that we invert range from 4-20 Hz.

FWI is an ill-posed problem meaning an infinite number of models can fit the data equally well (Virieux and Operto, 2009). Model regularization, included explicitly into the objective function, serves to stabilize the inversion and make it more well-posed. Furthermore, model regularization constrains updates by imposing prior assumptions on the model. The tuning of regularization hyperparameters is a costly procedure in FWI. In lieu of conventional

regularization, we constrain inversion updates with a form of gradient preconditioning known as anisotropic scaled Sobolev preconditioning (SSP) (Zuberi and Pratt, 2017). Anisotropic SSP involves applying a wavenumber domain filter with differing vertical and horizontal scale lengths to the gradient. In contrast to a Gaussian filter, SSP allows fine scale structure to feature into the inversion earlier by retaining high wavenumber features while emphasizing low wavenumber ones (Zuberi and Pratt, 2017). An example of raw and preconditioned gradients are presented in Figure 6.19. The SSP parameters μ_0, μ_x, and μ_z follow the definition in equation 16 of Zuberi and Pratt (2017); in all applications we use $\mu_0 = 1$. The SSP promotes the continuity of horizontal features in the gradient and effectively reduces the presence of acquisition related artefacts in the gradient. The SSP scale lengths μ_x and μ_z can be relaxed as the inversion progresses. In this application, we do not find a noticeable difference by relaxing the parameters within a frequency band. We find the best performance is achieved by gradually relaxing the vertical scale length (μ_z), while keeping the horizontal scale length (μ_x) fixed. Reducing both scale lengths increases the presence of acquisition related artefacts in the inverted model. Larger horizontal scale lengths produce preconditioned gradients that are preferentially smoothed in the horizontal direction, consistent with the geological structure suggested by the PSTM images. Each stage of our multi-scale inversion performs 10 preconditioned non-linear conjugate gradient iterations with a parabolic line search (Nocedal and Wright, 2006). Limiting the number of FWI iterations prevents overfitting of the data and mimics the effect of damping. The extent of model updates are also constrained by bound (box) constraints. Parameter values that lie outside of a permissible interval are projected to the boundaries of the user specified interval; the bound intervals are set using log information. This step helps to avoid non-physical updates and potential instabilities.

Source estimation is performed using the time-windowed version of Equation 2.36. In practice, we found windowing yields cleaner, more impulsive source signatures and prevents late arriving energy from appearing in the source. We select a time window that is sufficiently wide to prevent distortion of the source signature. An example of the windowed data are presented in Figure 6.20. Tighter windows are used for the elastic inversion to counteract the increased complexity in the waveforms. The initial synthetics are modelled with minimum-phase Ormsby wavelets that have a flat spectra in the bandwidths present in the data. An independent source wavelet is estimated for each shot.

During inversion, we found it sufficient to update the source wavelet after each frequency stage as opposed to after every FWI iteration. As a final step, we normalize each estimated wavelet by their RMS amplitudes in an effort to equalize the individual source contributions (Cheng et al., 2017). An example of the estimated wavelets before and after source normalization is displayed in Figure 6.21.

Sequence	Frequency band (Hz)	Damping (s)	Anisotropic SSP (μ_x, μ_z)
I/II	4-8	0.13, 0.25	1, 0.10
III/IV	4-12	0.09, 0.18	1, 0.08
V/VI	4-16	0.06, 0.13	1, 0.06
VIII/IX	4-20	0.05, 0.10	1, 0.05

Table 6.1: Parameter choices for multi-scale Acoustic FWI. Each frequency band performs inversion over two time windows.

Source estimation is performed after using each frequency stage as opposed to after every FWI iteration. Offsets below 500 m and above 3500 m are not considered during source estimation. A 0.15 s time window is also applied around the first-breaks of the data and synthetics for source estimation.

6.6 Results

6.6.1 Acoustic FWI

The acoustic inversion begins in the 4-8 Hz frequency band and advances in 4 Hz increments to a final frequency band of 4-20 Hz. Acoustic modelling generates pressure-component seismograms and uses a free surface boundary condition (with absorbing boundaries elsewhere). To fit pressure-component waveforms to velocity measurements in the data, we bury the receivers just beneath the surface following the approach of Plessix et al. (2012). Table 6.1 summarizes the inversion stages and parameters used for the acoustic inversion. While a variable-density model is used, it is not updated throughout the inversion. The inverted v_p models after each frequency stage are depicted in Figure 6.22. The 4-8 Hz inversion introduces a high-velocity layer to the velocity model at around 500 m depth. The continuity of the layer is interrupted for $x > 4.5$ km and is likely caused by 2 400 m source gaps in the acquisition for this area. As the inversion progresses to higher frequencies, the high-velocity layer becomes more refined; however, some imprints of the acquisition become more apparent towards to the boundaries of the model. We also observe a very thin, low-velocity channel at approximately 100 m depth between $x = 3 - 6$ km. In the subsequent elastic inversion, this feature does not appear leading us to believe that it is not genuine and is an artefact of the acoustic inversion. The artefact may represent leakage of low S-wave velocities in the near surface which could occur if shallow converted modes are present in the data. Since the acoustic modelling cannot account for the converted wave modes, the inversion will compensate by trying to fit these modes with erroneous structure.

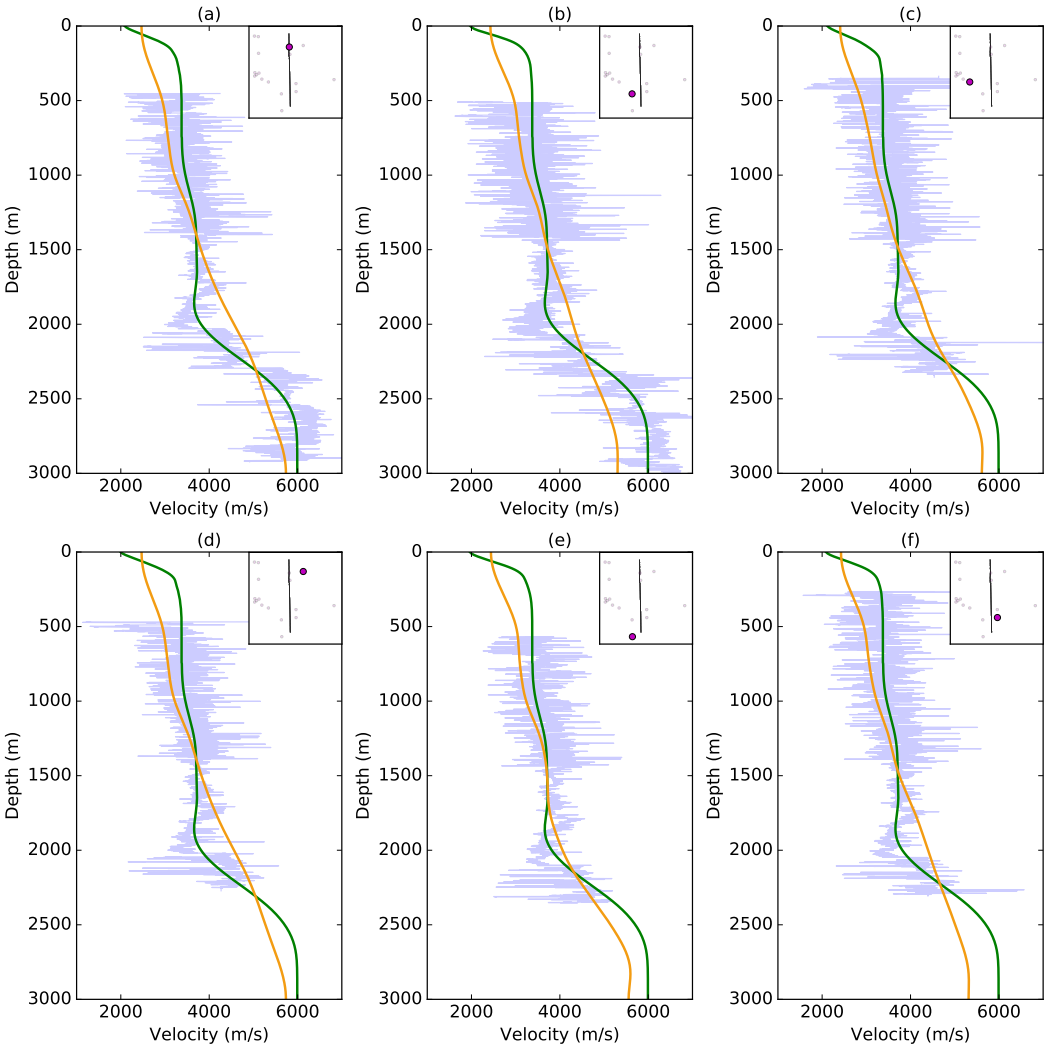

Figure 6.17: Sonic logs for deeper structure. The plots display depth profiles for the sonic logs (light blue), interval velocity (orange) and initial (green) P-wave velocity models. Inset maps display the locations of the logs relative to the survey line. Similar velocity trends are apparent in all the logs, consistent with a flat, layered subsurface. Inset maps display the locations of the logs relative to the survey line.

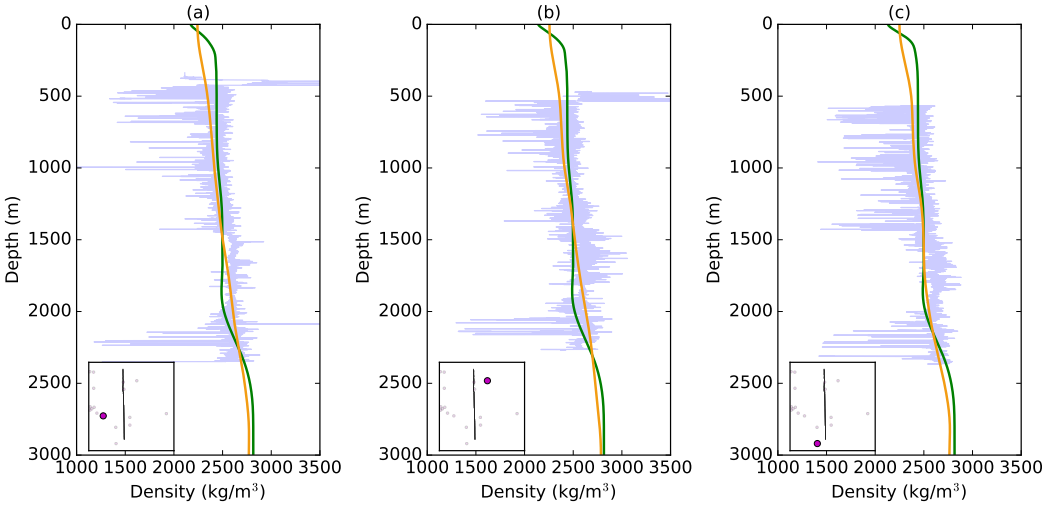

Figure 6.18: Density logs. The plots display depth profiles for the density logs (light blue), interval (orange) and initial (green) density models. The interval and initial density models are derived from Gardner's relation using the corresponding P-wave velocities. Inset maps display the locations of the logs relative to the survey line.

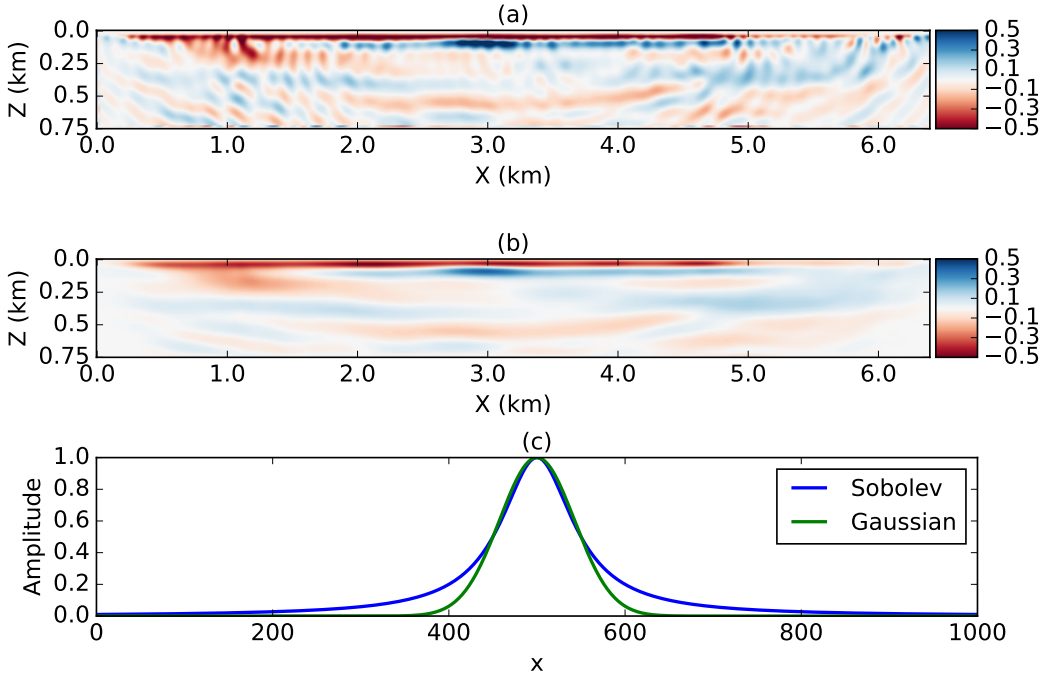

Figure 6.19: Comparison of (a) raw FWI gradient and (b) SSP preconditioned gradient ($\mu_0 = 1.0, \mu_x = 1.0, \mu_z = 0.1$). (c) A comparison of 1D Sobolev and Gaussian filters. The preconditioning attenuates acquisition artefacts and promotes horizontal continuity in the gradient.

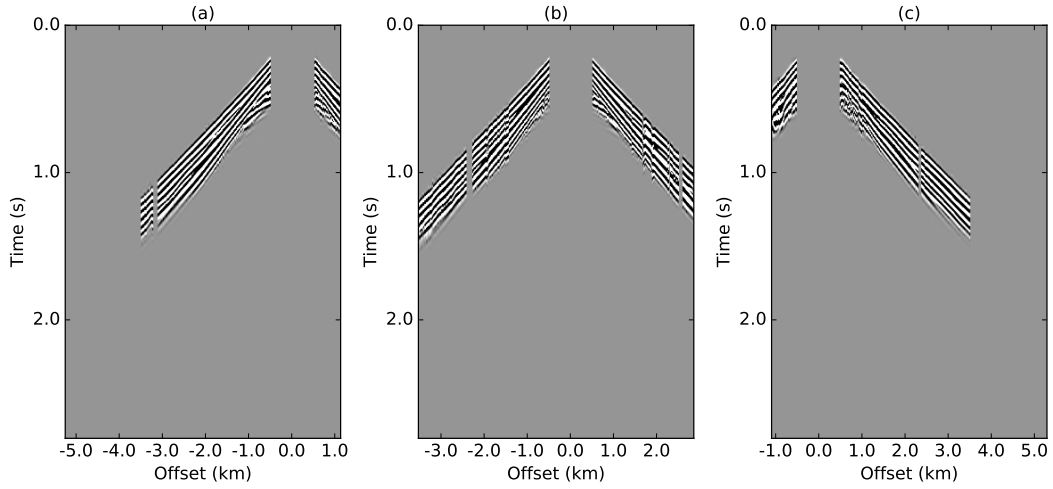

Figure 6.20: Windowed data used for source estimation in acoustic inversion. Offsets are limited to 3.5 km and a tapered time window is applied around the first breaks. (a) Shot #10 ($x = 1.1$ km), (b) Shot #94 ($x = 2.86$ km) and (c) Shot #100 ($x = 5.3$ km).

