Probabilistic waveform inversion: Quest for the law

Zur Erlangung des akademischen Grades

DOKTOR DER NATURWISSENSCHAFTEN (Dr. rer. nat.)

von der KIT-Fakultät für Physik des
Karlsruher Instituts für Technologie (KIT)

genehmigte

DISSERTATION

von

Master of Physics

Renat Shigapov

Tag der mündlichen Prüfung:	8. Februar 2019
Referent:	Prof. Dr. Thomas Bohlen
Korreferent:	Prof. Dr. Andreas Fichtner
Korreferent:	Prof. Dr. Jean Virieux

> Natural science is not only about proofs but more about the setting and resolution of problems. Mathematics is just a language that is useful for the setting, the resolutions, and especially for the proofs. [...]
>
> The research in empirical inference science requires searching for new models of inference (different from inductive inference, such as inference in Universum environment, transductive, selective, ad hoc inferences, and so on). They are currently not under the scope of interest of mathematicians since they do not yet have clear settings and clear resolutions (this is the main subject of research). Mathematicians will become interested in this subject later when new settings, new resolutions, and new ideas for proofs are found.
>
> The goal of the empirical inference discipline is to find these elements of the theory.

Vladimir Vapnik, "Estimation of Dependences Based on Empirical Data", p. 491, 2006.

> The idea that solving the inverse problem corresponds to obtaining one 'best' model requires revision. All the literature based on this paradigm suggests that solving an inverse problem consists somehow in 'extracting' a model from the data, using techniques reminiscent of those in signal processing. This is not necessarily the right point of view. What we should do is to base the inverse problem on a modified version of the popperian paradigm: data are not to be used to create a model, but, instead, to falsify models.
>
> I suggest that the setting, in principle, for an inverse problem should be as follows: use all available *a priori* information to sequentially create models of the system, potentially an infinite number of them. For each model, solve the forward modelling problem, compare the predictions to the actual observations and use some criterion to decide if the fit is acceptable or unacceptable, given the uncertainties in the observations and, perhaps, in the physical theory being used. The unacceptable models have been falsified, and must be dropped. The collection of all the models that have not been falsified represent the solution of the inverse problem.
>
> This concept of passing from a 'prior collection of models' to a 'posterior collection of models' will certainly be acceptable by the lovers of Bayes' paradigm, as the collections of models can be seen as samples of a prior probability distribution and samples of a posterior distribution. It should also please the believers in Popper's point of view. Although still far from unanimity, there are quests in this world too difficult to attain.

Albert Tarantola, "Popper, Bayes and the inverse problem", "Nature Physics" 2, 2006.

> The reciprocal relationship of epistemology and science is of noteworthy kind. They are dependent upon each other. Epistemology without contact with science becomes an empty scheme. Science without epistemology is - insofar as it is thinkable at all - primitive and muddled. However, no sooner has the epistemologist, who is seeking a clear system, fought his way through to such a system, than he is inclined to interpret the thought-content of science in the sense of his system and to reject whatever does not fit into his system. The scientist, however, cannot afford to carry his striving for epistemological systematic that far. He accepts gratefully the epistemological conceptual analysis; but the external conditions, which are set for him by the facts of experience, do not permit him to let himself be too much restricted in the construction of his conceptual world by the adherence to an epistemological system. He therefore must appear to the systematic epistemologist as a type of unscrupulous opportunist: he appears as realist insofar as he seeks to describe a world independent of the acts of perception; as idealist insofar as he looks upon the concepts and theories as free inventions of the human spirit (not logically derivable from what is empirically given); as positivist insofar as he considers his concepts and theories justified only to the extent to which they furnish a logical representation of relations among sensory experiences. He may even appear as Platonist or Pythagorean insofar as he considers the viewpoint of logical simplicity as an indispensable and effective tool of his research.

Albert Einstein in "Replies to Criticism", "Albert Einstein: Philosopher-Scientist" edited by P. A. Schilpp, 683-684, 1949.

Abstract

Probabilistic waveform inversion: Quest for the law

by Renat SHIGAPOV

Full-waveform inversion (FWI) is an algorithm (and a part of the measuring procedure in a wide sense) with the aim to find the governing law of a physical system using the partially measured physical fields with limited computational resources. A law is a forward theory equipped with the model parameters and the data parameters. The main characteristic of the law is the realizability assumption: the law explains all subsets of the measured data parameters and predicts all subsets of the unmeasured (in the given experiment) data parameters. To find the law, we have to guess a law (a forward theory and parametrization), measure some data parameters and check the realizability assumption.

To put it more precisely, I formulate a new probabilistic setting for inverse problems and full-waveform inversion. Instead of using the Bayes' theorem, the Tarantola-Valette conjunction or the principle of maximum entropy based on the prior information for the averaged quantities, I propose a principle of minimum relative information using the prior information for the non-averaged quantities. The Tarantola-Valette formula is obtained as a special case under the assumption that the theoretical and prior measures exist. Using the realizability assumption as a prior information, the principle of minimum relative information provides the parametric probabilistic solution with the arbitrary misfit functions. Maximization of the parametric probabilistic solution leads to a multiobjective minimization problem. All global Pareto optima are the sample points of the probabilistic solution with the highest values of the volumetric measure. Unfortunately, even a local multiobjective minimization problem is computationally intractable for FWI with many millions of model parameters.

To make it computationally attractive for large-scale FWI and to find at least a few local solutions of the multiobjective minimization problem, I implement the bilevel multiobjective waveform inversion (BMWI) using a single randomly chosen shot gather at each iteration. BMWI is a stochastic, nested algorithm with an adaptive parabolic line search and multiscale strategy. The computational cost per iteration is five forward modellings only. BMWI can worsen some of the single-shot misfit functions and the different random runs of BMWI converge to different points in the model manifold. I interpret these inverted models as the sample points of the probabilistic solution. I estimate the solution, uncertainty and sensitivity using the sample estimates of the mean, standard deviation and initial deviation of the sample points, respectively. Using the numerical examples with the Marmousi-2 model, I illustrate the potential of BMWI for automatic uncertainty and sensitivity analysis with just two-three sample points.

To test the idea with real-world data, I apply stochastic single-shot BMWI in a 2D acoustic finite-difference approximation to a 2D line of pressure data acquired in a shallow-water river delta with ocean bottom cables. I use minimal data preprocessing (only a new 3D-to-2D transform which is strictly valid in a linear-gradient medium), the linear gradient starting models and the diagonal preconditioners with a negligible regularization. I estimate the theoretical uncertainties due to the neglected 3D effects using the 3D-to-2D transforms. The uncertainties estimated by the random sequences of BMWI are higher than the uncertainties related to the 3D-to-2D transforms. I provide the estimates of the solution, uncertainty and sensitivity using up to fourteen sample points inverted with the different linear-gradient starting models, the differently 3D-to-2D-transformed real data sets and the different random sequences of descent directions. The uncertainty of sound velocities is the lowest in the central semicircle with the radius 3 km equal to half the length of the ocean bottom cable. The uncertainty of mass densities is the highest in the central semicircle. The sensitivity of the measuring procedure with respect to sound velocity and mass density is the highest in the central semicircle representing a footprint of the acquisition geometry. Outside the central semicircle the parameters are not falsifiable in the specified setting.

Full-waveform inversion is the quest for the *unique* governing law of the physical system under study. If the governing law is deterministic and the sample mean, standard deviation and initial deviation of the sample points represent the insufficient description of the solution, uncertainty and sensitivity, then the measuring procedure in a wide sense has to be improved.

Acknowledgements

Огромная благодарность моей Семье.

Russian period. The story starts in **Saint Petersburg**. I was lucky to study at Saint Petersburg State University and to meet there **Boris Markovich Kashtan** who gave me much more than knowledge. He shared with us, his students, a part of his freedom-loving and curious spirit, i.e., something untrainable by reading the folios and dissertations. I am deeply thankful to him. I thank **Vladimir Nikolaevich Troyan** for being a great mentor and teacher of inverse problems. I deeply thank **Anton Vetlugin**, **Ivan Abakumov** and **Vladimir Kazei** for unforgettable adventures and inspiring discussions. I thank **Denis Nasyrov**, **Maria Tatanova**, **Denis Anikiev**, **Dmitry Alexandrov**, **Andrey Ponomarenko** and **Pavel Znak** for many valuable discussions. I thank **Knut Richter** for explaining me the Stackelberg game.

Dutch period. The story continues in **The Hague**. I was lucky to work as an intern in the Shell's Rijswijk office. I deeply thank **Wim Mulder** for sharing many valuable thoughts and observations. Wim often asked me about my research question, motivating me to find a good one. **Fons ten Kroode** gave me that opportunity. **René-Édouard Plessix** inspired me by his working ability. **Alexander Droujinine** pushed my work forward and gave me many valuable suggestions. I thank **Boris Kuvshinov** and **Joost van der Neut** for inspiring discussions.

German period. The story shapes in **Karlsruhe**. I was lucky to have the kindest Doktorvater: I deeply thank **Thomas Bohlen** for giving me the freedom to work on topics of my interest, for encouraging my work, for motivating me and for saying with a smile "fight for your ideas". I thank **Claudia Payne** for administrative support and **Petra Knopf** for IT support. I deeply thank **André Kurzmann** for our attempts to understand FWI. I deeply thank **Tao Lei**: our discussions on FWI and mathematical optimization were simply necessary for existence of this work. I deeply thank **Yudi Pan** for his support, advice and our discussions on multiobjective FWI. My very special thanks go to **Thomas Hertweck** for being a mentor who is always in touch, for improving this work significantly and for sharing the knowledge about the industry, science and life. I highly value my discussions with **Thomas Forbriger**. **Laura Gaßner**, **Niklas Thiel** and **Nikolaos Athanasopoulos** improved my understanding of real-world FWI. I was happy to work in the same group with **Florian Wittkamp** and **Martin Pontius**, they always made my mood better. I thank the trio – **Fabian Kühn**, **Valérie Krampe** and **Daniel Krieger** – for inspiring me by being very good students. I thank **Tilman Steinweg** for concerts, **Eva Schroth** for chats and **Carlos Assis** for many very good questions. I am very grateful to **Andreas Fichtner** and **Jean Virieux** for agreeing to review this work and for their constructive, kind and encouraging comments.