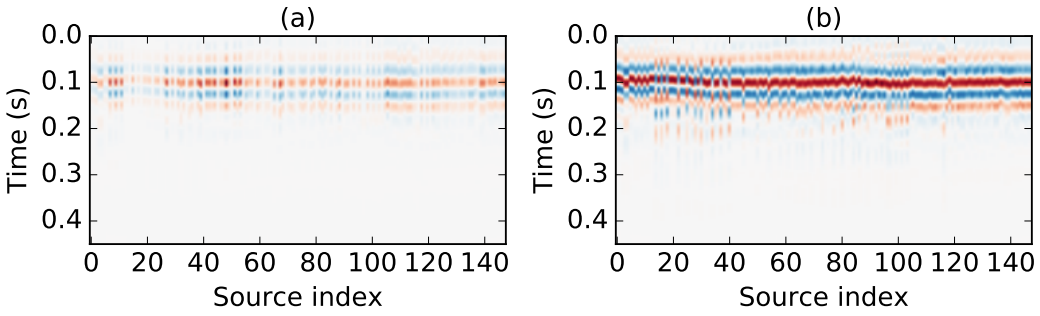

Figure 6.21: Source wavelets for independent shots (a) before and (b) after source normalization. Normalization equalizes the contribution from different sources.

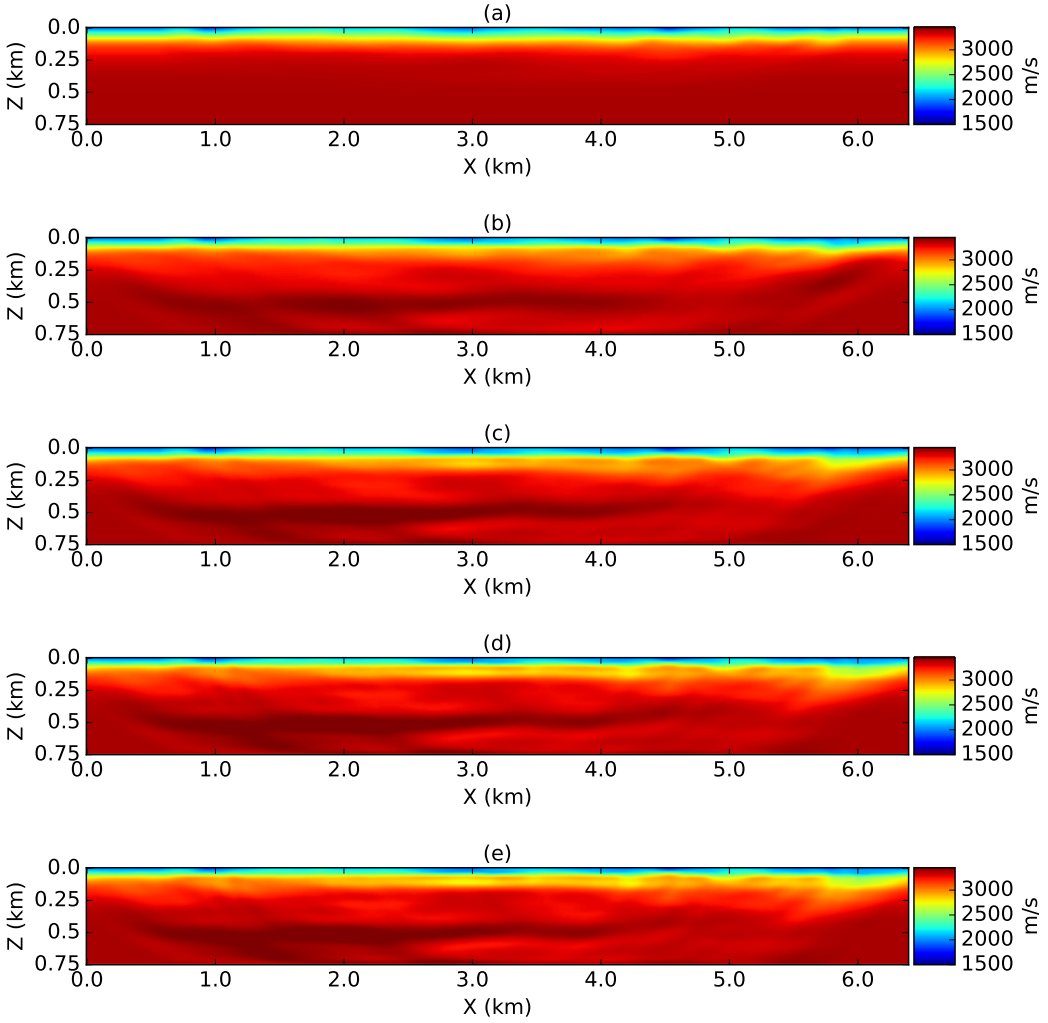

Figure 6.22: Inverted P-wave velocity models from acoustic FWI. (a) Initial model and model after (b) 4-8 Hz, (c) 4-12 Hz, (d) 4-16 Hz and (e) 4-20 Hz inversions.

Figure 6.23 displays a data comparison for acoustic synthetic data generated in the initial velocity model. The initial synthetics show a relatively good fit to the data and is expected due to the first-break fitting performed during tomography. The most notable mismatches occur in the wave coda arriving after the first breaks. In the later phases, the waveforms display less agreement and misalignment of various phases is apparent. After inversion, the waveform match is improved across all offsets (Figure 6.24). Examination of the source wavelet can be used as an additional QC measure. After inversion, the wavelets are expected to be more consistent between shots. While the change is not dramatic, the wavelets for shots 20-40 demonstrate improved similarity after inversion.

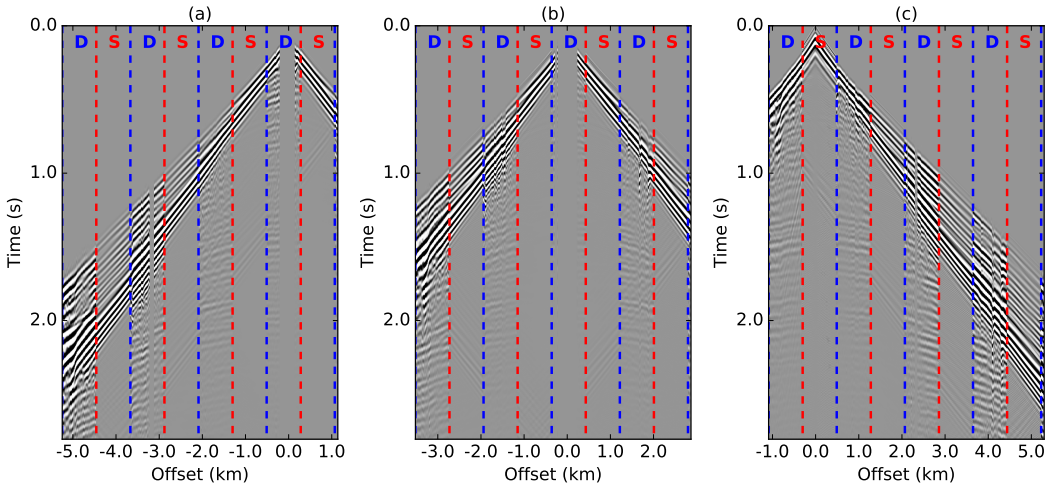

Figure 6.23: Data comparison for acoustic observations and synthetics generated in the initial FWI model. (a) Shot #10 ($x = 1.1$ km), (b) Shot #94 ($x = 2.86$ km) and (c) Shot #100 ($x = 5.3$ km). First breaks are well fit but later arrivals are less consistent at mid/long offsets.

6.6.2 Elastic FWI

Elastic inversion follows similar steps to the acoustic inversion with some additional changes. The synthetic data are velocity component seismograms recorded at the surface. A free surface boundary condition is implemented using stress-image methods (e.g., Levander (1988)). The elastic free surface produces high-amplitude surface waves in the synthetics. Due to errors in our S-wave velocity model, the modelled ground roll propagates faster than in the true data. While recent studies have explored waveform inversion of surface waves, we do not consider it here due to the large inconsistencies in the ground roll signature. Time windows and bottom mutes are used to exclude surface waves from the inversion. Due to the increased complexity of elastic inversion, we take more conservative steps in the multi-scale inversion. The frequency bands inverted range from 4-20 Hz, increasing in 2 Hz increments from a starting band of 4-8 Hz. Both v_p and v_s are considered for inversion parameters; density is updated at each iteration via Gardner's relation. Given that the inversion will be fitting phase-information in the vertical component data, we anticipate limited sensitivity to S-wave velocity structure. Nonetheless, we include it as an inversion parameter to mitigate potential crosstalk between the parameters. A summary of the inversion parameters is presented in Table 6.2.

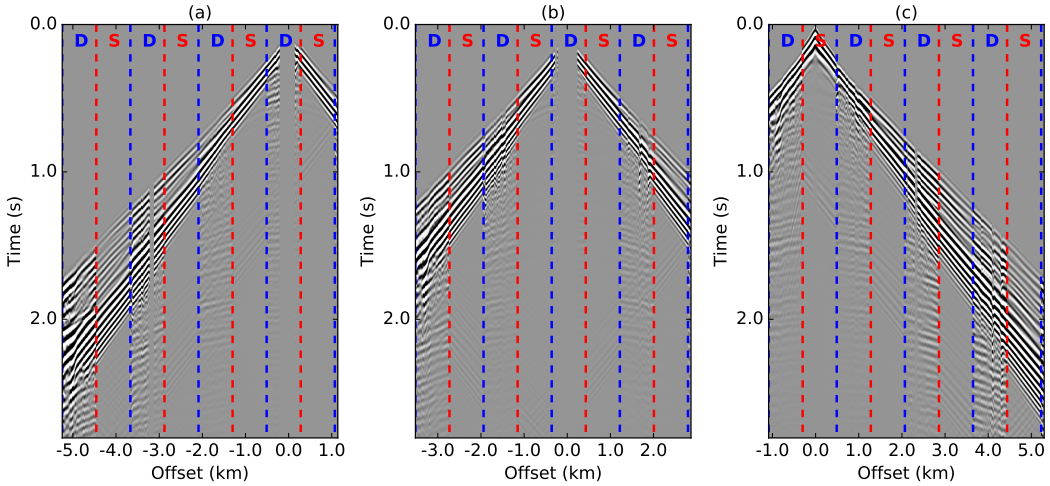

Figure 6.24: Data comparison for acoustic observations and synthetics generated in the final FWI model. (a) Shot #10 ($x = 1.1$ km), (b) Shot #94 ($x = 2.86$ km) and (c) Shot #100 ($x = 5.3$ km). The waveform fit has been improved in the mid to long offsets.

Sequence	**Frequency band (Hz)**	**Damping (s)**	**Anisotropic SSP** (μ_x, μ_z)
I	4-8	0.25	5, 0.10
II	4-10	0.2	4, 0.1
III	4-12	0.2	4, 0.1
IV	4-14	0.2	3, 0.1
V	4-16	0.2	1, 0.1
VI	4-18	0.2	1, 0.06
VII	4-20	0.1	1, 0.05

Table 6.2: Parameter choices for multi-scale elastic FWI.

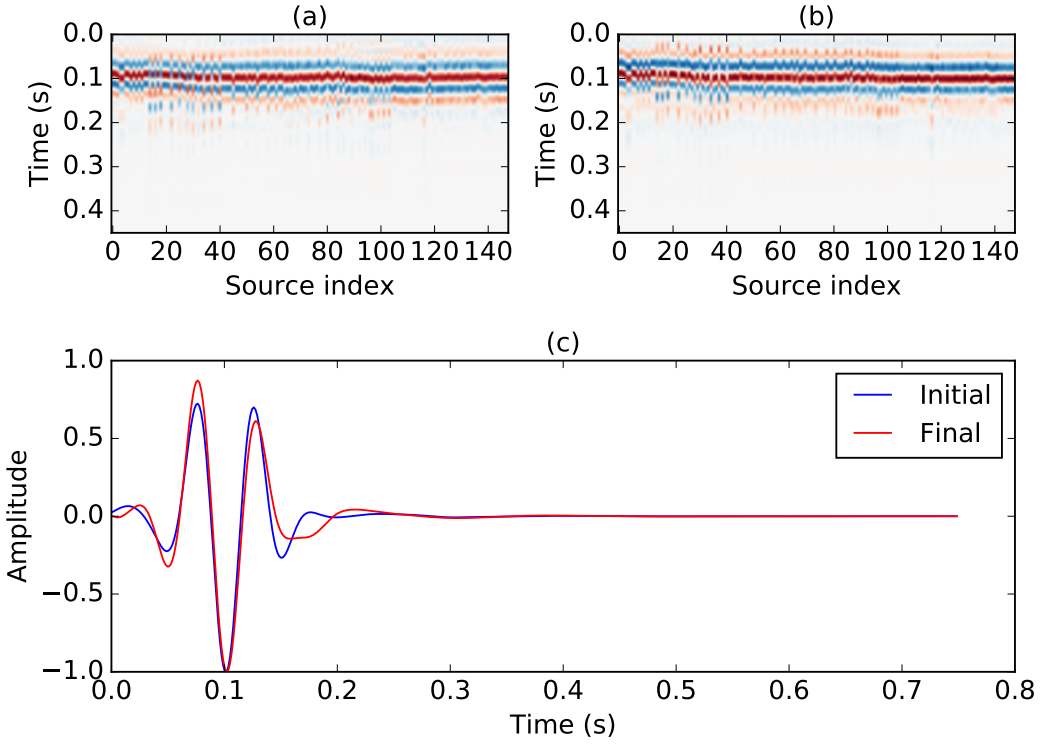

Figure 6.25: Estimated source wavelet (a) before and (b) after acoustic FWI. (c) Comparison of average wavelet before and after inversion.