Natalia danke ich für ihre Liebe.

Contents

To Albert Tarantola,

1949-2009,

a man of ideas

Introduction

Research question

> In general we look for a new law by the following process. First we guess it. Then we compute the consequences of the guess to see what would be implied if this law that we guessed is right. Then we compare the result of the computation to nature, with experiment or experience, compare it directly with observation, to see if it works. If it disagrees with experiment it is wrong. In that simple statement is the key to science. It does not make any difference how beautiful your guess is. It does not make any difference how smart you are, who made the guess, or what his name is – if it disagrees with experiment it is wrong. That is all there is to it.
>
> Richard Feynman, "The character of physical law"

The goal of physics is to find the law of nature. The goal of geophysics is to find the law of Earth. The principle of stationary action provides a rule for it: just guess a Lagrangian density $\mathcal{L}(\mathbf{f}(\mathbf{x}), \nabla_{\mathbf{x}}\mathbf{f}(\mathbf{x}), \mathbf{x})$ as a function of the fields $\mathbf{f}(\mathbf{x})$, its derivatives $\nabla_{\mathbf{x}}\mathbf{f}(\mathbf{x})$ and space-time coordinates $\mathbf{x}$ which optimizes the functional of action $\mathcal{S} = \int_{\mathbb{X}} \mathcal{L}(\mathbf{f}(\mathbf{x}), \nabla_{\mathbf{x}}\mathbf{f}(\mathbf{x}), \mathbf{x}) \sqrt{\det \hat{\mathbf{g}}(\mathbf{x})} d\mathbf{x}$, where $\hat{\mathbf{g}}(\mathbf{x})$ is a metric tensor of the space-time $\mathbb{X}$. The first variation of action yields the partial differential Euler-Lagrange equations for the fields. To make them *computable*, the fields with an infinite number of degrees of freedom are discretized. The disretized fields lead to the *finite* number of model $\mathbf{m}$ and data $\mathbf{d}$ parameters. The *numerical algorithm* which relates the model parameters and the data parameters is the *approximate forward theory*[1] or *forward solver*: $\mathcal{F}(\mathbf{m}) = \mathbf{d}$. If the forward theory and the

[1] Throughout this work I often call $\mathcal{F}$ a forward theory and $\mathcal{F}^{-1}$ an inverse theory because of tradition. Practically, even the analytical solution of the Euler-Lagrange equations is computed numerically with a finite precision. $\mathcal{F}$ is an algorithm. Its inverse $\mathcal{F}^{-1}$ is an algorithm as well.

values of parameters are *consistent* with the observed data, a candidate for the governing law of the physical system is found.

If we knew the governing law of the physical system "Earth" (the unique forward theory and the unique values of the model parameters and the data parameters), all secondary problems of geophysics – for instance, the location of hydrocarbons in exploration geophysics – would be solved. We are far away from that for at least three reasons.

First, an *experimental procedure* for any physical system has limitations. The physical fields are measured only in some points of space-time and never everywhere within a system under study. The non-injective (not one-to-one) and imprecise inverse theory $\mathcal{F}^{-1}$ is the main source of uncertainties: the smaller the number of the measured data parameters with respect to the total number of the data parameters required to describe a physical system, the more combinations of the parameters and theories can fit the observations.

Second, a physically realizable *computational system* has limited resources. With limited computational resources we can find at most only a finite number of combinations of the parameters and theories which fit the observations at least partially.

Third, the choice of a *formal system* (rules of inference and axioms) is not unique. We can set up the problem of finding the governing law of a physical system in many different ways and get different answers. The notion "solution" differs in probabilistic and deterministic formulations.

Full-waveform inversion is an algorithm (and a part of the measuring procedure in a wide sense) with the aim to find the governing law of a physical system using the partially measured physical fields with limited computational resources. The measuring procedure in a wide sense includes the experimental procedure, the computational and formal systems. Any measurement is inexact and has to be equipped with information on uncertainty and sensitivity.

This leads to the research question of this work: *What are the "solution", "uncertainty" and "sensitivity" in full-waveform inversion?*

In the next two sections I formulate the standard formal systems for solving the inverse problems (the Tarantola-Valette setting and the principle of maximum entropy) and present a short introduction into full-waveform inversion. After getting a feeling for the research question, I formulate the thesis statement.

Probabilistic inverse problems

An inverse problem can be set as: 1) a problem of mathematical optimization in the deterministic and frequentist interpretations (Aster et al., 2005; Backus and Gilbert, 1968; Kirsch, 1996; Menke, 2018; Parker, 1994; Tikhonov, 1963) and 2) a problem of inference or decision-making in the Bayesian, Tarantolian and Jaynesian interpretations (Calvetti and Somersalo, 2018; Dashti and Stuart, 2017; Jaynes, 1984; Kaipio and Somersalo, 2006; Scales and Snieder, 2000; Stuart, 2010; Tarantola, 2005).

I consider an inverse problem as a measurement (Mosegaard and Tarantola, 2002). Following Tarantola (2007) I interpret any measuring act as an attempt to obtain information on the position of the point $\mathbf{q}$, characterized by the quantities (or coordinates) $\{q_1, q_2, ...\}$ in the abstract quality (or parameter) manifold[2] $\mathbb{Q}$ equipped with the metric tensor $\mathbf{g}(\mathbf{q})$ and the volume density $v(\mathbf{q}) = \sqrt{\det \mathbf{g}(\mathbf{q})}$. The quantities $\{q_1, q_2, ...\} = \{d_1, d_2, ..., m_1, m_2, ..\}$ are the coordinates of a point $\mathbf{q}$ in the quality manifold $\mathbb{Q}$, where $\{d_1, d_2, ...\}$ are the data quantities and $\{m_1, m_2, ...\}$ are the model quantities. The data quantities $\{d_1, d_2, ...\}$ are the coordinates of a point $\mathbf{d}$ in the data manifold $\mathbb{D}$. The model quantities $\{m_1, m_2, ...\}$ are the coordinates of a point $\mathbf{m}$ in the model manifold $\mathbb{M}$. To parametrize a physical system means to choose the coordinates in the quality manifold. A forward theory $\mathcal{F}$ is the mapping $\mathcal{F}(\mathbf{m}) = \mathbf{d}$. An inverse theory $\mathcal{F}^{-1}$ is the inverse mapping $\mathcal{F}^{-1}(\mathbf{d}) = \mathbf{m}$.

Any measuring procedure starts with a *formulation stage* (BIPM et al., 2009, 2008): specification of the input quantities $\mathbf{q}_{in}$, the output quantities $\mathbf{q}_{out}$, the theory relating the quantities $\mathbf{q}_{in}$ and $\mathbf{q}_{out}$ and the joint probability distribution over $\mathbf{q}_{in}$. At the *calculation stage* the propagation of uncertainty from $\mathbf{q}_{in}$ to $\mathbf{q}_{out}$ through the theory is performed (BIPM et al., 2008) or the theoretical information on $\mathbf{q}_{in}$ is combined with the prior information on $\mathbf{q}_{in}$ in the Tarantola-Valette setting (Tarantola and Valette, 1982b). The principle of maximum entropy can be used to assign the probability density to the input quantities.

[2] Albert Tarantola introduced the terms "physical quantity", "physical quality" and "quality manifold" in Tarantola (2006a).

The Tarantola-Valette setting

> The Tarantola-Valette inverse theory is not a theory because it's not falsifiable. It is a framework.
>
> Andreas Fichtner, personal communication

At least three versions of the Tarantola-Valette setting exist. The classic formulation (Tarantola, 2005, 1987; Tarantola and Valette, 1982b) is based on *conjunction* between the theoretical $\hat{\Theta}(\mathbf{q})$ and prior $\hat{\rho}(\mathbf{q})$ *measure densities* defined over the parameter manifold $\mathbb{Q}$: $\hat{\sigma}(\mathbf{q}) = \hat{\rho}(\mathbf{q}) \wedge \hat{\Theta}(\mathbf{q})$. Alternatively, Tarantola formulated the setting using the *product* of the theoretical $\Theta(\mathbf{q})$ and prior $\rho(\mathbf{q})$ *volumetric measures* $\sigma(\mathbf{q}) = \rho(\mathbf{q})\Theta(\mathbf{q})$ in the Wolfram MathWorld note[3] and using *intersection* of the *volumetric measures* in the unfinished book "Mapping of Probabilities" (Tarantola, 2007).

Tarantola and Valette proposed a probabilistic setting for inverse problems without the Bayes' theorem and conditional probabilities in the seminal paper "Inverse problems = Quest for information" (Tarantola and Valette, 1982b). The Bayesian setting is a special case of the Tarantola-Valette setting if the conditional probabilities are introduced.

I formulate the setting using the measure densities and the volumetric measures[4]. Following Tarantola (2007), I call the operation between the measures as "intersection" instead of "conjunction".

The Tarantola-Valette setting is based on three axioms.

The first axiom specifies an abstract geometric description of a physical system: the parameter or quality manifold $\mathbb{Q}$. A physical system can be described with a *finite* number of parameters (or quantities) and the specific values of these parameters $\{q_1, q_2, ...\}$ define the coordinates of a point $\mathbf{q}$ in the abstract parameter manifold $\mathbb{Q}$ which has a notion of volume. The metric tensor $\mathbf{g}(\mathbf{q})$ specifies the volume density $v(\mathbf{q}) = \sqrt{\det \mathbf{g}(\mathbf{q})}$.

The second axiom introduces a probabilistic description of a physical system or a subjective degree of knowledge on the state of a physical system. The degree of knowledge on the values of quantities $\{q_1, q_2, ...\}$ can be described using a real, *non-negative*[5] function $\hat{m}(\mathbf{q})$, called a measure density, or using the volumetric measure $m(\mathbf{q})$. The measure density and the volumetric measure are related via the volume density $\hat{m}(\mathbf{q}) = m(\mathbf{q})v(\mathbf{q})$, so that the measure $M(A)$ of any subset A of $\mathbb{Q}$ is the integral $M(A) = \int_A \hat{m}(\mathbf{q})d\mathbf{q} = \int_A m(\mathbf{q})v(\mathbf{q})d\mathbf{q}$.

The third axiom defines a basic operation between measures. The intersection $\cap$ of the measure densities $\hat{m}_1(\mathbf{q})$ and $\hat{m}_2(\mathbf{q})$ is

$$\hat{m}_1(\mathbf{q}) \cap \hat{m}_2(\mathbf{q}) = \frac{\hat{m}_1(\mathbf{q})\hat{m}_2(\mathbf{q})}{v(\mathbf{q})}. \tag{1}$$

The intersection of the corresponding volumetric measures $m_1(\mathbf{q})$ and $m_2(\mathbf{q})$ is

$$m_1(\mathbf{q}) \cap m_2(\mathbf{q}) = m_1(\mathbf{q})m_2(\mathbf{q}). \tag{2}$$

The definitions are consistent:

$$\frac{\hat{m}_1(\mathbf{q}) \cap \hat{m}_2(\mathbf{q})}{v(\mathbf{q})} = m_1(\mathbf{q}) \cap m_2(\mathbf{q}) = \frac{\hat{m}_1(\mathbf{q})}{v(\mathbf{q})}\frac{\hat{m}_2(\mathbf{q})}{v(\mathbf{q})}. \tag{3}$$

The solution of an inverse problem is the a-posteriori measure density $\hat{\sigma}(\mathbf{q})$ given by the intersection of the prior $\hat{\rho}(\mathbf{q})$ and theoretical $\hat{\Theta}(\mathbf{q})$ measure densities

$$\hat{\sigma}(\mathbf{q}) = \hat{\rho}(\mathbf{q}) \cap \hat{\Theta}(\mathbf{q}) = \frac{\hat{\rho}(\mathbf{q})\hat{\Theta}(\mathbf{q})}{v(\mathbf{q})}, \tag{4}$$

[3] http://mathworld.wolfram.com/InverseProblem.html (last accessed January 4, 2019)

[4] Tarantola (2005) mentioned that for pedagogical reasons he formulated the setting in terms of the probability densities only, although for research he preferred to use the volumetric probabilities.

[5] In Tarantola and Valette (1982b) and Tarantola (2005) the measure density is a positive function. Tarantola (2005) cited the Radon-Nikodym theorem, but it requires only the non-negative measure densities. In the unfinished book of Tarantola (2007) the Radon-Nikodym theorem with the non-negative measure densities is used.

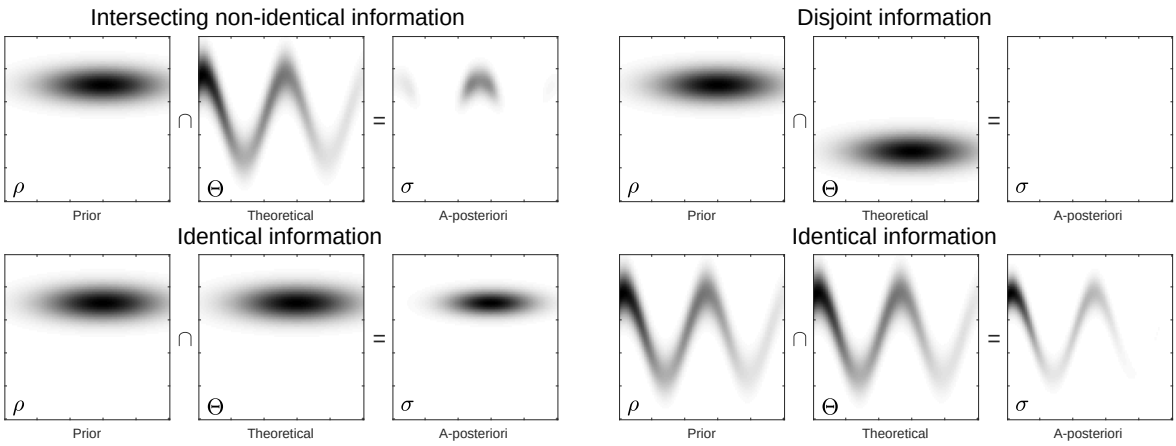


FIGURE 1: Schematic illustration: Intersection ∩ of the prior and theoretical volumetric measures in the Tarantola-Valette setting.

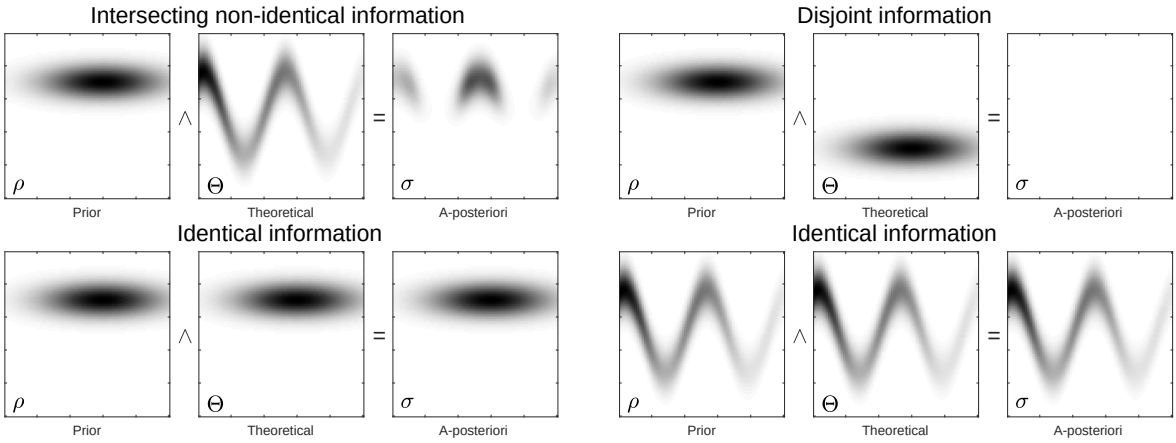


FIGURE 2: Schematic illustration: Conjunction ∧ of the prior and theoretical volumetric measures.

or the a-posteriori volumetric measure $\sigma(\mathbf{q})$ given by the intersection of the prior $\rho(\mathbf{q})$ and theoretical $\Theta(\mathbf{q})$ volumetric measures

$$\sigma(\mathbf{q}) = \rho(\mathbf{q}) \cap \Theta(\mathbf{q}) = \rho(\mathbf{q})\Theta(\mathbf{q}). \tag{5}$$

The two equivalent representations of the solution are related by the volume density: $\hat{\sigma}(\mathbf{q}) = \sigma(\mathbf{q})v(\mathbf{q})$. The problem is always well-posed: the probabilistic solution exists and is unique and stable. The formulation is invariant with respect to the change of coordinates.

The a-posteriori measure density $\hat{\sigma}_{\mathbb{M}}(\mathbf{m})$ and the a-posteriori volumetric measure $\sigma_{\mathbb{M}}(\mathbf{m})$ defined over the model manifold can be obtained by integrating the formulas (4) and (5) over the data manifold $\mathbb{D}$:

$$\sigma_{\mathbb{M}}(\mathbf{m}) = \rho(\mathbf{m})L(\mathbf{m}), \qquad \hat{\sigma}_{\mathbb{M}}(\mathbf{m}) = \hat{\rho}(\mathbf{m})\hat{L}(\mathbf{m}), \tag{6}$$

where $L(\mathbf{m})$ and $\hat{L}(\mathbf{m})$ are the likelihood functions.

In contrast to the formula (4) formulated with the volume density $v(\mathbf{q})$, Tarantola used $\mu(\mathbf{q}) = v(\mathbf{q})/V$, where V is the volume of the parameter manifold. The normalized measure density $\mu(\mathbf{q})$ represents null information (Tarantola and Valette, 1982b) and is called the homogeneous measure density (Mosegaard and Tarantola, 2002). For an infinite volume V the measure density $\mu(\mathbf{q})$ is undefined and Tarantola suggested to use $\mu(\mathbf{q}) \sim v(\mathbf{q})$. The uniform volumetric measure is $u(\mathbf{q}) = 1$. The uniform measure density is $\hat{u}(\mathbf{q}) = v(\mathbf{q})$, which is constant only for a flat quality manifold and Cartesian coordinates.

Tarantola used the term "conjunction" for the product of the volumetric measures which does not have an idempotent property. Figure 1 schematically illustrates the conjunction of a few different states of information[6]. In contrast, the logical conjunction and disjunction are idempotent[7] $x \wedge x = x$ and $x \vee x = x$ for any x. Moreover, the choice of conjunction and disjunction[8] in the Tarantola-Valette setting does not respect de Morgan's laws: $x \wedge y = \neg(\neg x \vee \neg y)$ and $x \vee y = \neg(\neg x \wedge \neg y)$, where $\neg$ is the logical negation.

A possible choice of the conjunction and disjunction which respects the idempotent property and de Morgan's laws is $\wedge = \min$ and $\vee = \max$. An idempotent conjunction can also be defined using the square root operation. Then, the a-posteriori measure density $\hat{\sigma}(\mathbf{q})$ and the a-posteriori volumetric measure $\sigma(\mathbf{q})$ have the same simple form

$$\hat{\sigma}(\mathbf{q}) = \hat{\rho}(\mathbf{q}) \wedge \hat{\Theta}(\mathbf{q}) = \sqrt{\hat{\rho}(\mathbf{q})\hat{\Theta}(\mathbf{q})}, \tag{7}$$

$$\sigma(\mathbf{q}) = \rho(\mathbf{q}) \wedge \Theta(\mathbf{q}) = \sqrt{\rho(\mathbf{q})\Theta(\mathbf{q})}. \tag{8}$$

Figure 2 schematically illustrates these formulas. The square root operation preserves the physical dimension of the volumetric measure in contrast to the intersection operation in the Tarantola-Valette setting which squares the physical dimension of the volumetric measure.