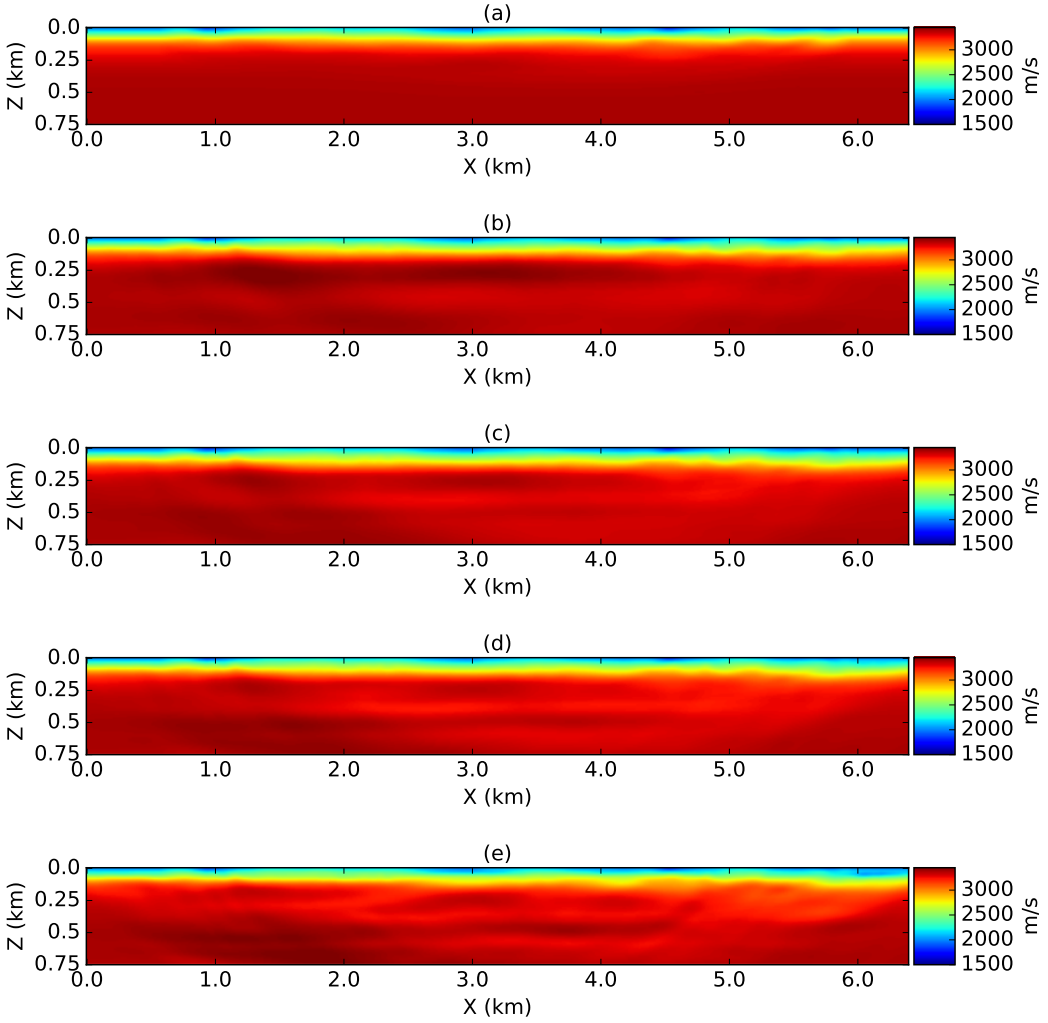

Figure 6.26: Inverted P-wave velocity models from elastic FWI. (a) Initial model and model after (b) 4-8 Hz, (c) 4-12 Hz, (d) 4-16 Hz and (e) 4-20 Hz inversions.

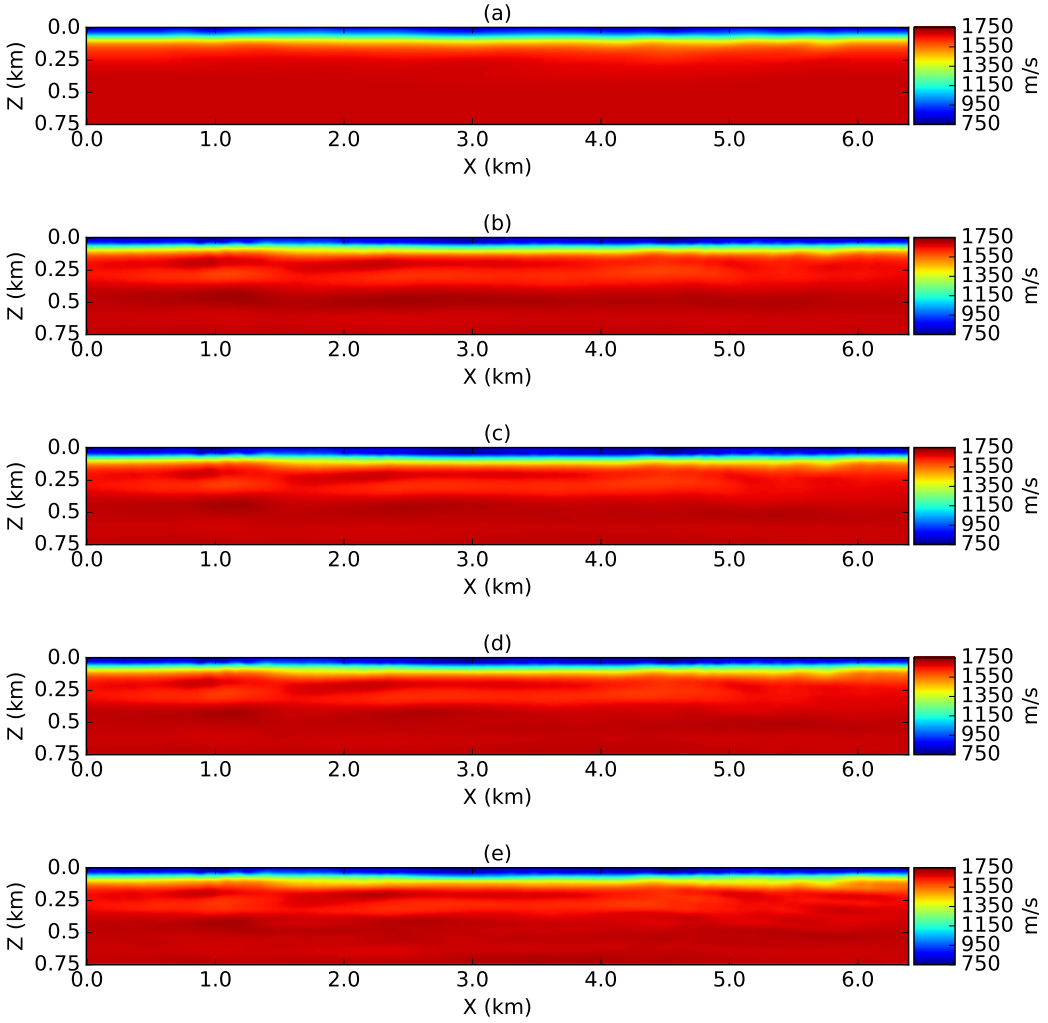

Figure 6.27: Inverted S-wave velocity models from elastic FWI. (a) Initial model and model after (b) 4-8 Hz, (c) 4-12 Hz, (d) 4-16 Hz and (e) 4-20 Hz inversions.

The inverted P- and S-wave velocity models are displayed in Figures 6.26 and 6.27, respectively. The v_p updates exhibit longer-wavelength updates than their acoustic counterparts owing to the increased bandwidth of the processed elastic data. The high v_p perturbation that appeared in the acoustic inversion is also observed in the elastic inversion at higher frequency stages (after 4-16 Hz band). The amplitude of the high velocity layer is reduced relative to the acoustic v_p model, yet it exhibits improved lateral coherency and more well defined layered structure. It is not clear whether the difference arises from the switch to elastic modelling or the additional low-frequency information in the data. The shallow channel feature observed in the acoustic v_p inversion is not apparent here. The v_s model

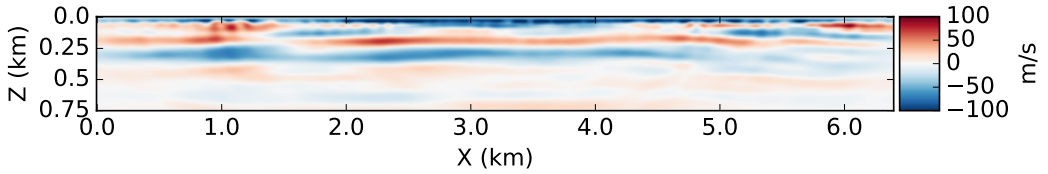

Figure 6.28: Difference between initial and inverted S-wave velocity model. The inversion has poor sensitivity to S-wave structure in general; however, we do observe a decrease in shallow velocities consistent with the data.

updates appear to be limited to depths below 400 m. Due to the lack of constraints, it is challenging to comment on the validity of the v_s updates. We note that the inversion reduces S-wave velocities for depths below 100 m (Figure 6.28). The reduction in velocity is consistent with low velocities indicated by the slow moveout of the ground roll in the data. The final source wavelets display improved consistency, across sources, compared to the initial wavelet. Comparisons of the data before and after inversion are displayed in Figures 6.29 and 6.30, respectively. Similar to the acoustic case, the first-breaks in the synthetics match reasonably well with the observations. The waveform fit for later arriving phases is improved across all offsets after inversion.

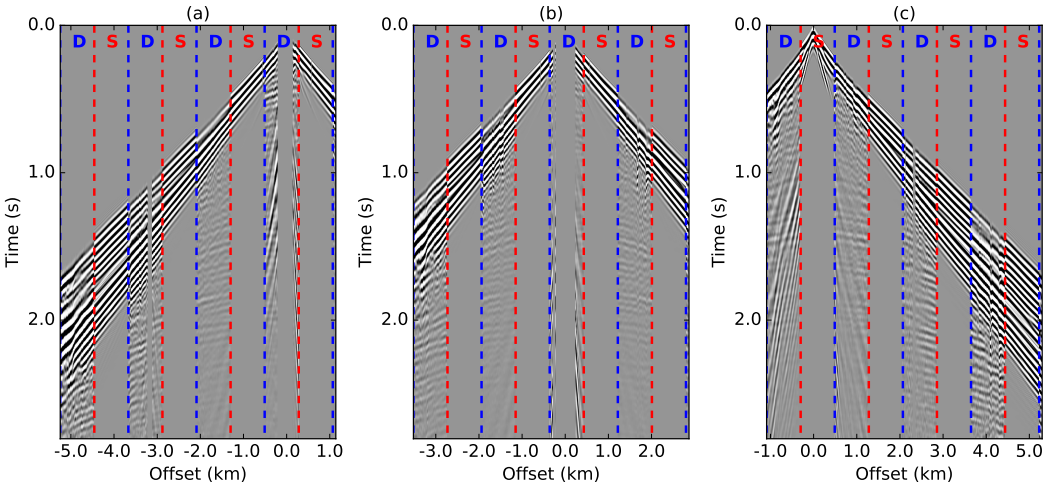

Figure 6.29: Data comparison for elastic observations and synthetics generated in the initial elastic FWI model. (a) Shot #10 (x = 1.1 km), (b) Shot #94 (x = 2.86 km) and (c) Shot #100 (x = 5.3 km). Synthetic ground roll has been muted to prevent obfuscating the sections.

As an initial QC independent of the inversion, we compare the inverted P-wave velocity models from acoustic and elastic inversion to a depth-converted PSTM image in Figure 6.32. The purpose of the comparison is to identify any structural similarities in the velocity

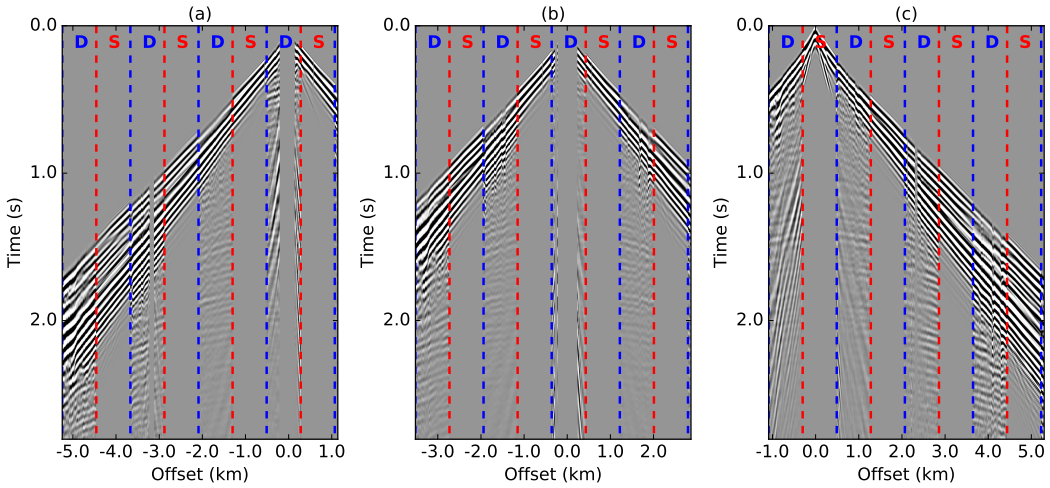

Figure 6.30: Data comparison for elastic observations and synthetics generated in the final elastic FWI model. (a) Shot #10 ($x = 1.1$ km), (b) Shot #94 ($x = 2.86$ km) and (c) Shot #100 ($x = 5.3$ km).

model and the seismic image. In the comparison, a dashed cyan line marks the approximate depth of the base of the high-velocity layer. The presence of a strong reflection event in the image (Figure 6.32c) offers some confidence to the validity inverted models.

6.6.3 Elastic FWI (w/ reflections)

Thus far, we have applied conventional FWI by attempting to fit refracted arrivals in the data. Given the nature of the acquisition and subsurface, the diving waves in this study have limited penetration depths. Reflections penetrate deeper into the subsurface and can therefore be used to constrain deeper structure. The challenge with using reflections in land data is that they are typically masked by high amplitude ground roll. Fitting reflection data requires for the ground roll to be removed. One approach is to apply ground roll attenuation to both the data and the modelled synthetics at each FWI iteration prior to gradient computations. We explored this approach but found it to be impractical due to the differing ground roll signatures in the data and the synthetics. FK filtering is also prone to producing artefacts that could compromise the FWI updates.