I present a simple derivation of the formulas (7)-(8) and (4)-(5) in chapter 1 using a new probabilistic setting for inverse problems. I also show a simple modification of the Tarantola-Valette formula which does not square the physical dimension of the a-posteriori volumetric measure.

The Tarantola-Valette setting does not provide a rule to assign the prior and theoretical measures. This is the main problem in any probabilistic setting.

The principle of maximum entropy

> The principle of maximum entropy is not an oracle telling which predictions *must* be right; it is a rule for inductive reasoning that tells us which predictions *are most strongly indicated by our present information*.
>
> Edwin T. Jaynes, "Probability theory: The logic of science"

The principle of maximum entropy proposed by Edwin Thompson Jaynes (1957a,b) is used to assign the measure densities based on the partial knowledge on the averaged quantities. It can also be used as an updating rule for the measure densities instead of the Tarantola-Valette setting and the Bayes' formula. Shore and Johnson (1980) have shown that the principles of maximum entropy and maximum relative entropy are the correct methods for inductive inference when new information is given in terms of expected values. Ulrych and Sacchi (2005) discussed the principle of maximum entropy with many geophysical examples.

According to the principle of maximum entropy the measure density $\hat{\sigma}(\mathbf{q})$ or the volumetric measure $\sigma(\mathbf{q})$ which represent the current stage of knowledge on the quantity $\mathbf{q}$ is the one corresponding to the maximum entropy constrained by the prior information. The prior information is given as a mathematical expectation of a certain function of $\mathbf{q}$. The constrained optimization problem is solved using the method of Lagrange multipliers.

The famous Tarantolian "null information" $\mu(\mathbf{q})$ proportional to the volume density $v(\mathbf{q})$ was introduced by Jaynes in 1963 as "a state of total ignorance" to make the Shannon's definition of

[6] Some people say that the problem of "non-idempontent conjunction" in the Tarantola-Valette setting does not exist, because the rule of Tarantola and Valette cannot be used to combine the redundant or identical states of information. The logic of these people is based on the following quote from Tarantola and Valette (1982b): "The conjunction, as defined above, must be used to combine two states of information only if these states of information have been obtained independently, as for example, for two independent physical measurements on a given set of parameters, or for combining experimental and theoretical information". However, the notions "independent" and "non-identical" are not the same.

[7] The principle of idempotence was first introduced into logic by Gottfried Wilhelm Leibniz (Gabbay et al., 2012).

[8] Tarantola (2005) introduced the disjunction as $\hat{m}_1(\mathbf{q}) \vee \hat{m}_2(\mathbf{q}) = \frac{1}{2}(\hat{m}_1(\mathbf{q}) + \hat{m}_2(\mathbf{q}))$ and $m_1(\mathbf{q}) \vee m_2(\mathbf{q}) = \frac{1}{2}(m_1(\mathbf{q}) + m_2(\mathbf{q}))$.

entropy invariant under the change of variables (Jaynes, 1963). Because of the lack of invariance, the classic Shannon's definition of entropy cannot be used for the continuous case.

The entropy is a mathematical expectation of the information $I(\sigma(\mathbf{q})) = -\log \sigma(\mathbf{q})$ contained in $\sigma(\mathbf{q})$:

$$H(\sigma) = -\int_{\mathbb{Q}} v(\mathbf{q})\sigma(\mathbf{q}) \log \sigma(\mathbf{q}) d\mathbf{q}, \tag{9}$$

where $v(\mathbf{q})$ is the volume density and $\sigma(\mathbf{q})$ is the volumetric measure.

The entropy of the measure density $\hat{\sigma}(\mathbf{q})$ is

$$H(\hat{\sigma}) = -\int_{\mathbb{Q}} \hat{\sigma}(\mathbf{q}) \log \frac{\hat{\sigma}(\mathbf{q})}{v(\mathbf{q})} d\mathbf{q}. \tag{10}$$

The two definitions of entropy are consistent, i.e., $H(\sigma) = H(\hat{\sigma})$, because the volumetric measure and the measure density are related $\hat{\sigma}(\mathbf{q}) = v(\mathbf{q})\sigma(\mathbf{q})$.

If we know about a quantity $\mathbf{q}$ that some functions $f_i(\mathbf{q})$ have the mathematical expectations $\int_{\mathbb{Q}} f_i(\mathbf{q})\hat{\sigma}(\mathbf{q})d\mathbf{q} = F_i$ with $i = 1, .., N$ and the measure density is normalized $\int_{\mathbb{Q}} \hat{\sigma}(\mathbf{q})d\mathbf{q} = 1$, then the principle of maximum entropy leads to the measure density in the exponential form

$$\hat{\sigma}(\mathbf{q}) = \frac{v(\mathbf{q})}{Z(\lambda_1, ..., \lambda_N)} \exp\left(\lambda_1 f_1(\mathbf{q}) + ... + \lambda_N f_N(\mathbf{q})\right), \tag{11}$$

with the partition function

$$Z(\lambda_1, ..., \lambda_N) = \int_{\mathbb{Q}} v(\mathbf{q}) \exp\left(\lambda_1 f_1(\mathbf{q}) + ... + \lambda_N f_N(\mathbf{q})\right) d\mathbf{q}, \tag{12}$$

and the values of parameters λ_k defined from the equations

$$F_k = \frac{\partial Z(\lambda_1, ..., \lambda_N)}{\partial \lambda_k}, \qquad k = 1, .., N. \tag{13}$$

The principle of maximum entropy leads to the exponential family of probability distributions. The exponential form of $\hat{\sigma}(\mathbf{q})$ is a basis to formulate the Bayes' theorem in the infinite dimensional setting as shown by Stuart (2010) and Dashti and Stuart (2017).

The formulas (9) and (10) contain the logarithms of dimensional quantities. To get rid of that, the principle of maximum entropy can be replaced by the principle of maximum relative entropy for the the volumetric measure $\sigma(\mathbf{q})$ with respect to the uniform volumetric measure $u(\mathbf{q})$:

$$H(\sigma) = -\int_{\mathbb{Q}} v(\mathbf{q})\sigma(\mathbf{q}) \log \frac{\sigma(\mathbf{q})}{u(\mathbf{q})} d\mathbf{q}, \tag{14}$$

and for the measure density $\hat{\sigma}(\mathbf{q})$ with respect to the uniform measure density $\hat{u}(\mathbf{q})$:

$$H(\hat{\sigma}) = -\int_{\mathbb{Q}} \hat{\sigma}(\mathbf{q}) \log \frac{\hat{\sigma}(\mathbf{q})}{\hat{u}(\mathbf{q})} d\mathbf{q}. \tag{15}$$

In chapter 1 I introduce a new principle – the principle of minimum relative information – to assign the measure densities using the prior information for the *non-averaged* quantities.

Sampling a sample and the sample points

The Tarantola-Valette formulas (4), (5) and (6) provide a simple probabilistic setting for inverse problems. The prior volumetric measures can be assigned using the principle of maximum entropy or using some assumptions about the symmetry of the problem. The probabilistic solution of an inverse problem is the a-posteriori volumetric measure $\sigma_{\mathbb{M}}(\mathbf{m})$ over the model manifold $\mathbb{M}$ or the a-posteriori volumetric measure $\sigma(\mathbf{q})$ over the quality manifold $\mathbb{Q}$.

In full-waveform inversion applications the number of dimensions of the model manifold $\mathbb{M}$ is $10^6 - 10^8$ and is continuing to grow. Due to the limited computational resources we cannot define

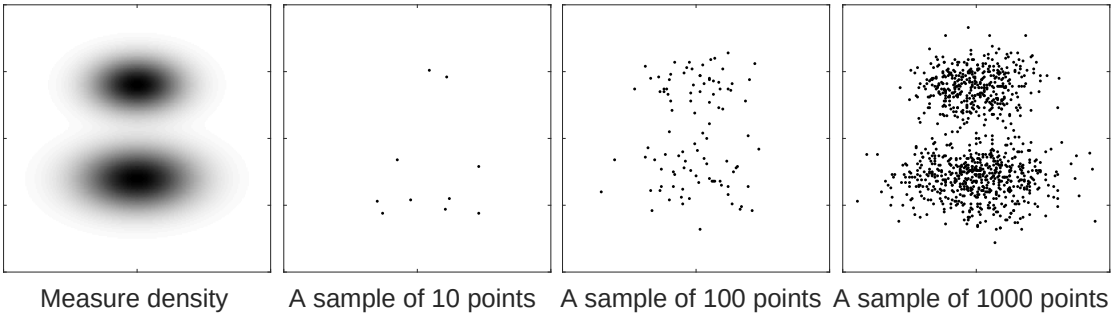


FIGURE 3: Sampling a two-modal measure density with 10, 100 and 1000 sample points.

an *arbitrarily dense* grid of points in $\mathbb{M}$ and for *every* point $\mathbf{m}$ compute the value of $\sigma_{\mathbb{M}}(\mathbf{m})$. The systematic sampling is computationally infeasible in large-dimensional spaces.

Instead, the a-posteriori measure density can be *sampled* by the *finite*, relatively small number of *sample points* so that the number of the sample points in *any* sub-volume divided by the total number of the sample points is proportional to the measure of this sub-volume. A *sample* is a set of the *sample points*[9]. Figure 3 shows a two-modal 2D measure density sampled with 10, 100 and 1000 sample points. An algorithm performing sampling is called the sampler, sampling algorithm or sampling method. The choice of a sampler depends on the available computational resources.

If sampling is performed, the solution of an inverse problem is a collection of models, the sample points of the a-posteriori measure density. Instead of plotting the mean values, Tarantola suggested to plot all sample points. He called it the movie strategy (Tarantola, 2005).

If the a-posteriori measure density is sampled with N sample points $\{\mathbf{m}_1, ..., \mathbf{m}_N\}$ in $\mathbb{M}$, then the sample mean is

$$\langle \mathbf{m} \rangle = \frac{1}{N} \sum_{k=1}^{N} \mathbf{m}_k. \tag{16}$$

The sample covariance matrix is

$$\langle \mathbf{C} \rangle = \frac{1}{N-1} \sum_{k=1}^{N} (\mathbf{m}_k - \langle \mathbf{m} \rangle)(\mathbf{m}_k - \langle \mathbf{m} \rangle)^{\top}. \tag{17}$$

The sample standard deviation is

$$\mathbf{V} = \sqrt{\frac{1}{N-1} \sum_{k=1}^{N} (\mathbf{m}_k - \langle \mathbf{m} \rangle)^2}, \tag{18}$$

where the operations are understood in a component-wise sense.

If the computation of $\sigma_{\mathbb{M}}(\mathbf{m})$ at any point $\mathbf{m}$ in $\mathbb{M}$ is inexpensive (*not* the case in this work), a random (Monte Carlo) method can be used to generate the samples of $\sigma_{\mathbb{M}}(\mathbf{m})$. The standard Monte Carlo samplers are the Metropolis-Hastings algorithm (Hastings, 1970; Metropolis et al., 1953) and the Gibbs sampler (Geman and Geman, 1984). If the computation of derivatives of $\sigma_{\mathbb{M}}(\mathbf{m})$ at any point $\mathbf{m}$ in $\mathbb{M}$ is inexpensive (*not* the case in this work), a Hamiltonian Monte Carlo method speeds up the sampling (Duane et al., 1987; Neal, 2011). Keilis-Borok and Yanovskaja (1967) and Press (1968) are the first examples of the Monte Carlo methods in geophysical inverse problems. The modern references in the geophysical context are (Mosegaard and Tarantola, 1995) and (Sambridge and Mosegaard, 2002).

[9]The relevant debate between Albert Tarantola and his colleagues about the notions "sample", "sample point" and "equiprobable sample points" can be found on Tarantola's web-page: http://www.ipgp.fr/~tarantola/Files/Professional/Diverse/SamplePoints (last accessed January 4, 2019). Tarantola defined the "random point", "independent random point", "sample point" and "independent sample point". Then, a sample is a collection of independent sample points. In this work I avoid to define the notions "random point" and "independent random point". I prefer to define the "sample points" as the points whose number in *any* sub-volume divided by the total number of the sample points is proportional to the measure of this sub-volume. The choice of a sampler – random or deterministic – is irrelevant with this definition.

For conceptual understanding of the first chapter in this work, it is sufficient to focus on one of the simplest sampling algorithms – the rejection method. The rejection sampler generates the sample points $\{\mathbf{q}_1, \mathbf{q}_2, ...\}$ of the uniform volumetric measure $u(\mathbf{q}) = 1$. The sample point $\mathbf{q}_k$ is accepted with the probability $\sigma(\mathbf{q}_k)/\max(\sigma(\mathbf{q}))$, where $\max(\sigma(\mathbf{q}))$ denotes the maximum value of the volumetric measure $\sigma(\mathbf{q})$ in $\mathbb{Q}$. All accepted points are the sample points of $\sigma(\mathbf{q})$.

According to Tarantola (2005), the main problems in random sampling of the probability densities in a large-dimensional space are 1) locating the regions of significant probability and 2) sampling the regions densely enough.

Tarantola (2006b) suggested to replace the *verification* of the models by *falsification* of the models in inverse problems. He suggested to use the following setting: to create various models (possibly an infinite number of them) and to use the data to falsify the models using *some criterion* which was not specified in the paper.

In his unfinished book Tarantola (2007) clarified the criterion in the section "The Bayes-Popper problem". He related the criterion to the Tarantola-Valette setting: all models that have been generated as the sample points of the prior volumetric measure $\rho(\mathbf{m})$, but have not been accepted as the sample points of the theoretical volumetric measure $\Theta(\mathcal{F}(\mathbf{m}))$, have been *falsified* (Tarantola, 2007, p. 38).

In chapter 2 I show that a stochastic, nested algorithm applied to the subsets of the data can be used to locate the regions of significant probability by inverting the subsets of model parameters (at least in the context of marine seismic experiments with ocean bottom cables).

Full-waveform inversion

To set up the problem of FWI, we guess a forward theory $\mathcal{F}$ and a finite number of parameters $\mathbf{q} = \{\mathbf{m}, \mathbf{d}\}$ describing the physical system under study and measure some data parameters $\mathbf{d}_o$.

In a deterministic setting FWI is formulated as a minimization of a misfit function between the measured $\mathbf{d}_o$ and modelled $\mathcal{F}_o(\mathbf{m})$ data parameters:

$$\min \ell(\mathcal{F}_o(\mathbf{m}), \mathbf{d}_o), \tag{19}$$

where ℓ is an arbitrary (possibly regularized) misfit.

The corresponding a-posteriori volumetric measure is

$$\sigma_{\mathbb{M}}(\mathbf{m}) = \exp(-\alpha \ell(\mathcal{F}_o(\mathbf{m}), \mathbf{d}_o)), \tag{20}$$

where α is an arbitrary positive parameter. The minimization problem (19) corresponds to the maximization of the volumetric measure (20).

The non-linear minimization problem (19) is solved using any local optimization algorithm (Nocedal and Wright, 1999) with an initial guess $\mathbf{m}_0$:

$$\mathbf{m}_{k+1} = \mathbf{m}_k + \alpha_k \boldsymbol{\delta}\mathbf{m}_k(\ell(\mathcal{F}_o(\mathbf{m}_k), \mathbf{d}_o)), \qquad k = 0, 1, ..., N-1, \tag{21}$$

where k is the number of iteration, α_k is a step length at the k-th iteration, $\boldsymbol{\delta}\mathbf{m}_k = \boldsymbol{\delta}\mathbf{m}_k(\ell(\mathcal{F}_o(\mathbf{m}_k), \mathbf{d}_o))$ is a descent direction at the k-th iteration and N is the total number of iterations specified by a certain stopping criterion.

Geometrically speaking, the single run of FWI is a sequence of points $\{\mathbf{m}_0, ..., \mathbf{m}_N\}$ in the model manifold $\mathbb{M}$ and the move from the point $\mathbf{m}_k$ to the point $\mathbf{m}_{k+1}$ is performed along a descent direction $\boldsymbol{\delta}\mathbf{m}_k$ times the value of a step length α_k ($k = 0, ..., N-1$). The performance of an FWI algorithm depends on the initial guess $\mathbf{m}_0$, the descent directions $\{\boldsymbol{\delta}\mathbf{m}_0, ..., \boldsymbol{\delta}\mathbf{m}_{N-1}\}$ and the step lengths $\{\alpha_0, ..., \alpha_{N-1}\}$. The final model parameters $\mathbf{m}_N$ are interpreted as an estimate of the solution.

The history of FWI (see Figure[10] 4, the monograph by Fichtner (2011) and the reviews by Virieux and Operto (2009) and Virieux et al. (2017)) has started in the 1980-s with three seminal contributions made by Albert Tarantola and his collaborators[11]. First, they presented a probabilistic

[10] The figure is dramatically subjective, very incomplete and, unfortunately, imprecise.

[11] This view on the history of FWI differs from the one which starts from the works presented by 1) Bamberger et al. (1982) where the inversion of the normal incidence seismograms was formulated as the optimal control problem based on the theory of Lions (1968) and 2) Lailly (1983) where the seismic inverse problem was formulated as a sequence of before stack migrations.

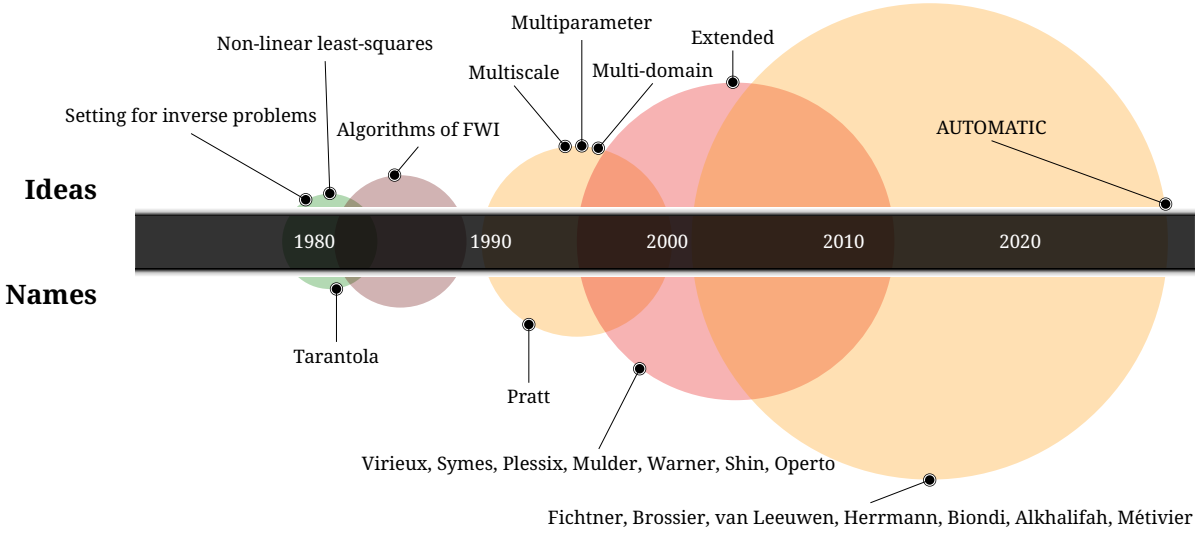


FIGURE 4: A brief history of full-waveform inversion.

setting for inverse problems with an imprecisely known theory (Tarantola and Valette, 1982b). Second, they formulated the solution of non-linear inverse problems using the generalized least squares criterion under the assumption of Gaussian probability distributions (Tarantola and Valette, 1982a). Third, Tarantola presented three classic papers on FWI in acoustic (Tarantola, 1984), elastic (Tarantola, 1986) and anisotropic viscoelastic media (Tarantola, 1988) where he explained how the gradient of a misfit function can be computed with only two forward modellings per shot gather. The method of computing the gradient of a misfit function without the Fréchet derivatives is called the adjoint-state method[12] (Chavent, 1974; Plessix, 2006; Tromp et al., 2005).