To access the reflections, we implemented the modified free-surface boundary condition proposed by Plessix and Perez Solano (2015). The modified boundary condition alters the zero normal-stress condition, $\boldsymbol{\sigma} \cdot \hat{\mathbf{n}}|_{z=0} = 0$. In the conventional condition, $\sigma_{xz} = \sigma_{zz} = 0$, whereas in the updated condition the σ_{xz} term is replaced by its vertical derivative such that $\frac{\partial \sigma_{xz}}{\partial z} = 0$; the σ_{zz} term is not altered. The modification prevents ground roll from being

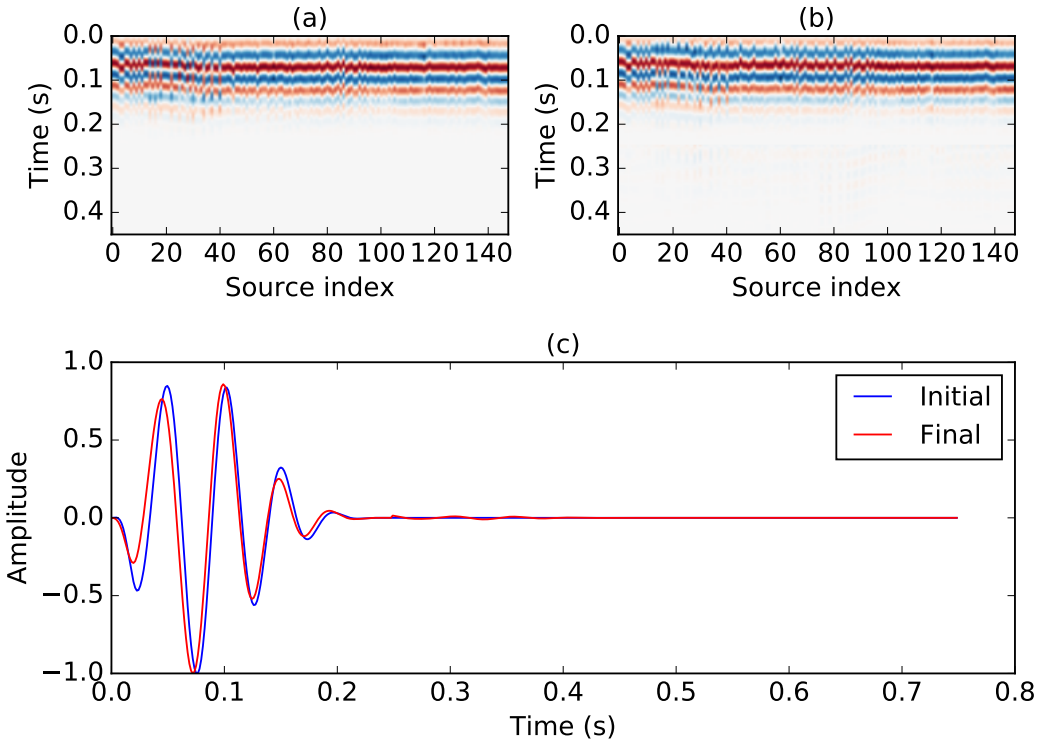

Figure 6.31: Estimated source wavelet (a) before and (b) after elastic FWI. (c) Comparison of average wavelet before and after inversion.

generated in the synthetic data, albeit with some limitations. The reflection coefficient of PP waves at the free surface has a fixed value of -1 and is no longer angle dependent. Similarly, the SS reflection coefficient has a value of 1 and is most often in the opposite direction to that produced with a true free surface. Plessix and Perez Solano (2015) state that the approximation is best suited for small angles and when near-surface v_p/v_s ratios are above 2. The error in SS reflection coefficients is justified by the fact that we focus on fitting P-wave data on the vertical component.

For the inversion, offsets above 2.25 km are muted and ground roll in the data is removed prior to inversion via FK filtering. The inversion domain is extended to 2 km depth. The initial model is updated in the upper 750 m with the near-surface inversion results from elastic FWI. Inversion is performed over 4-12, 4-16 and 4-20 Hz frequency bands. Low SNR means that the reflection events are not distinguishable at lower frequency bands. Each frequency stage performs 20 preconditioned NLCG iterations. Since reflections contribute to higher-wavenumber features in the gradient, we use larger Sobolev scale length in the preconditioner to promote longer-wavelength updates. A summary of the inversion parameters

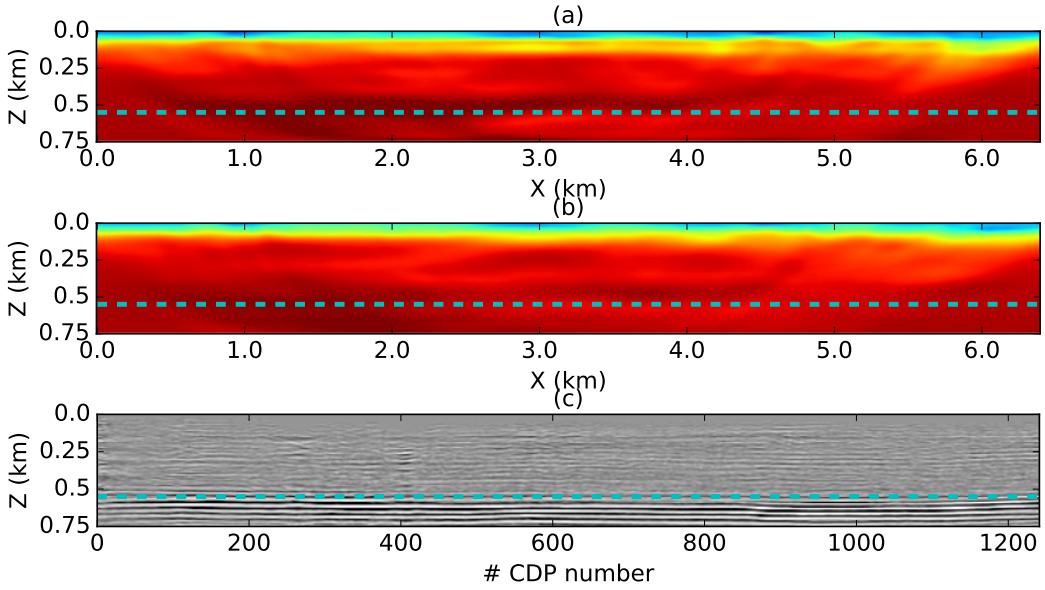

Figure 6.32: Comparison of inverted v_p models with a depth-converted PSTM image. The final FWI models from (a) acoustic and (b) elastic inversion exhibit a high-velocity perturbation at approximately 0.5 km depth (cyan line). The structure appears to coincide with a strong reflector apparent in the depth-converted PSTM image.

Sequence	**Frequency band (Hz)**	**Damping (s)**	**Anisotropic SSP** $(\boldsymbol{\mu_x}, \boldsymbol{\mu_z})$
I	4-12	-	5, 0.75
II	4-16	-	5, 0.25
III	4-20	-	3, 0.1

Table 6.3: Parameter choices for multi-scale elastic FWI including reflections.

is presented in Table 6.3.

The initial and final inverted v_p and v_s models are displayed in Figures 6.33 and 6.34, respectively. The inclusion of the reflection data into the inversion has improved the lateral continuity of the high-velocity layer at 500 m depth in the v_p model. Additional layers become apparent below the high-velocity perturbation. A secondary high velocity perturbation is observed at 800 m depth; however, it appears that the inversion was not able to construct its assumed true lateral extent. No significant updates appear below 1 km depth. The v_s model again shows very limited updates as expected. The inversion does appear to

further decrease the S-wave velocities in the upper 250 m. A comparison of data and final synthetics is displayed in Figure 6.35. After inversion, some weak reflections events are now apparent in the near offsets up to 1 s; within this range the data and synthetics are well aligned.

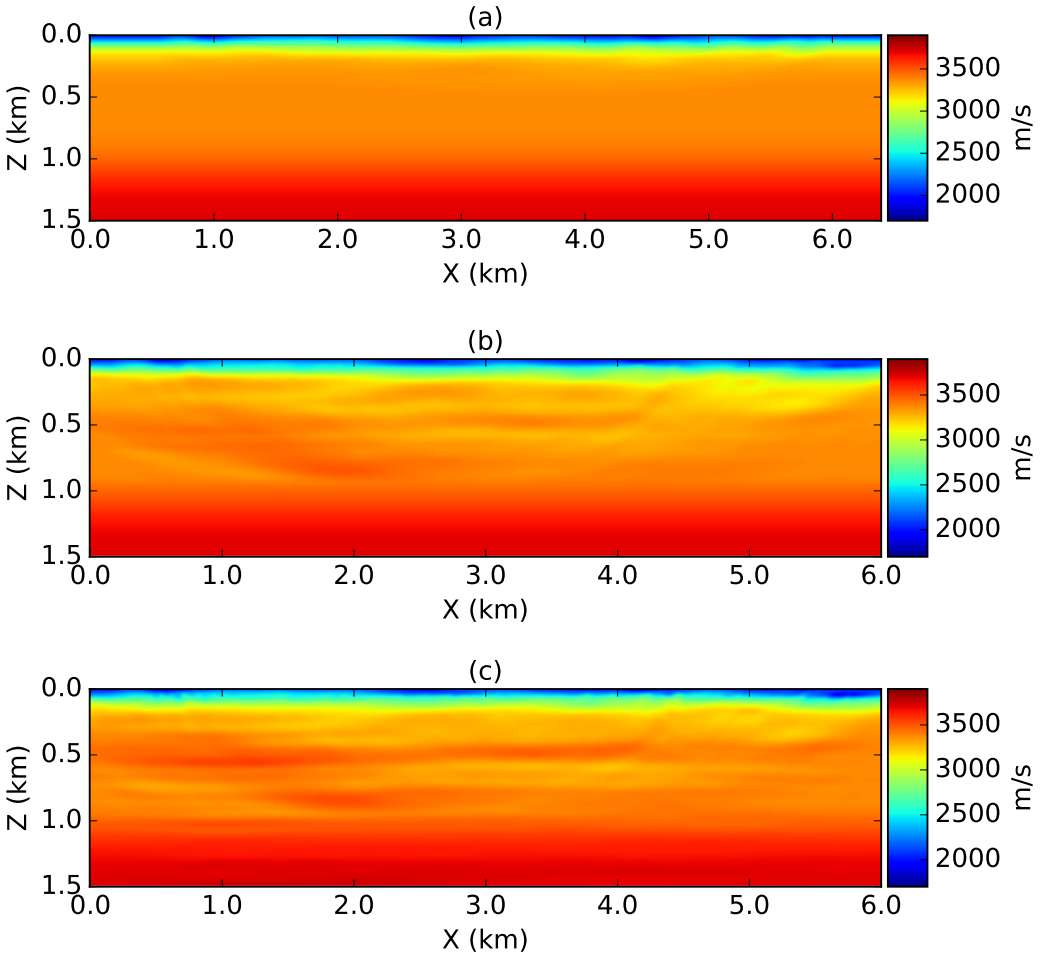

Figure 6.33: Estimated P-wave velocity models. (a) Initial model after tomography and well analysis. (b) Model after near-surface elastic FWI. (c) Model after elastic inversion using reflections and a modified free-surface boundary condition. Incorporating reflections has helped to improve the lateral continuity of the apparent reflectors.

6.7 Validation

For the final validation of the inverted model, we compare the inversion results to sonic log data and investigate the impact of the updated velocities on their respective RTM images. Figure 6.36 plots two sonic longs close to the survey line. Since the initial model was partially derived from a smoothed sonic log, the agreement is already good before inversion. In the first log, some velocity perturbations have been added to the initial model. It is unclear whether the inverted model can be considered an improved fit to the sonic log. The second log is positioned closer to the edge of the geometry where acquisition gaps and reduced illumination result in negligible updates.

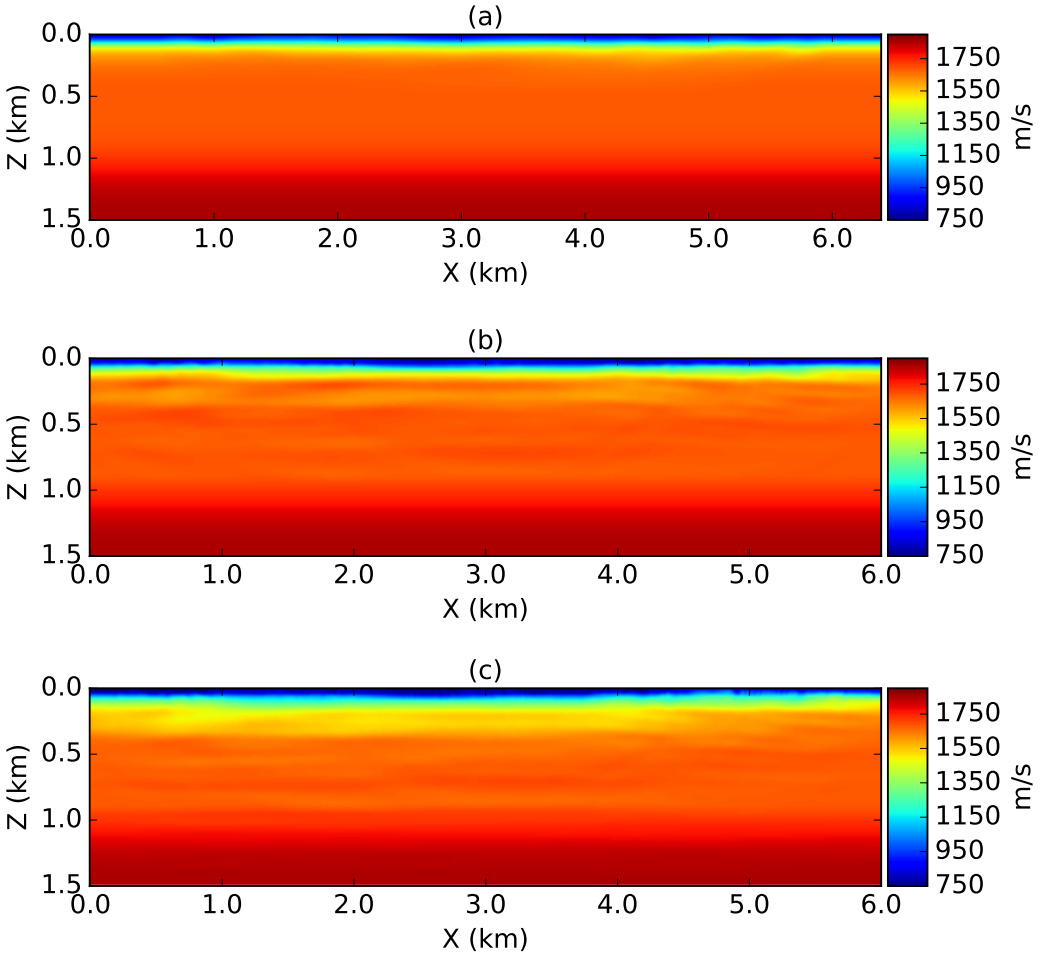

Figure 6.34: Estimated S-wave velocity models. (a) Initial model after tomography and well analysis. (b) Model after near-surface elastic FWI. (c) Model after elastic inversion using reflections and a modified free-surface boundary condition. Incorporating reflections has further reduced shallow S-wave velocities.