Gauthier et al. (1986) showed the first ever numerical implementation of FWI using the time-domain finite-difference code presented earlier by Jean Virieux (1984). Another early synthetic examples of FWI are presented by Mora (1988, 1987). The first applications to real data are (Crase et al., 1990; Crase et al., 1992; Pica et al., 1990). The significant development of FWI in 1990-s is due to Gerhard Pratt: the shift into the frequency domain, a matrix-vector formulation, reduction of the computational cost and application to laboratory and cross-hole data (Pratt, 1990, 1999; Pratt and Shipp, 1999; Pratt and Worthington, 1990; Pratt et al., 1998, 1996). Bunks et al. (1995) proposed the multiscale strategy. Brenders and Pratt (2007a,b) presented a successful reconstruction in a blind test. The first examples of 3D FWI are shown by Ben-Hadj-Ali et al. (2008), Fichtner et al. (2009), Plessix (2009), and Warner et al. (2008). The multiparameter aspects of FWI are discussed by Plessix et al. (2013) and Operto et al. (2013). Brossier et al. (2009) and Sears et al. (2008) applied elastic FWI to field data. The anisotropic FWI is presented by Alkhalifah (2014), Köhn et al. (2015), Plessix and Cao (2011), and Warner et al. (2013). The viscoacoustic (Askan et al., 2007; Kamei and Pratt, 2013; Malinowski et al., 2011; Yang et al., 2018a) and viscoelastic (Brossier, 2011; Charara et al., 2000; Yang et al., 2016a) FWI are challenging due to almost unavoidable cross-talks between sound velocity and attenuation coefficient (Hak and Mulder, 2010, 2011; Mulder and Hak, 2009). The truncated-Newton FWI is implemented by Lei et al. (2018), Métivier et al. (2014), and Yang et al. (2018a). The regularized FWI is shown by Asnaashari et al. (2013) and Guitton (2012). The memory-reduced FWI is based on a few ideas: the wavefields can be compressed (Boehm et al., 2016; Sun and Fu, 2013), recomputed from the snapshots-checkpoints (Komatitsch et al., 2016; Symes, 2007) and reverse-propagated from the saved boundary values (Clapp, 2008; Dussaud et al., 2008). These ideas can be combined (Yang et al., 2016b). The extended waveform inversion is discussed by Symes (2015, 2017). The spatial scale of FWI applications varies between a few mm and cm in medical imaging (Agudo et al., 2018; Goncharsky et al., 2014; Pratt, 2018; Pratt et al., 2007) and the hundreds of km in the project on a multiscale collaborative Earth model (Afanasiev et al., 2016; Fichtner et al., 2018).

[12] In the nuclear reactor calculations the closely related perturbation method with the forward and adjoint operators was developed in 1953-1956 by Marchuk (1959). Gurii Ivanovich Marchuk was the last president of the Soviet Academy of Sciences and is famous for his contributions to the numerical methods and the perturbation methods with adjoint operators. His most relevant for FWI-community monographs are Marchuk (1995) and Marchuk et al. (1996).

Fundamental problems and no free lunch

Full-waveform inversion suffers from two fundamental problems. The first one is the lack of priors about the problem in a wide sense: imprecise knowledge of the forward theory $\mathcal{F}$, inadequate parametrization $\mathbf{q}$, limitations of the experimental procedure hidden in the observed data parameters $\mathbf{d}_o$[13] and poor initial knowledge of the values of model parameters $\mathbf{m}_0$. The second fundamental problem is the lack of computational resources which limits the choice of a forward solver, parametrization and the method of mathematical optimization. If the computational resources are unlimited, a brute force approach provides the theories and model parameters which fit the observed data parameters with any desired accuracy.

The "no-free-lunch" theorem (Wolpert and Macready, 1997, 1995) in a finite-dimensional space and for the algorithms which do not visit (sample) the same point more than once can be stated as: all algorithms for searching an extremum of a misfit function perform exactly the same when averaged over all possible misfit functions. In the context of FWI this implies that for any FWI algorithm there exist a pathological setting where the algorithm fails and another algorithm will outperform it in this setting. The choice of an FWI-algorithm should be adapted to the problem, i.e., to a physical system under study and to the measuring procedure including the computational and formal systems.

Any challenge in FWI is a consequence of the lack of priors about the problem and the lack of computational resources. For example, *cycle skipping*: like any local optimization algorithm, FWI gets trapped in a local minimum if the forward solver $\mathcal{F}$ evaluated with an initial estimate of the model parameters $\mathbf{m}_0$ does not predict the measured wavefields $\mathbf{d}_o$ within half a wave-cycle. To reduce the effect of cycle skipping, the multiscale strategy, non-ℓ_2 misfits and extended inversion are used. The *multiscale strategy* is the inversion of the low-frequency data first and then inversion of the higher-frequency data. It helps if the low-frequency components of the data are measured in the experiment (Bunks et al., 1995; Fichtner et al., 2013). In the controlled-source experiments the low frequency components of the data are either damaged by noise or absent due to technological limits (ten Kroode et al., 2013). The *non-ℓ_2 misfits* include the envelope misfit (Bozdağ et al., 2011; Fichtner et al., 2008; Wu et al., 2014), the Wasserstein or Kantorovich-Rubinstein norms (Engquist and Froese, 2014; Lellmann et al., 2014; Métivier et al., 2016a,b; Métivier et al., 2018; Yang and Engquist, 2018; Yang et al., 2018b) and the correlation-, convolution- and deconvolution-based misfits (Luo and Sava, 2011; van Leeuwen and Mulder, 2010). The *non-physical extension* of the physical fields in space, time and in space-time and modifications of the misfit functions to minimize the difference between the non-physical and physical fields are developed by Biondi and Almomin (2014, 2013), Fu and Symes (2017a,b), Huang et al. (2017, 2018a,b), Symes (2008, 2009), van Leeuwen and Herrmann (2013b), Warner and Guasch (2016), and Zhu and Fomel (2016).

Another challenge is the *resolution* of FWI which depends on the quality of the descent directions related to the number and spatial distribution of the sources and receivers. Increasing resolution and decreasing computational cost of FWI are conflicting tasks. The resolution of FWI is limited by half a wavelength in the presence of many sources and receivers. The computational cost of FWI is proportional to the number of shot gathers and can be reduced using the stochastic simultaneous sources (Castellanos et al., 2015; Guitton and Diaz, 2012; Krebs et al., 2009; Moghaddam et al., 2013; van Leeuwen et al., 2011) and hybrid deterministic-stochastic shot-subsampling methods (Friedlander and Schmidt, 2012; Li et al., 2012; van Leeuwen and Herrmann, 2013a). The simultaneous sources reduce the resolution of FWI due to the cross-talks between different shot gathers.

In chapters 2 and 3 I show that a stochastic, nested algorithm with a single randomly chosen shot gather per iteration is a simple alternative to the algorithms mentioned above.

Solution, uncertainty and sensitivity

Until recently the main focus in FWI was to find one best-fit model without any uncertainty analysis. A few criteria were used to *falsify* the model: 1) Comparison of the time-domain (Brenders and Pratt, 2007b; Ravaut et al., 2004) and frequency-domain (Bleibinhaus et al., 2009; Malinowski et al., 2011) seismograms. 2) Consistency of the source time functions corresponding to the different shot gathers (Gao et al., 2007; Kamei and Pratt, 2013; Smithyman et al., 2009). 3) Overlay of the inverted models by the 3D depth migrated results (Plessix et al., 2013). 4) Comparison of a 1D velocity profile

[13]The data parameters $\mathbf{d}_o$ are measured in the limited number of space-time points and never everywhere within a physical system under study.

from the final velocity model with the sonic log data (Malinowski et al., 2011; Takougang and Calvert, 2011). 5) Re-migration with a new velocity model (Plessix et al., 2012). 6) Flatness of the common-image gathers (Plessix et al., 2012). 7) Consistency of the inverted velocity model with the prior geological information (Kamei and Pratt, 2013).

The uncertainty analysis in FWI is just at the initial stage of development and can be performed using the parameterized and low-rank approximations of the Hessian operator (Bui-Thanh et al., 2013; Fichtner and Trampert, 2011a,b), random probing (Fichtner and van Leeuwen, 2015) and ensemble transform Kalman filters (Thurin et al., 2017a; Thurin et al., 2017b). Lim (2016) used the randomized maximum likelihood sampler for the acoustic equation with less than 10^3 parameters. Fang (2018) and Fang et al. (2018) sampled the local Gaussian approximation of the posterior distribution for the acoustic equation with less than 10^5 parameters. A review on uncertainty analysis in tomography is given by Rawlinson et al. (2014) and Osypov et al. (2013).

The sensitivity analysis in FWI is often understood as the trade-off analysis between the different parameter classes by showing the scattered amplitudes (radiation patterns) from the different diffractors (Alkhalifah and Plessix, 2014; Gholami et al., 2013; Kazei and Alkhalifah, 2018). I give a different definition of sensitivity in chapter 2.

In chapters 2 and 3 I show how the stochastic single-shot FWI algorithm estimates the solution and uncertainty and sensitivity of the measuring procedure. The number of unknown parameters is more than 10^6 in chapter 2 and is more than 10^7 in chapter 3.