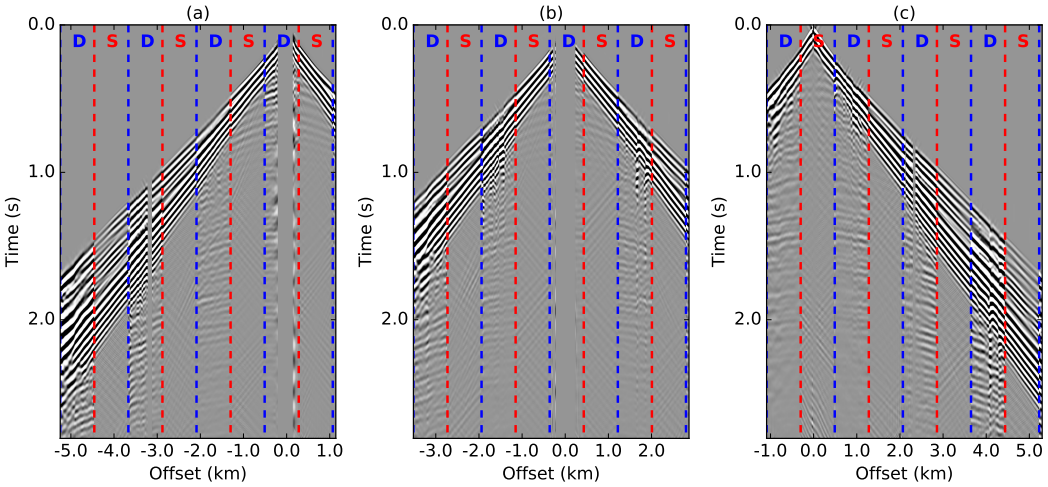

Figure 6.35: Data comparison for elastic observations and synthetics generated in the final elastic FWI model. (a) Shot #10 ($x = 1.1$ km), (b) Shot #94 ($x = 2.86$ km) and (c) Shot #100 ($x = 5.3$ km). Some of the early arriving reflection events, visible at near offsets, now appear in the synthetic data and demonstrate good agreement to the data.

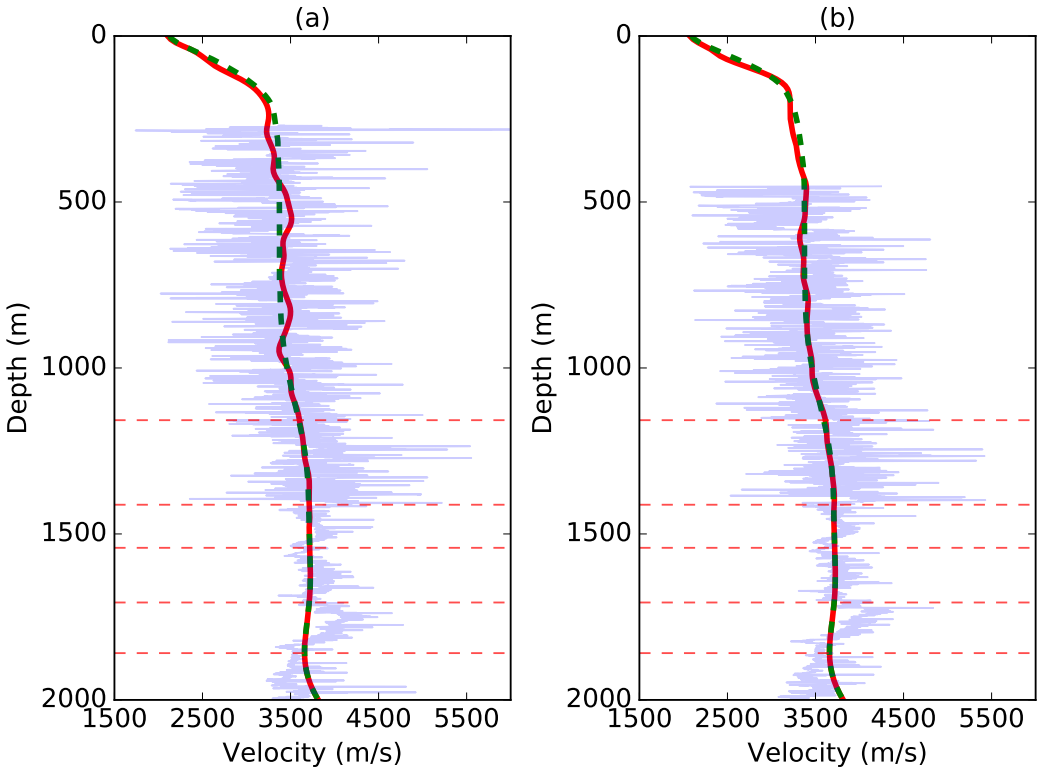

Figure 6.36: Sonic log validation. (a) Well at $x = 2.5$ km. (b) Well at $x = 5.2$ km. The sonic logs and initial model are depicted as pale blue and dashed green lines, respectively. The inverted P-wave velocities appear as bold red lines. Horizontal dashed red lines mark formation tops picked from the log data. FWI has added small perturbations to the background model. No discernible updates are achieved below 1 km depth.

The primary use of high-resolution FWI models is to facilitate improved depth imaging. We compute 25 Hz RTM images for 4 different velocity models: the interval velocity, the initial model, the model after elastic near-surface inversion and the final inversion model after elastic inversion including reflections. A comparison of the various RTM results is displayed in Figure 6.37. To assess the correctness of the seismic images, the positions of various formation tops are also plotted. The RTM image computed with the interval velocity (Figure 6.37) is clearly inadequate. Deeper reflectors show prominent undulations and shallow reflectors are disrupted and defocused. The image is noticeably improved with the initial model. Shallow reflectors exhibit improved lateral coherency and the flatness of the deeper reflectors has improved. The formation tops now are now aligned with prominent reflectors in the image. After near-surface FWI, the flatness of the reflectors between $x = 5 - 6$ km is slightly improved. We observed no noticeable difference between the RTM images produced

with the acoustic and elastic near-surface v_p models. The final RTM, computed with the reflection-based inverted model, shifts the reflector positioning slightly but shows no distinct improvement in the image. The velocity models used to produce the RTM images in Figure 6.37 are kinematically very similar, hence the lack of change in the various images.

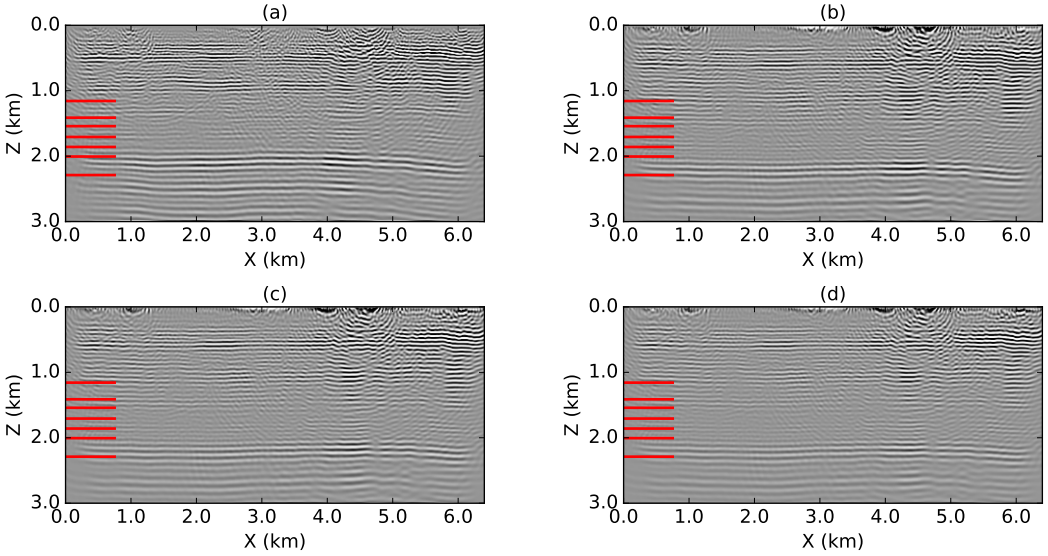

Figure 6.37: Acoustic RTM images computed using the (a) interval, (b) initial, (c) inverted elastic (near-surface) and (d) inverted elastic (w/ reflections) P-wave velocity models. The red ticks mark geological formation tops. The reflectors in the image exhibit poor alignment with the formation tops in the interval velocities. Significant undulations exist in the layers, further confirming the inaccuracy of the velocity model. The initial model significantly improves the flatness of the reflectors with their positioning with respect to the formation boundaries. After near-surface FWI, some uplift is observed in the deep reflectors improving their flatness. The update after including reflections has negligible impact on the RTM image.

Before concluding, we discuss some potential concerns that were not addressed by the proposed FWI workflow. Topography was neglected in this study due to the small variations in elevation across the survey line. For more pronounced topography, neglecting it can introduce time shifts (in synthetic data) due to mispositioned sources and receivers. Such time shifts could result in erroneous velocity structure being introduced into the model. The more egregious omissions in this study are those of anisotropy and attenuation. The geological evidence suggests that both factors could be prevalent in the survey area. By assuming an isotropic model, our inversion may under/overestimate velocity perturbations, as well as introduce erroneous features as a consequence of attenuation/anisotropy parameters mapping into the velocity model. Velocity errors can arise because the inversion is forced to fit the data by modifying only isotropic acoustic or elastic parameters. In reality,

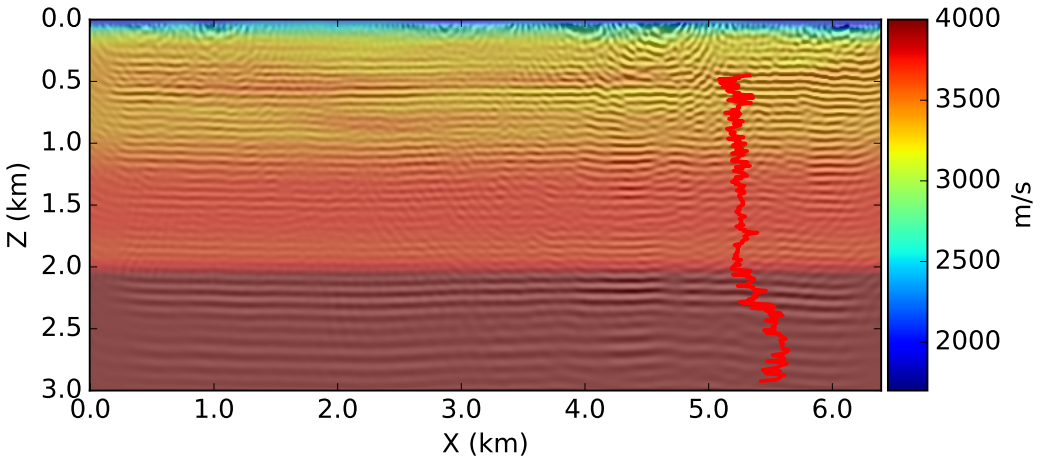

Figure 6.38: Final inverted P-wave velocity model after elastic inversion with an RTM overlay.

features of the data may be associated with changes in anisotropy/attenuation. Without further exploration, it is difficult to assess if, or to what extent, we have introduced errors by omitting attenuation and anisotropy; future studies in the area should consider including them.

The amplitude corrections we have adopted exhibit some redundancy/competing effects. For example, source normalization largely equalizes amplitude variations arising from variable source signatures. As such, the source correction term in the surface-consistent amplitude correction has limited influence. Similarly, a secondary geometrical spreading type correction arises from the amplitude matching of data and synthetics in the surface-consistent amplitude correction step. This effect occurs because amplitude differences (at mid-wide offsets), between data and synthetics, persist after the 3D-to-2D correction. Ultimately, our emphasis on a phase-based inversion of velocity parameters reduces the importance of amplitude information in the inversion. If accurate amplitude information is necessary, an inversion workflow should reconcile issues pertaining to geometrical spreading, along with inconsistent source and receiver signatures. Geometrical spreading concerns can be alleviated with 2.5D or 3D modelling; however, this will increase the computational cost. More research is required to determine a proper approach to dealing with variable source and receiver signatures on land data. Signature in this context refers to amplitude and phase variations that occur between adjacent/nearby receivers. Variable signatures may be linked to instrument variation, but can also be due to complex near-surface heterogeneities. Future research, should explore the potential link between near-surface structure and source/receiver signatures and whether the signatures can be inverted for.

6.8 Conclusions

We have presented applications of acoustic and elastic FWI on a 2D land dataset from the WCSB. Initial velocity models are developed through a combination of first-break traveltime tomography and structural constraints from well log data. An initial interpretation of the subsurface is obtained by combining geological knowledge of the study area with PSTM images which indicate a finely layered flat subsurface. Interleaved shale and sandstone layers at approximately 2 km depth are the primary sources of hydrocarbons in the region and potential target areas. Prior to FWI, the quality of the low frequencies in the data are assessed. Poor SNR observed at frequency bands below 4 Hz prompts us to only consider frequencies above 4 Hz. In preprocessing, the data are prepared differently for acoustic and elastic inversions. For both cases, the data undergo bandpass filtering, resampling, trace editing, 3D-to-2D amplitude corrections and surface consistent amplitude corrections. FK filtering is also included in the acoustic inversion to remove ground roll from the data.

We perform multi-scale FWI using the global correlation objective function to emphasize the fitting of phase information in the data. A combination of time windowing and gradient preconditioning ensures proper convergence of the inversion. For the acoustic inversion only P-wave velocities are updated whereas both P- and S-wave velocities are updated during elastic inversion. We observe reasonable updates down to depths of 750 m. FWI introduces a high-velocity v_p perturbation in both acoustic and elastic inversions. The elastic FWI result demonstrates improved structural coherence and remove a shallow channel artefact that is present in the acoustic result. Updates to v_s are limited, but generally demonstrate a decrease in shallow velocities relative to the starting model. After inversion, the data fit is improved in both test cases.

To extend the penetration depth of FWI below that of the diving waves, we include reflections into the inversion. This is done in the elastic case by implementing a modified free surface boundary condition in the waveform simulations. The modified boundary avoids generating surface waves, thus makes reflections visible. Ground roll is removed from the observed data prior to inversion. By including reflections, we improve the lateral continuity of the reflector introduced by near-surface FWI. Some additional reflectors also become more apparent down to depths of 1 km.