Thesis statement

The goal of full-waveform inversion is to find the unique governing law of the physical system under study. The idea of uniqueness of the governing law was essential for the development of modern science[14]. What is the main characteristic of the law? It is the realizability assumption[15]: the law explains all subsets of the measured data parameters and predicts all subsets of the unmeasured (in the given experiment) data parameters.

The thesis statement is simple: To find the governing law, guess a law (a forward solver and parametrization), measure some data parameters and check the realizability assumption. The rest of this work is nothing more than explanation of this idea in some details.

In chapter 1 I formulate a new probabilistic setting for inverse problems and full-waveform inversion. I propose a principle of minimum relative information using the prior information for the non-averaged quantities. Using the realizability assumption as a prior information, the principle of minimum relative information gives the parametric probabilistic solution with the arbitrary misfit functions. Maximization of the parametric probabilistic solution leads to a multiobjective minimization problem. All global Pareto optima are the sample points of the probabilistic solution with the highest values of the volumetric measure. However, even a local multiobjective minimization problem is computationally intractable for FWI with $10^6 - 10^8$ model parameters.

In chapter 2 I implement the bilevel multiobjective waveform inversion using a single randomly chosen shot gather at each iteration. The different random sequences of descent directions in BMWI converge to different points in the model manifold. I interpret these inverted points as the sample points of the probabilistic solution. I estimate the solution, uncertainty and sensitivity using the sample estimates of the mean, standard deviation and initial deviation of the sample points, respectively.

In chapter 3 I apply the stochastic single-shot BMWI in a 2D acoustic finite-difference approximation to the real-world pressure data acquired in a marine experiment with ocean bottom cables. I provide the estimates of the solution, uncertainty and sensitivity using different linear-gradient starting models, two differently 3D-to-2D-transformed real data sets and different random sequences of descent directions.

[14] Ari Ben-Menahem (2009) has written in his "Historical Encyclopedia of Natural and Mathematical Sciences": "The main postulate of science is the *unity of nature*: nature is one; and therefore, *science is one*. Finally, the fact that simultaneous discoveries have been made by different groups of workers, in different settings, organizations and nations, demonstrate that *mankind is one*: one mankind through one science is unfolding the mysteries of one nature."

[15] The realizability assumption in the statistical learning theory (Shalev-Shwartz and Ben-David, 2014) is formulated for the averaged quantity: the mathematical expectation of the loss function is zero.

Chapter 1

Probabilistic setting

1.1 Introduction

I present a new probabilistic setting for inverse problems and full-waveform inversion[1]. I do not use the Tarantola-Valette setting or a Bayesian framework or the integral principle of maximum entropy based on the prior information for the *averaged* quantities.

I propose the principle of minimum relative information based on the prior information for the *non-averaged* quantities. The Tarantola-Valette setting can be obtained as a special case under the assumption that the prior and theoretical measures exist. Using the *realizability assumption* as a prior information, the principle of minimum relative information leads to the parametrized volumetric measure over the model manifold with arbitrary parameters and arbitrary misfits. Maximization of the obtained volumetric measure leads to a multiobjective minimization problem with arbitrary misfits.

The formal system developed here *as any other formal system* is non-falsifiable in the Popperian sense. But it is very simple, reduces to the well-known settings in special cases and provides a useful tool for uncertainty analysis used in chapters 2 and 3 with synthetic and real-world data.

1.2 Description of a system under study

The system under study is a physical system and measuring procedure in a wide sense. A physical system is a region of space-time governed by the *unique* law of physics, i.e., the forward theory $\mathcal{F}$ equipped with the *finite* number of model $\mathbf{m}$ and data $\mathbf{d}$ parameters. The measuring procedure includes the experimental procedure as well as the computational and formal systems. A computational system is a region of space-time with an ability to compute according to the law of physics. A formal system is a set of symbols, grammar, axioms and rules of inference.

The starting point for both probabilistic and deterministic inverse problems is a geometric description of a physical system under study.

[1] I was inspired by Tarantola (2005), Vapnik (2013) and Jaynes (2003).

1.2.1 Geometric description: physical qualities and quantities

The field-theoretical formalism is a simple and elegant tool to study the physical systems. To describe a physical system, a vector field $\mathbf{f}(\mathbf{x})$ as a function of space-time coordinates $\mathbf{x}$ is introduced. The *principle of stationary action* provides a rule to find the governing law of the physical system. We have to *guess* the Lagrangian density $\mathcal{L}(\mathbf{f}(\mathbf{x}), \nabla_{\mathbf{x}}\mathbf{f}(\mathbf{x}), \mathbf{x})$ which optimizes the functional of action

$$\mathcal{S} = \int_{\mathbb{X}} \mathcal{L}(\mathbf{f}(\mathbf{x}), \nabla_{\mathbf{x}}\mathbf{f}(\mathbf{x}), \mathbf{x}) \sqrt{\det \hat{\mathbf{g}}(\mathbf{x})} d\mathbf{x}, \tag{1.1}$$

where $\hat{\mathbf{g}}(\mathbf{x})$ is a metric tensor of the space-time $\mathbb{X}$.

The first variation of action yields the partial differential Euler-Lagrange equations for the field:

$$\frac{\partial \mathcal{L}}{\partial f_k} - \frac{\partial}{\partial x_l}\left(\frac{\partial \mathcal{L}}{\partial f_{k,l}}\right) = 0, \quad k = 1, ..., N_f, \quad l = 1, ..., N_x, \tag{1.2}$$

where N_f is the number of scalar components of the vector field, $N_x = 4$ is the number of dimensions of space-time $\mathbb{X}$, the derivatives of the field $f_{k,l} = \frac{\partial f_k}{\partial x_l}$ are treated as the independent variables and the summation over the repeated index l is assumed.

Any continuous symmetry of the Lagrangian density $\mathcal{L}$ is related to a conservation law (Noether, 1918). If the Lagrangian density is invariant with respect to time translation, the energy is conserved. The invariance of the Lagrangian density under spatial translation leads to conservation of momentum. If the Lagrangian density is invariant under angular rotation about an axis, the angular momentum about the same axis is conserved.

To compute the Euler-Lagrange equations, the field is discretized. Then, the physical system is described using the *finite* number of model $\mathbf{m}$ and data $\mathbf{d}$ quantities. The forward theory is a numerical algorithm which relates the model and data quantities: $\mathcal{F}(\mathbf{m}) = \mathbf{d}$. The inverse theory $\mathcal{F}^{-1}$ is a numerical algorithm which relates the data and model quantities: $\mathcal{F}^{-1}(\mathbf{d}) = \mathbf{m}$.

The study of a physical system can be further divided into four steps.

1) *Parametrization (problem statement):* Discovery of a set of physical quantities $\{q_1, q_2, ...\}$ characterizing the system under study. These quantities are the coordinates of a point $\mathbf{q}$ in the abstract quality manifold $\mathbb{Q}$ equipped with a metric tensor $\mathbf{g}(\mathbf{q})$. The quantities $\{q_1, q_2, ...\} = \{m_1, m_2, ..., d_1, d_2, ...\}$ consist of the model $\{m_1, m_2, ...\}$ and data $\{d_1, d_2, ...\}$ quantities. The model $\{m_1, m_2, ...\}$ and data $\{d_1, d_2, ...\}$ quantities are the coordinates of the points $\mathbf{m}$ and $\mathbf{d}$ in the model $\mathbb{M}$ and data $\mathbb{D}$ manifolds, respectively.

2) *Learning (problem of induction):* Discovery of the forward theory $\mathcal{F}$ which maps the model quantities into the data quantities: $\mathcal{F}(\mathbf{m}) = \mathbf{d}$, and discovery of the inverse theory $\mathcal{F}^{-1}$ which maps the data quantities into the model quantities: $\mathcal{F}^{-1}(\mathbf{d}) = \mathbf{m}$.

3) *Modelling (forward problem):* Computing the data quantities $\mathbf{d}$ for the given model quantities $\mathbf{m}$ and forward theory $\mathcal{F}$: $\mathcal{F}(\mathbf{m}) = \mathbf{d}$.

4) *Inversion (inverse problem):* Inferring the model quantities $\mathbf{m}$ for the given data quantities $\mathbf{d}$ using the inverse theory $\mathcal{F}^{-1}$: $\mathcal{F}^{-1}(\mathbf{d}) = \mathbf{m}$.

According to Tarantola (2005), the goal of parametrization is to discover 1) the *minimal* set of parameters describing a physical system and 2) the *least*-dimensional quality manifold. This is a long-term goal. Keeping in mind an extended full-waveform inversion, it might be practically useful to consider a higher dimensional, *extended* quality manifold $\mathbb{Q}_e$ with more quantities $\mathbf{q}_e$ required to describe it.

The quality manifold $\mathbb{Q}$ is a metric manifold and has a notion of volume. The volume of any sub-volume A of $\mathbb{Q}$:

$$V(A) = \int_A dV \tag{1.3}$$

is independent of a particular choice of a coordinate system over $\mathbb{Q}$. If some coordinates $\{q_1, q_2, ..., q_n\}$ are chosen, the volume element is

$$dV = v(\mathbf{q}) dq_1 dq_2 ... dq_n, \tag{1.4}$$

where the volume density $v(\mathbf{q})$ in the given coordinates is equal to the square root of the determinant of the metric tensor $\mathbf{g}(\mathbf{q})$:

$$v(\mathbf{q}) = \sqrt{\det \mathbf{g}(\mathbf{q})}. \tag{1.5}$$

The squared distance between the point $\mathbf{q}$ with coordinates $\{q_1, q_2, ...\}$ and the point with coordinates $\{q_1 + dq_1, q_2 + dq_2, ...\}$ is

$$ds^2 = g_{ij}(\mathbf{q}) dq_i dq_j, \tag{1.6}$$

where the summation over the repeated indexes is assumed.