The deficiencies in the data (lack of wide offsets and low frequencies) limit the utility of this dataset for FWI. While FWI updates are able to provide increased resolution in the velocity model, the updates produced only small improvements in the RTM images. This is because the initial model already captured the correct kinematics of the data. Pushing FWI to high resolutions in this study results in velocity updates that enter the scale lengths of migration. High-resolution FWI is potentially useful if it allows direct interpretation;

however, this often relies on having knowledge on multiple parameters. Although the v_s model is updated in reasonable directions, we are unable to achieve similar high-resolution. Furthermore, the limited penetration depth prevents us from updating the model in the target area. Complex near surface structure remains challenging to fit with FWI due to the influence of multiple parameters in this region.

CHAPTER 7

Conclusions

7.1 Summary

Full waveform inversion is a powerful inversion technique used to estimate subsurface material parameters from seismic data. Increased accessibility and capabilities of high-performance computing systems, coupled with continued algorithmic advancements have led to the adoption of FWI for large-scale 3D inversion in global and exploration seismology. FWI is routinely applied in modern seismic processing workflows for oil and gas exploration in complex subsurface environments. In spite of recent successes, FWI remains troubled by fundamental problems stemming from the ill-posed nature of the inverse problem. The non-linearity of the objective function mandates a good initial velocity model to ensure proper convergence. The sensitivity and computational cost of the algorithm mean that applications on real data typically require careful monitoring and extensive quality control checks.

In chapter 2, we review the formulation of FWI as a PDE-constrained optimization problem. Some details related to the adjoint-state method and a generic FWI algorithm are included. We also establish and demonstrate some of the challenges of multi-parameter FWI.

In chapter 3, we explore the extension of source-encoding methods to multi-parameter FWI. Source encoding encompasses a range of methods that use simultaneous sources to reduce the number of PDE solves required per FWI iteration. In the presence of source-encoding, the FWI gradient exhibits cross-talk noise associated with interactions between wavefields from different sources. Cross-talk noise also appears in the source-encoded Hessian. We demonstrate that cross-talk noise (in the Hessian) can be attenuated randomizing the source encoding at each iteration. By suppressing cross-talk noise, similar parameter trade-off behaviour is observed for conventional and source-encoded FWI. A shortcoming of source

encoding methods is presented in the case of a data-driven inversion. We demonstrate that the inability to perform time/offset windowing to individual simulated shots, when using encoding, compromises the inversion leading to unsatisfactory inversion results for a particular test case.

In chapter 4, we propose a subsampled truncated Newton method for FWI. The method uses second-order stochastic optimization to reduce the computational cost of second-order optimization methods. Source subsampling is applied during the computation of Hessian-vector products to reduce the number of PDE solves required for the inner CG iterations. The STN approach demonstrates convergence rates and resolving power comparable to conventional truncated Newton methods, while incurring a computational cost closer to first-order gradient methods. A non-uniform sampling scheme is also proposed to identify sources that have a greater contribution to the summed Hessian-vector products. We perform numerical tests for two synthetic models and demonstrate that non-uniform sampling is beneficial when illumination in the subsurface is highly irregular. In the examples, irregular illumination is caused by lateral heterogeneities and structure that causes strong scattering in isolated regions of the model.

Chapter 5 presents a local resolution analysis and a linearized Bayesian inversion for acoustic FWI. Both analyses are supported by approximating the Hessian as a superposition of Kronecker products. The factorization allows for Hessian-vector products to be computed efficiently through a series of matrix multiplications involving (relatively) small matrices. The Kronecker factors are interpreted as vertical and horizontal point-spread functions. In the local resolution analysis, we apply the Hessian to spike perturbations at various points in the subsurface to extract horizontal and vertical resolution lengths. A linearized Bayesian inversion is performed using a structurally-informed model prior. The prior model distribution and data likelihood terms are assumed to be Gaussian. We use a low-rank approximation of the prior-preconditioned Hessian to enable efficient sampling of the posterior distribution. Samples estimated from the posterior distribution are used to compute standard deviations and 95% confidence intervals on the inverted v_p model. Through the analyses, we observe that resolution degrades towards the boundaries of the acquisition/model where illumination is limited.

Chapter 6 presents applications of acoustic and elastic FWI on a 2D land dataset from the WCSB. We devise a workflow that covers initial model building, data processing, and multi-scale inversion. A near-surface P-wave velocity model (to 750 m depth) is obtained from traveltime tomography. Deeper regions of the initial v_p model are inferred from sonic log data; log data are also used to inform the initial v_s model. A preliminary analysis of the subsurface, using PSTM and knowledge of geological formations, identifies a finely layered subsurface composed of interleaving shale and sandstone layers. We apply minimal

data processing for acoustic and elastic inversion. The processing consists of bandpass filtering, resampling, trace editing, 3D-to-2D amplitude corrections, and surface-consistent amplitude corrections. For the acoustic inversion, an additional ground-roll attenuation step is performed via FK filtering. During FWI, we apply a multi-scale approach and use the global correlation objective function to emphasize the fitting of phase information in the data. Time windowing and gradient preconditioning assist with proper convergence of the inversion. Acoustic and elastic inversion yield reasonable updates to P-wave velocity to depths of 750 m. Both inversions introduce a high-velocity v_p perturbation. Elastic FWI result demonstrates improved structural coherence and removes a shallow channel artefact apparent in the acoustic result. Updates to v_s are limited, but generally demonstrate a decrease in shallow velocities relative to the starting model. After inversion, the data fit is improved in both test cases. In the elastic inversion, reflections are included by using a modified boundary condition. The inclusion of reflections improves the lateral continuity of the reflector introduced by near-surface FWI. Some additional reflectors also become more apparent down to depths of 1 km.

In this final section, I offer some of my personal curiosities for future FWI research. Numerous alternative objective functions (e.g., traveltime-based, matching-filter objectives, optimal transport methods etc.) have been proposed to reduce the non-convexity of FWI and improve the robustness of convergence. At a basic level, these objectives succeed by simplifying the representation of the data. Comparing the similarities and shortcomings of these various algorithms could be useful in identifying more fundamental factors controlling convergence behaviour. Extended inversion algorithms have demonstrated impressive resilience to poor initial models in FWI; however, most current algorithms are too computationally expensive to be practical. An interesting parallel exists between deep neural networks and extended modelling algorithms. In both cases, models are over-parametrized to the extent that it becomes trivial, or easy, to fit the data. Models can be subsequently simplified to something more reasonable. In deep learning, this over-parametrization seems important to navigating the non-convex optimization landscape. It remains to be seen whether concepts from either discipline can assist the other.

Deep learning has emerged as a powerful tool for computer vision and natural language processing. While some efforts in translating deep learning algorithms to exploration seismology have been made, transferability of trained networks over a wide range of datasets has not been demonstrated. An interesting avenue of research could explore deep learning as a means to augment full waveform inversion/seismic imaging. One might envision training a network that acts as a projection-like operator, projecting model updates onto a learned set of feasible models. Generative models that estimate samples of a learned probability distribution may be useful for exploring prior, or true, model distributions. The challenge in these scenarios is producing suitable datasets that are diverse enough to be representative

of true Earth models.

My exposure to real data FWI has made it abundantly clear that the data dictate what FWI can accomplish. As a response to this, I believe we should establish better constraints on what subsurface properties the data can realistically estimate. Formally, this involves characterizing the null space in FWI. Proper diagnosis of the null space should allow us to identify which parameters our data are sensitive to, whether there are deficiencies in the data that could be targeted, and whether we are overextending the "reach" of our data. Ideally, such analyses would be done inexpensively prior to inversion to help inform the choice of physics model and FWI workflow. The ability to quantify the null space, and uncertainties, will be necessary to provide rigorous validation of multi-parameter models. If FWI is ever to be used for quantitative interpretation, such analyses are essential. Finally, during my land FWI application, we explored the inclusion of surface-consistent amplitude scalars. A more comprehensive approach might utilize surface-consistent deconvolution operators to account for phase and amplitude variations between source and receiver signatures. Surface-consistent deconvolution operators can also capture complexities in the data associated with complex near-surface structure. How such operators interact with high-resolution, near-surface FWI remains to be seen.

Bibliography

Akcelik, V., G. Biros, and O. Ghattas, 2002, Parallel Multiscale Gauss-Newton-Krylov Methods for Inverse Wave Propagation: SC '02: Proceedings of the 2002 ACM/IEEE Conference on Supercomputing, 41–41.

Aki, K., and P. G. Richards, 2002, Quantitative seismology: University Science Books. Geology (University Science Books): Seismology.

Alkhalifah, T., and R.-E. Plessix, 2014, A recipe for practical full-waveform inversion in anisotropic media: An analytical parameter resolution study: Geophysics, **79**, R91–R101.

Aminzadeh, F., J. Brac, and T. Kunz, 1997, SEG/EAGE 3-D salt and overthrust models: SEG/EAGE 3-D Modeling Series No. 1, SEG.

Anagaw, A. Y., and M. D. Sacchi, 2014, Comparison of multifrequency selection strategies for simultaneous-source full-waveform inversion: Geophysics, **79**, R165–R181.

Backus, G., and F. Gilbert, 1968, The Resolving Power of Gross Earth Data: Geophysical Journal International, **16**, 169–205.

Backus, G., F. Gilbert, and E. C. Bullard, 1970, Uniqueness in the inversion of inaccurate gross earth data: Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences, **266**, 123–192.

Backus, G. E., 1962, Long-wave elastic anisotropy produced by horizontal layering: Journal of Geophysical Research (1896-1977), **67**, 4427–4440.

Ben-Hadj-Ali, H., S. Operto, and J. Virieux, 2011, An efficient frequency-domain full waveform inversion method using simultaneous encoded sources: Geophysics, **76**, R109–R124.

Beylkin, G., and R. Burridge, 1990, Linearized inverse scattering problems in acoustics and elasticity: Wave Motion, **12**, 15 – 52.

Biondi, B., and A. Almomin, 2014, Simultaneous inversion of full data bandwidth by tomographic full-waveform inversion: Geophysics, **79**, WA129–WA140.

Bollapragada, R., R. Byrd, and J. Nocedal, 2016, Exact and Inexact Subsampled Newton Methods for Optimization: arXiv e-prints, arXiv:1609.08502.

Bosch, M., P. Barton, S. C. Singh, and I. Trinks, 2005, Inversion of traveltime data under a statistical model for seismic velocities and layer interfaces: Geophysics, **70**, R33–R43.

Bottou, L., 2010, Large-Scale Machine Learning with Stochastic Gradient Descent: Proceedings of COMPSTAT'2010, Physica-Verlag HD, 177–186.

Bottou, L., F. Curtis, and J. Nocedal, 2018, Optimization methods for large-scale machine learning: SIAM Review, **60**, 223–311.

Bozdag, E., J. Trampert, and J. Tromp, 2011, Misfit functions for full waveform inversion based on instantaneous phase and envelope measurements: Geophysical Journal International, **185**, 845–870.

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.

Brossier, R., S. Operto, and J. Virieux, 2009, Seismic imaging of complex onshore structures by 2D elastic frequency-domain full-waveform inversion: Geophysics, **74**, WC105–WC118.

———, 2010, Which data residual norm for robust elastic frequency-domain full waveform inversion?: Geophysics, **75**, R37–R46.

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.

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

Byrd, R., G. Chin, W. Neveitt, and N. Jorge, 2011, On the use of stochastic Hessian information in optimization methods for machine learning: SIAM Journal on Optimization, **21**, 977–995.

Byrd, R. H., G. M. Chin, J. Nocedal, and Y. Wu, 2012, Sample size selection in optimization methods for machine learning: Mathematical Programming, **134**, 127–155.

Capdeville, Y., Y. Gung, and B. Romanowicz, 2005, Towards global earth tomography using the spectral element method: a technique based on source stacking: Geophysical Journal International, **162**, 541–554.

Capdeville, Y., and L. Mtivier, 2018, Elastic full waveform inversion based on the homogenization method: theoretical framework and 2-D numerical illustrations: Geophysical Journal International, **213**, 1093–1112.

Cary, P. W., and G. A. Lorentz, 1993, Four component surface consistent deconvolution: Geophysics, **58**, 383–392.

Castellanos, C., L. Métivier, S. Operto, R. Brossier, and J. Virieux, 2015, Fast full waveform inversion with source encoding and second-order optimization methods: Geophysical Journal International, **200**, 718–742.

Chavent, G., F. Clément, and S. Gómez, 1994, *in* Automatic determination of velocities via migration-based traveltime waveform inversion: A synthetic data example: 1179–1182.

Chen, P., L. Zhao, and T. Jordan, 2007, Full 3D Tomography for Crustal Structure of the Los Angeles Region: Bulletin of the Seismological Society of America, **97**, 1094–1120.

Cheng, X., K. Jiao, D. Sun, Z. Xu, D. Vigh, and A. El-Emam, 2017, High-resolution radon preconditioning for full-waveform inversion of land seismic data: Interpretation, **5**, SR23–SR33.

Choi, Y., and T. Alkhalifah, 2012, Application of multi-source waveform inversion to marine streamer data using the global correlation norm: Geophysical Prospecting, **60**, 748–758.

Claerbout, J. F., 1971, Toward a unified theory of reflector mapping: Geophysics, **36**, 467–481.

Crase, E., A. Pica, M. Noble, J. McDonald, and A. Tarantola, 1990, Robust elastic nonlinear waveform inversion: Application to real data: Geophysics, **55**, 527–538.

Crase, E., C. Wideman, M. Noble, and A. Tarantola, 1992, Nonlinear elastic waveform inversion of land seismic reflection data: Journal of Geophysical Research: Solid Earth, **97**, 4685–4703.

Dahlen, F. A., S.-H. Hung, and G. Nolet, 2000, Frchet kernels for finite-frequency traveltimesi. theory: Geophysical Journal International, **141**, 157–174.

Dai, W., P. Fowler, and G. T. Schuster, 2012, Multi-source least-squares reverse time migration: Geophysical Prospecting, **60**, 681–695.

Dix, C. H., 1955, Seismic velocities from surface measurements: Geophysics, **20**, 68–86.

Drineas, P., R. Kannan, and M. W. Mahoney, 2006, Fast Monte Carlo Algorithms for Matrices I: Approximating Matrix Multiplication: SIAM Journal on Computing, **36**, 132–157.

E., P., and G. Ribiere, 1969, Note sur la convergence de méthodes de directions conjuguées: ESAIM: Mathematical Modelling and Numerical Analysis - Modélisation Mathématique et Analyse Numérique, **3**, 35–43.

Eisenstat, S. C., and H. F. Walker, 1996, Choosing the Forcing Terms in an Inexact Newton Method: SIAM Journal on Scientific Computing, **17**, 16–32.

Ely, G., A. Malcolm, and O. V. Poliannikov, 2018, Assessing uncertainties in velocity models and images with a fast nonlinear uncertainty quantification method: Geophysics, **83**, R63–R75.

Epanomeritakis, I., V. Akelik, O. Ghattas, and J. Bielak, 2008, A Newton-CG method for large-scale three-dimensional elastic full-waveform seismic inversion: Inverse Problems, **24**, 034015.

Erdogdu, M. A., and A. Montanari, 2015, Convergence rates of sub-sampled Newton methods: arXiv e-prints, arXiv:1508.02810.

Etgen, J., and C. Regone, 2005, Strike shooting, dip shooting, widepatch shooting Does prestack depth migration care? A model study: SEG Technical Program Expanded Abstracts 1998, 66–69.

Fang, Z., C. D. Silva, R. Kuske, and F. J. Herrmann, 2018, Uncertainty quantification for inverse problems with weak partial-differential-equation constraints: Geophysics, **83**, R629–R647.

Feng, Z., B. Guo, and G. T. Schuster, 2018, Multiparameter deblurring filter and its application to elastic migration and inversion: Geophysics, **0**, 1–49.

Fichtner, A., 2010, Full Seismic Waveform Modelling and Inversion: Springer Verlag.

Fichtner, A., H.-P. Bunge, and H. Igel, 2006, The adjoint method in seismology: Physics of the Earth and Planetary Interiors, **157**, 86 – 104.

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

Fichtner, A., and T. v. Leeuwen, 2015, Resolution analysis by random probing: Journal of Geophysical Research: Solid Earth, **120**, 5549–5573. (2015JB012106).

Fichtner, A., and J. Trampert, 2011a, Hessian kernels of seismic data functionals based upon adjoint techniques: Geophysical Journal International, **185**, 775–798.

———, 2011b, Resolution analysis in full waveform inversion: Geophysical Journal International, **187**, 1604–1624.

Forgues, E., and G. Lambaré, 1997, Parameterization study for acoustic and elastic ray+born inversion: Journal of seismic exploration, **6**, 253–277.

French, S., V. Lekic, and B. Romanowicz, 2013, Waveform Tomography Reveals Channeled Flow at the Base of the Oceanic Asthenosphere: Science, **342**, 227–230.

Friedlander, M., and M. Schmidt, 2012, Hybrid deterministic-stochastic methods for data fitting: SIAM Journal on Scientific Computing, **34**, A1380–A1405.

Gallagher, K., M. Sambridge, and G. Drijkoningen, 1991, Genetic algorithms: An evolution from monte carlo methods for strongly non-linear geophysical optimization problems: Geophysical Research Letters, **18**, 2177–2180.

Gao, W., G. Matharu, and M. D. Sacchi, 2020, Fast least-squares reverse time migration via a superposition of kronecker products: Geophysics, **85**, S115–S134.

Gardner, G. H. F., L. W. Gardner, and A. R. Gregory, 1974, Formation velocity and density - the diagnostic basics for stratigraphic traps: Geophysics, **39**, 770–780.

Gauthier, O., J. Virieux, and A. Tarantola, 1986, Twodimensional nonlinear inversion of seismic waveforms: Numerical results: Geophysics, **51**, 1387–1403.

Gholami, Y., R. Brossier, S. Operto, A. Ribodetti, and J. Virieux, 2013, Which parameterization is suitable for acoustic vertical transverse isotropic full waveform inversion? Part 1: Sensitivity and trade-off analysis: Geophysics, **78**, R81–R105.

Gomes, A., and N. Chazalnoel, 2017, *in* Extending the reach of full-waveform inversion with reflection data: Potential and challenges: 1454–1459.

Guitton, A., G. Ayeni, and E. Diaz, 2012, Constrained full-waveform inversion by model

reparameterization: Geophysics, **77**, R117–R127.

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.

Hale, D., 2014, Implementing an anisotropic and spatially varying matérn model covariance with smoothing filters: Presented at the .

Hansen, P., 1998, Rank-deficient and discrete ill-posed problems: Society for Industrial and Applied Mathematics.

Huang, Y., and G. T. Schuster, 2012, Multisource least-squares migration of marine streamer and land data with frequency-division encoding: Geophysical Prospecting, **60**, 663–680.

Hutchinson, M., 1990, A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines: Communications in Statistics - Simulation and Computation, **19**, 433–450.

Ibrahim, A., and M. D. Sacchi, 2014, Simultaneous source separation using a robust radon transform: Geophysics, **79**, V1–V11.

Innanen, K. A., 2014, Seismic avo and the inverse hessian in precritical reflection full waveform inversion: Geophysical Journal International, **199**, 717–734.

Ji, Y., S. C. Singh, and B. E. Hornby, 2000, Sensitivity study using a genetic algorithm: inversion of amplitude variations with slowness: Geophysical Prospecting, **48**, 1053–1073.

Kamei, R., T. Miyoshi, R. G. Pratt, M. Takanashi, and S. Masaya, 2015, Application of waveform tomography to a crooked-line 2d land seismic data set: Geophysics, **80**, B115–B129.

Köhn, D., D. De Nil, A. Kurzmann, A. Przebindowska, and T. Bohlen, 2012, On the influence of model parametrization in elastic full waveform tomography: Geophysical Journal International, **191**, 325–345.

Komatitsch, D., and R. Martin, 2007, An unsplit convolutional perfectly matched layer improved at grazing incidence for the seismic wave equation: Geophysics, **72**, SM155–SM167.

Korta, N., A. Fichtner, and V. Sallarcs, 2013, Block-diagonal approximate hessian for preconditioning in full waveform inversion.

Krebs, J., C. Ober, T. Smith, J. Overfelt, S. Collis, G. V. Winckel, B. V. B. Waanders, N. Downey, and D. Aldridge, 2016, Synthetic study of raw-data FWI applied to visco-TTI-elastic data, *in* SEG Technical Program Expanded Abstracts 2016: Society of Exploration Geophysicists, 1179–1183.

Krebs, J. R., J. E. Anderson, D. Hinkley, R. Neelamani, S. Lee, A. Baumstein, and M.-D. Lacasse, 2009, Fast full-wavefield seismic inversion using encoded sources: Geophysics, **74**, WCC177–WCC188.

Kuhn, H. W., and A. W. Tucker, 1951, Nonlinear Programming: University of California

Press.

Lailly, P., 1983, The seismic inverse problem as a sequence of before stack migrations: Conference on Inverse Scattering-Theory and Application, Society for Industrial and Applied Mathematics, Expanded Abstracts, 206–220.

Levander, A. R., 1988, Fourth-order finite-difference P-SV seismograms: Geophysics, **53**, 1425–1436.

Li, Y., B. Biondi, R. Clapp, and D. Nichols, 2016, Integrated vti model building with seismic data, geologic information, and rock-physics modeling part 1: Theory and synthetic test: Geophysics, **81**, C177–C191.

Liu, D. C., and J. Nocedal, 1989, On the limited memory bfgs method for large scale optimization: Mathematical Programming, **45**, 503–528.

Luo, S., and P. Sava, 2011, *in* A deconvolution based objective function for wave equation inversion: 2788–2792.

Luo, Y., and G. T. Schuster, 1991, Wave-equation traveltime inversion: Geophysics, **56**, 645–653.

Marquardt, D., 1963, An algorithm for least-squares estimation of nonlinear parameters: Journal of the Society for Industrial and Applied Mathematics, **11**, 431–441.

Marquering, H., F. Dahlen, and G. Nolet, 1999, Three-dimensional sensitivity kernels for finite-frequency traveltimes: the banana-doughnut paradox: Geophysical Journal International, **137**, 805–815.

Martin, G. S., R. Wiley, and K. J. Marfurt, 2006, Marmousi2: An elastic upgrade for Marmousi: The Leading Edge, **25**, 156–166.

Matharu, G., and M. D. Sacchi, 2018, Source encoding in multiparameter full waveform inversion: Geophysical Journal International, **214**, 792–810.

Menke, W., 1984, Geophysical data analysis: discrete inverse theory: Academic Press.

Métivier, L., A. Allain, R. Brossier, Q. Marigot, E. Oudet, and J. Virieux, 2018, Optimal transport for mitigating cycle skipping in full-waveform inversion: A graph-space transform approach: Geophysics, **83**, R515–R540.

Métivier, L., R. Brossier, Q. Mérigot, E. Oudet, and J. Virieux, 2016, Measuring the misfit between seismograms using an optimal transport distance: application to full waveform inversion: Geophysical Journal International, **205**, 345–377.

Métivier, L., R. Brossier, S. Operto, and J. Virieux, 2015, Acoustic multi-parameter FWI for the reconstruction of P-wave velocity, density and attenuation: preconditioned truncated Newton approach: SEG Technical Program Expanded Abstracts 2015, 1198–1203.

Métivier, L., R. Brossier, J. Virieux, and S. Operto, 2013, Full Waveform Inversion and the Truncated Newton Method: SIAM Journal on Scientific Computing, **35**, B401–B437.

Modrak, R., and J. Tromp, 2016, Seismic waveform inversion best practices: regional, global and exploration test cases: Geophysical Journal International, **206**, 1864–1889.

Modrak, R. T., D. Borisov, M. Lefebvre, and J. Tromp, 2018, SeisFlows - Flexible waveform inversion software: Computers & Geosciences, **115**, 88 – 95.

Moghaddam, P. P., H. Keers, F. J. Herrmann, and W. A. Mulder, 2013, A new optimization approach for source-encoding full-waveform inversion: Geophysics, **78**, R125–R132.

Mora, P., 1987, Nonlinear twodimensional elastic inversion of multioffset seismic data: Geophysics, **52**, 1211–1228.

———, 1988, Elastic wavefield inversion of reflection and transmission data: Geophysics, **53**, 750–759.

———, 1989, Inversion = migration + tomography: Geophysics, **54**, 1575–1586.

Mosegaard, K., and A. Tarantola, 1995, Monte carlo sampling of solutions to inverse problems: Journal of Geophysical Research: Solid Earth, **100**, 12431–12447.

Mosegaard, K., and P. D. Vestergaard, 1991, A simulated annealing approach to seismic model optimization with sparse prior information1: Geophysical Prospecting, **39**, 599–611.

Mossop, G., and c. Shetsen, I., 1994, *in* Geological atlas of the Western Canada Sedimentary Basin: Canadian Society of Petroleum Geologists and Alberta Research Council.

Nocedal, J., and S. J. Wright, 2006, Numerical optimization, 2nd ed.: Springer. Springer Series in Operations Research and Financial Engineering.

Operto, S., Y. Gholami, V. Prieux, A. Ribodetti, R. Brossier, L. Métivier, and J. Virieux, 2013, A guided tour of multiparameter full-waveform inversion with multicomponent data: From theory to practice: The Leading Edge, **32**, 1040–1054.

Pan, W., Y. Geng, and K. A. Innanen, 2018, Interparameter trade-off quantification and reduction in isotropic-elastic full-waveform inversion: synthetic experiments and hussar land data set application: Geophysical Journal International, **213**, 1305–1333.

Pan, W., K. A. Innanen, G. F. Margrave, M. C. Fehler, X. Fang, and J. Li, 2016, Estimation of elastic constants for hti media using gauss-newton and full-newton multiparameter full-waveform inversion: Geophysics, **81**, R275–R291.

Pica, A., J. P. Diet, and A. Tarantola, 1990, Nonlinear inversion of seismic reflection data in a laterally invariant medium: Geophysics, **55**, 284–292.

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.

Plessix, R.-E., G. Baeten, J. W. de Maag, F. ten Kroode, and Z. Rujie, 2012, Full waveform inversion and distance separated simultaneous sweeping: a study with a land seismic data set: Geophysical Prospecting, **60**, 733–747.

Plessix, R.-E., and Q. Cao, 2011, A parametrization study for surface seismic full waveform inversion in an acoustic vertical transversely isotropic medium: Geophysical Journal International, **185**, 539–556.

Plessix, R.-E., and C. A. Perez Solano, 2015, Modified surface boundary conditions for elastic waveform inversion of low-frequency wide-angle active land seismic data: Geophysical Journal International, **201**, 1324–1334.

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.

Pratt, R. G., C. Shin, and G. J. Hick, 1998, Gauss-Newton and full Newton methods in frequency-space seismic waveform inversion: Geophysical Journal International, **133**, 341–362.

Pratt, R. G., Z.-M. Song, P. Williamson, and M. Warner, 1996, Two-dimensional velocity models from wide-angle seismic data by wavefield inversion: Geophysical Journal International, **124**, 323–340.

Pratt, R. G., and M. H. Worthington, 1990, Inverse theory applied to multi-source cross-hole tomography.: Geophysical Prospecting, **38**, 287–310.

Prieux, V., R. Brossier, S. Operto, and J. Virieux, 2013a, Multiparameter full waveform inversion of multicomponent ocean-bottom-cable data from the Valhall field. Part 1: imaging compressional wave speed, density and attenuation: Geophysical Journal International, **194**, 1640–1664.

———, 2013b, Multiparameter full waveform inversion of multicomponent ocean-bottom-cable data from the valhall field. part 2: imaging compressive-wave and shear-wave velocities: Geophysical Journal International, **194**, 1665–1681.

Raknes, E. B., B. Arntsen, and W. Weibull, 2015, Three-dimensional elastic full waveform inversion using seismic data from the Sleipner area: Geophysical Journal International, **202**, 1877–1894.

Ravaut, C., S. Operto, L. Improta, J. Virieux, A. Herrero, and P. Dell'Aversana, 2004, Multiscale imaging of complex structures from multifold wide-aperture seismic data by frequency-domain full-waveform tomography: application to a thrust belt: Geophysical Journal International, **159**, 1032–1056.

Rawlinson, N., A. Fichtner, M. Sambridge, and M. K. Young, 2014, Chapter one - seismic tomography and the assessment of uncertainty: Elsevier, volume **55** *of* Advances in Geophysics, 1 – 76.

Rawlinson, N., and W. Spakman, 2016, On the use of sensitivity tests in seismic tomography: Geophysical Journal International, **205**, 1221–1243.

Romero, L. A., D. C. Ghiglia, C. C. Ober, and S. A. Morton, 2000, Phase encoding of shot records in prestack migration: Geophysics, **65**, 426–436.

Roosta-Khorasani, F., and M. W. Mahoney, 2016a, Sub-Sampled Newton Methods I: Globally Convergent Algorithms: arXiv e-prints, arXiv:1601.04737.

———, 2016b, Sub-Sampled Newton Methods II: Local Convergence Rates: arXiv e-prints, arXiv:1601.04738.

Routh, P., J. Krebs, S. Lazaratos, A. Baumstein, S. Lee, Y. H. Cha, I. Chikichev, N. Downey, D. Hinkley, and J. Anderson, 2011, Encoded simultaneous source fullwavefield inversion for spectrally shaped marine streamer data: SEG Technical Program Expanded Abstracts 2011, 2433–2438.

Sambridge, M., 1999a, Geophysical inversion with a neighbourhood algorithms - I. Searching a parameter space: Geophysical Journal International, **138**, 479–494.

———, 1999b, Geophysical inversion with a neighbourhood algorithms-II. Appraising the ensemble: Geophysical Journal International, **138**, 727–746.

Sambridge, M., and G. Drijkoningen, 1992, Genetic algorithms in seismic waveform inversion: Geophysical Journal International, **109**, 323–342.

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

Schuster, G. T., X. Wang, Y. Huang, W. Dai, and C. Boonyasiriwat, 2011, Theory of multisource crosstalk reduction by phase-encoded statics: Geophysical Journal International, **184**, 1289–1303.

Sears, T., S. Singh, and P. Barton, 2008, Elastic full waveform inversion of multi-component OBC seismic data: Geophysical Prospecting, **56**, 843–862.

Sears, T. J., P. J. Barton, and S. C. Singh, 2010, Elastic full waveform inversion of multicomponent ocean-bottom cable seismic data: Application to alba field, u. k. north sea: Geophysics, **75**, R109–R119.

Sedova, A., G. T. Royle, T. Allemand, G. Lambaré, and O. Hermant, 2019, High-frequency acoustic land full-waveform inversion: a case study from the sultanate of oman: EAGE First break, 75–81.

Sen, M. K., and P. L. Stoffa, 1991, Nonlinear one-dimensional seismic waveform inversion using simulated annealing: Geophysics, **56**, 1624–1638.

Shin, C., S. Jang, and D.-J. Min, 2001, Improved amplitude preservation for prestack depth migration by inverse scattering theory: Geophysical Prospecting, **49**, 592–606.

Shin, C., and D.-J. Min, 2006, Waveform inversion using a logarithmic wavefield: Geophysics, **71**, R31–R42.

Shipp, R. M., and S. C. Singh, 2002, Two-dimensional full wavefield inversion of wide-aperture marine seismic streamer data: Geophysical Journal International, **151**, 325–344.

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

Solano, C. P., and R.-E. Plessix, 2019, Velocity-model building with enhanced shallow resolution using elastic waveform inversion: An example from onshore oman: Geophysics, **84**, R989–R1000.

Stoffa, P. L., and M. K. Sen, 1991, Nonlinear multiparameter optimization using genetic algorithms: Inversion of plane-wave seismograms: Geophysics, **56**, 1794–1810.

Stopin, A., R.-E. Plessix, and S. A. Abri, 2014, Multiparameter waveform inversion of a large wide-azimuth low-frequency land data set in oman: Geophysics, **79**, WA69–WA77.

Sun, B., and T. Alkhalifah, 2018, *in* The application of an optimal transport to a preconditioned data matching function for robust waveform inversion: 5168–5172.

Symes, W. W., 2007, Reverse time migration with optimal checkpointing: Geophysics, **72**, SM213–SM221.

———, 2008, Migration velocity analysis and waveform inversion: Geophysical Prospecting, **56**, 765–790.

Taner, M. T., and F. Koehler, 1981, Surface consistent corrections: Geophysics, **46**, 17–22.

Tang, Y., 2009, Target-oriented wave-equation least-squares migration/inversion with phase-encoded Hessian: Geophysics, **74**, WCA95–WCA107.

Tang, Y., and S. Lee, 2010, Preconditioning full waveform inversion with phase-encoded Hessian, *in* SEG Technical Program Expanded Abstracts 2010: Society of Exploration Geophysicists, 1034–1038.

———, 2015, Multi-parameter full wavefield inversion using non-stationary point-spread functions, *in* SEG Technical Program Expanded Abstracts 2015: Society of Exploration Geophysicists, 1111–1115.

Tape, C., Q. Liu, A. Maggi, and J. Tromp, 2009, Adjoint Tomography of the Southern California Crust: Science, **325**, 988–992.

Tarantola, A., 1984a, Inversion of seismic reflection data in the acoustic approximation: Geophysics, **49**, 1259–1266.

———, 1984b, Linearized inversion of seismic reflection data: Geophysical Prospecting, **32**, 998–1015.

———, 1986, A strategy for nonlinear elastic inversion of seismic reflection data: Geophysics, **51**, 1893–1903.

———, 2005, Inverse problem theory and methods for model parameter estimation: Society for Industrial and Applied Mathematics.

Thurin, J., R. Brossier, and L. Métivier, 2019, Ensemble-based uncertainty estimation in full waveform inversion: Geophysical Journal International, **219**, 1613–1635.

Trampert, J., A. Fichtner, and J. Ritsema, 2013, Resolution tests revisited: the power of random numbers: Geophysical Journal International, **192**, 676–680.

Trinh, P., R. Brossier, L. Lemaistre, L. Métivier, and J. Virieux, 2019, 3d elastic fwi with a non-linear model constraint: Application to a real complex onshore dataset: **2019**, 1–5.

Trinh, P., R. Brossier, L. Mtivier, J. Virieux, and P. Wellington, 2017, Bessel smoothing filter for spectral-element mesh: Geophysical Journal International, **209**, 1489–1512.

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

Valenciano, A. A., B. Biondi, and A. Guitton, 2006, Target-oriented wave-equation inversion:

Geophysics, **71**, A35–A38.

van Leeuwen, T., and F. J. Herrmann, 2013a, Fast waveform inversion without source-encoding: Geophysical Prospecting, **61**, 10–19.

———, 2013b, Mitigating local minima in full-waveform inversion by expanding the search space: Geophysical Journal International, **195**, 661–667.

Van Leeuwen, T., and W. A. Mulder, 2010, A correlation-based misfit criterion for wave-equation traveltime tomography: Geophysical Journal International, **182**, 1383–1394.

van Vossen, R., A. Curtis, and J. Trampert, 2006, Surface-consistent amplitude corrections for single or multicomponent sources and receivers using reciprocity and waveform inversion: Geophysical Journal International, **165**, 311–322.

Vigh, D., X. Cheng, K. Jiao, Z. Xu, and D. Sun, 2018, *in* Essential steps for successful full-waveform inversion using land data: 1128–1132.

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.

Vigh, D., and E. W. Starr, 2008, 3D prestack plane-wave, full-waveform inversion: Geophysics, **73**, VE135–VE144.

Virieux, J., 1986, P-SV wave propagation in heterogeneous media: Velocity-stress finite-difference method: Geophysics, **51**, 889–901.

Virieux, J., and S. Operto, 2009, An overview of full-waveform inversion in exploration geophysics: Geophysics, **74**, WCC1–WCC26.

Wang, Y., L. Dong, Y. Liu, and J. Yang, 2016, 2D frequency-domain elastic full-waveform inversion using the block-diagonal pseudo-Hessian approximation: Geophysics, **81**, R247–R259.

Warner, M., and L. Guasch, 2016, Adaptive waveform inversion: Theory: Geophysics, **81**, R429–R445.

Warner, M., A. Ratcliffe, T. Nangoo, J. Morgan, A. Umpleby, N. Shah, V. Vinje, I. Štekl, L. Guasch, C. Win, G. Conroy, and A. Bertrand, 2013, Anisotropic 3d full-waveform inversion: Geophysics, **78**, R59–R80.

Wu, R., and K. Aki, 1985, Scattering characteristics of elastic waves by an elastic heterogeneity: Geophysics, **50**, 582–595.

Xu, P., F. Roosta-Khorasani, and M. W. Mahoney, 2017, Newton-Type Methods for Non-Convex Optimization Under Inexact Hessian Information: arXiv e-prints, arXiv:1708.07164.

Xu, P., J. Yang, F. Roosta-Khorasani, C. Ré, and M. W. Mahoney, 2016, Sub-sampled newton methods with non-uniform sampling, *in* Advances in Neural Information Processing Systems 29: Curran Associates, Inc., 3000–3008.

Xu, S., D. Wang, F. Chen, G. Lambaré, and Y. Zhang, 2012, *in* Inversion on Reflected

Seismic Wave: 1–7.

Yang, P., R. Brossier, L. Mtivier, and J. Virieux, 2016, Wavefield reconstruction in attenuating media: A checkpointing-assisted reverse-forward simulation method: Geophysics, **81**, R349–R362.

Yang, P., R. Brossier, L. Mtivier, J. Virieux, and W. Zhou, 2018a, A Time-Domain Preconditioned Truncated Newton Approach to Visco-acoustic Multiparameter Full Waveform Inversion: SIAM Journal on Scientific Computing, **40**, B1101–B1130.

Yang, Y., and B. Engquist, 2018, Analysis of optimal transport and related misfit functions in full-waveform inversion: Geophysics, **83**, A7–A12.

Yang, Y., B. Engquist, J. Sun, and B. F. Hamfeldt, 2018b, Application of optimal transport and the quadratic wasserstein metric to full-waveform inversion: Geophysics, **83**, R43–R62.

Yee, K., 1966, Numerical solution of initial boundary value problems involving maxwell's equations in isotropic media: IEEE Transactions on Antennas and Propagation, **14**, 302–307.

Zhao, L., T. H. Jordan, and C. H. Chapman, 2000, Three-dimensional Fréchet differential kernels for seismicdelay times: Geophysical Journal International, **141**, 558–576.

Zhou, W., R. Brossier, S. Operto, and J. Virieux, 2015, Full waveform inversion of diving & reflected waves for velocity model building with impedance inversion based on scale separation: Geophysical Journal International, **202**, 1535–1554.

Zhu, H., E. Bozdag, and J. Tromp, 2015, Seismic structure of the European upper mantle based on adjoint tomography: Geophysical Journal International, **201**, 18–52.

Zuberi, M. A., and R. G. Pratt, 2017, Mitigating nonlinearity in full waveform inversion using scaled-Sobolev pre-conditioning: Geophysical Journal International, **213**, 706–725.

APPENDIX A

Shaping covariance operator

In this section, we define structure tensors and provide details about the shaping covariance operator. Equations are expressed in terms of continuous scalar fields (that are a function of position $\mathbf{x}$); however, they can be readily adapted to discrete systems. Where suitable, we drop the spatial dependence of variables for brevity. Structure tensors $\mathbf{T}(\mathbf{x})$ are outer products of the gradient operator applied to an arbitrary scalar field $p(\mathbf{x})$:

$$\mathbf{T}(\mathbf{x}) = (\nabla p(\mathbf{x}))(\nabla p(\mathbf{x}))^T. \tag{A.1}$$

Structure tensors are symmetric positive-definite matrices that are commonly used to extract orientation information from images. In the 2D case, $\mathbf{T}$ is a 2×2 matrix defined at each position $\mathbf{x}$. In the discrete setting, a structure tensor would be defined for each discrete point, or pixel in the case of images. For each position, the Eigen decomposition of $\mathbf{T}$ can be written as

$$\mathbf{T} = \lambda_1 \mathbf{v}_1 \mathbf{v}_1^T + \lambda_2 \mathbf{v}_2 \mathbf{v}_2^T, \tag{A.2}$$

where λ_1 and λ_2 are the eigenvalues of $\mathbf{T}$ associated with the eigenvectors $\mathbf{v}_1$ and $\mathbf{v}_2$. Hale (2014) defines a smoothing or diffusion tensor $\mathbf{D}(\mathbf{x})$. The diffusion tensor shares the same eigenvectors as $\mathbf{T}$ but alters the eigenvalues. In our implementation, we define $\mathbf{D}$ as

$$\mathbf{D} = \kappa_1 \mathbf{v}_1 \mathbf{v}_1^T + \kappa_2 \mathbf{v}_2 \mathbf{v}_2^T, \tag{A.3}$$

where $\kappa_1 = \alpha$ and $\kappa_2 = \alpha + (1 - \alpha) \exp\left(-\frac{\lambda_1}{\lambda_1 - \lambda_2}\right)$; α is a small value taken to be 0.01. By rearranging the eigenvalues, the largest eigenvectors of $\mathbf{D}$ ($\mathbf{v}_2$) are oriented parallel to coherent structures in the scalar field q.

To begin the definition of the smoothing covariance, we define the scalar fields $p(\mathbf{x})$ and

$q(\mathbf{x})$ that are related via

$$q(\mathbf{x}) = \mathbf{C_M}(\mathbf{x})p(\mathbf{x}), \tag{A.4}$$

where $\mathbf{C_M}(\mathbf{x})$ denotes a covariance operator. Hale (2014) equates the application of the smoothing covariance operator, to the solution of the following anisotropic PDE:

$$|\mathbf{D}|^{-\frac{1}{4}}(\mathbf{x})(1 - \alpha\nabla \cdot \mathbf{D}(\mathbf{x}) \cdot \nabla)^l(1 - \beta\nabla \cdot \mathbf{D}(\mathbf{x}) \cdot \nabla)|\mathbf{D}|^{-\frac{1}{4}}(\mathbf{x})q(\mathbf{x}) = \gamma p(\mathbf{x}), \tag{A.5}$$

where α and β are scalars that characterize the covariance operator (see Hale (2014) for more details). We reiterate that $\mathbf{D}$ is a non-stationary diffusion tensor. In this study, we select $l = 2, \alpha = 1$ and $\beta = 0$ and do not further experiment with these parameters. Hale (2014) proposes to solve equation Equation A.5 by sequentially solving a series of linear problems:

$$q_0(\mathbf{x}) = |\gamma^2\mathbf{D}(\mathbf{x})|^{\frac{1}{4}}p(\mathbf{x}) \tag{A.6}$$

$$(1 - \nabla \cdot \mathbf{D}(\mathbf{x}) \cdot \nabla)q_1(\mathbf{x}) = q_0(\mathbf{x}) \tag{A.7}$$

$$(1 - \nabla \cdot \mathbf{D}(\mathbf{x}) \cdot \nabla)q_2(\mathbf{x}) = q_1(\mathbf{x}) \tag{A.8}$$

$$q(\mathbf{x}) = |\gamma^2\mathbf{D}(\mathbf{x})|^{\frac{1}{4}}q_2(\mathbf{x}). \tag{A.9}$$

The scalar γ is a scaling term that is proportional to a user-defined scale length r_0 ($\gamma \propto 4\pi r_0^2$); further specifics can be found in Hale (2014). The characteristic scale length r_0 defines the spatial range of the covariance operator. Larger scale lengths coincide with larger spatial correlations in the covariance operator. Equations A.7 and A.8 are solved with linear conjugate gradient iterations. The covariance operator can be decomposed as $\mathbf{C_M} = \mathbf{F}\mathbf{F}^T$. The result of solving Equation A.6 followed by A.7 represents the application of $\mathbf{F}$. Likewise, the solution of A.7 followed by A.8 represents the application of $\mathbf{F}^T$.