1.2.2 Probabilistic description: measures, information and entropy

> "My greatest concern was what to call it. I thought of calling it 'information,' but the word was overly used, so I decided to call it 'uncertainty.' When I discussed it with John von Neumann, he had a better idea. Von Neumann told me, 'You should call it entropy, for two reasons. In the first place your uncertainty function has been used in statistical mechanics under that name, so it already has a name. In the second place, and more important, no one knows what entropy really is, so in a debate you will always have the advantage.' "
>
> Claude E. Shannon to Myron Tribus
> (Tribus and McIrvine, 1971)

Any state of information or subjective degree of believe about the values of the physical quantities $\mathbf{q}$ can be described using the measure, the measure density and the volumetric measure over the quality manifold $\mathbb{Q}$.

The finite-dimensional quality manifold $\mathbb{Q}$ has subsets. A measure over the quality manifold $\mathbb{Q}$ is a *non-negative* real-valued function $M(A)$ of any subset A of $\mathbb{Q}$ which satisfies two axioms:

1. $M(\emptyset) = 0$, where $\emptyset$ is the empty set.
2. For disjoint subsets $A_1, A_2, ...$: $M(\bigcup_j A_j) = \Sigma_j M(A_j)$, where $\bigcup$ denotes the union.

The measure M can be zero only where the volume measure V is zero, i.e., the measure M is *absolutely continuous* with respect to the volume measure V, where $V(A) = \int_A dV = \int_A v(\mathbf{q}) d\mathbf{q}$.

For any measure M there exists a *non-negative*[2] function $\hat{m}(\mathbf{q})$, called the measure density, and for any subset A of $\mathbb{Q}$

$$M(A) = \int_A \hat{m}(\mathbf{q}) d\mathbf{q}. \tag{1.7}$$

The volumetric measure is $m(\mathbf{q})$ and for any subset A of $\mathbb{Q}$:

$$M(A) = \int_A m(\mathbf{q}) v(\mathbf{x}) d\mathbf{q}. \tag{1.8}$$

The measure density and the volumetric measure are related via the volume density: $\hat{m}(\mathbf{q}) = v(\mathbf{q}) m(\mathbf{q})$. The measure is a dimensionless function. The volumetric measure and the measure density have physical dimensions.

[2] Due to the Radon-Nykodim theorem (Tarantola, 2007).

Apart from the absolutely continuous measures, there exist so-called singular measures. The state of perfect knowledge on quantity $\mathbf{q}$ corresponds to the singular measure. In this case we definitely know that $\mathbf{q} = \mathbf{q}_t$. The singular measure density is:

$$\hat{h}(\mathbf{q}) = \delta(\mathbf{q}; \mathbf{q}_t). \tag{1.9}$$

The singular volumetric measure is

$$h(\mathbf{q}) = \frac{\delta(\mathbf{q}; \mathbf{q}_t)}{v(\mathbf{q})}. \tag{1.10}$$

The singular measure is

$$H(A) = \begin{cases} 1, & \mathbf{q}_t \subset A, \\ 0, & \mathbf{q}_t \not\subset A. \end{cases} \tag{1.11}$$

The state of total ignorance (also known as the reference state of information, null information or homogeneous state of information) is the uniform volumetric measure:

$$u(\mathbf{q}) = 1. \tag{1.12}$$

The uniform measure density is

$$\hat{u}(\mathbf{q}) = v(\mathbf{q}), \tag{1.13}$$

which is not necessarily constant.

The uniform measure of any subset A of $\mathbb{Q}$

$$U(A) = \int_A u(\mathbf{q}) v(\mathbf{q}) d\mathbf{q} = \int_A \hat{u}(\mathbf{q}) d\mathbf{q} \tag{1.14}$$

is equal to the volume of the subset A, but it is a dimensionless quantity.

The information (also known as surprisal and self-information) contained in the volumetric measure $\sigma(\mathbf{q})$ is

$$I(\sigma(\mathbf{q})) = -\log \sigma(\mathbf{q}). \tag{1.15}$$

To simplify the rest of this chapter, I assume that $\sigma(\mathbf{q}) \in [0, 1]$ for any $\mathbf{q} \in \mathbb{Q}$. Then $I(\sigma(\mathbf{q})) \geq 0$ for any $\sigma(\mathbf{q})$ and the minimal information is the null information $I = 0$ corresponding to the uniform measure $u(\mathbf{q}) = 1$ and the uniform measure density $\hat{u}(\mathbf{q}) = v(\mathbf{q})$.

The information contained in the measure density $\hat{\sigma}(\mathbf{q})$ is

$$I(\hat{\sigma}(\mathbf{q})) = -\log \frac{\hat{\sigma}(\mathbf{q})}{v(\mathbf{q})}. \tag{1.16}$$

The definitions satisfy the equality $I(\sigma(\mathbf{q})) = I(\hat{\sigma}(\mathbf{q}))$, but contain the logarithms of a dimensional quantity.

The entropy is a mathematical expectation of the information

$$S(\sigma) = \int v(\mathbf{q}) \sigma(\mathbf{q}) I(\sigma(\mathbf{q})) d\mathbf{q} = -\int v(\mathbf{q}) \sigma(\mathbf{q}) \log \sigma(\mathbf{q}) d\mathbf{q}, \tag{1.17}$$

or equivalently:

$$S(\hat{\sigma}) = \int \hat{\sigma}(\mathbf{q}) I(\hat{\sigma}(\mathbf{q})) d\mathbf{q} = -\int \hat{\sigma}(\mathbf{q}) \log \frac{\hat{\sigma}(\mathbf{q})}{v(\mathbf{q})} d\mathbf{q}. \tag{1.18}$$

To get rid of the problem with the logarithms of a dimensional quantity, consider the relative information contained in $\sigma(\mathbf{q})$ with respect to $u(\mathbf{q})$:

$$I(\sigma(\mathbf{q}), u(\mathbf{q})) = -\log \frac{\sigma(\mathbf{q})}{u(\mathbf{q})}. \tag{1.19}$$

The relative information contained in $\hat{\sigma}(\mathbf{q})$ with respect to $\hat{u}(\mathbf{q})$ is

$$I(\hat{\sigma}(\mathbf{q}), \hat{u}(\mathbf{q})) = -\log \frac{\hat{\sigma}(\mathbf{q})}{\hat{u}(\mathbf{q})}. \tag{1.20}$$

The relative entropy is a mathematical expectation of relative information:

$$S(\sigma, u) = \int v(\mathbf{q})\sigma(\mathbf{q})I(\sigma(\mathbf{q}), u(\mathbf{q}))d\mathbf{q} = -\int v(\mathbf{q})\sigma(\mathbf{q}) \log \frac{\sigma(\mathbf{q})}{u(\mathbf{q})} d\mathbf{q}, \tag{1.21}$$

or

$$S(\hat{\sigma}, \hat{u}) = \int \hat{\sigma}(\mathbf{q})I(\hat{\sigma}(\mathbf{q}), \hat{u}(\mathbf{q}))d\mathbf{q} = -\int \hat{\sigma}(\mathbf{q}) \log \frac{\hat{\sigma}(\mathbf{q})}{\hat{u}(\mathbf{q})} d\mathbf{q}. \tag{1.22}$$

The definitions of relative entropy contain the logarithms of a dimensionless quantity and $S(\sigma, u) = S(\hat{\sigma}, \hat{u})$.

1.3 The principle of minimum relative information

The state of information about the values of the physical quantities $\mathbf{q}$ is described by the volumetric measure $\sigma(\mathbf{q})$ and the measure density $\hat{\sigma}(\mathbf{q})$ defined over the quality manifold $\mathbb{Q}$. The principle of maximum entropy provides a rule to assign a measure density using the prior information on the averaged quantities.

It might happen that only information on a certain *non-averaged* function of quantity is available. For this case I propose the principle of minimum relative information: the current state of knowledge about the values of the physical quantities $\mathbf{q}$ corresponds to the minimum relative information constrained by prior information on the quantity.

Next, I consider two cases. First, the prior information is given in the form of an arbitrary vector function of quantity. Second, the prior information is given in the form of the two volumetric measures over the quality manifold, i.e., the prior and theoretical volumetric measures.

1.3.1 The non-averaged value of a vector function of quantity is known

First, I show what happens if, instead of the principle of minimum relative information, the principle of minimum information is used. The information $I(\sigma(\mathbf{q}))$ constrained by the prior information $\mathbf{f}(\mathbf{q}) = \mathbf{a}$, where $\mathbf{f}$ is an arbitrary vector function of $\mathbf{q}$ and $\mathbf{a}$ is a constant vector, has to be minimal. The method of Lagrange multipliers gives for any $\mathbf{q} \in \mathbb{Q}$:

$$I(\sigma(\mathbf{q})) + \mathbf{w}^\top(\mathbf{a} - \mathbf{f}(\mathbf{q})) = 0, \tag{1.23}$$

with

$$\sigma(\mathbf{q}) = e^{-\mathbf{w}^\top \mathbf{f}(\mathbf{q}) + \mathbf{w}^\top \mathbf{a}}, \tag{1.24}$$

or

$$I(\hat{\sigma}(\mathbf{q})) + \mathbf{w}^\top(\mathbf{a} - \mathbf{f}(\mathbf{q})) = 0, \tag{1.25}$$

with

$$\hat{\sigma}(\mathbf{q}) = e^{-\mathbf{w}^\top \mathbf{f}(\mathbf{q}) + \mathbf{w}^\top \mathbf{a}}. \tag{1.26}$$

These formulas cannot be correct because of the improper physical dimensions caused by the definition of information containing the logarithm of a dimensional quantity.

The principle of minimum relative information solves the problem. The relative information $I(\sigma(\mathbf{q}), u(\mathbf{q}))$ constrained by prior information has to be minimal. The method of Lagrange multipliers gives for any $\mathbf{q} \in \mathbb{Q}$:

$$I(\sigma(\mathbf{q}), u(\mathbf{q})) + \mathbf{w}^\top(\mathbf{a} - \mathbf{f}(\mathbf{q})) = 0, \tag{1.27}$$

with

$$\frac{\sigma(\mathbf{q})}{u(\mathbf{q})} = e^{-\mathbf{w}^\top \mathbf{f}(\mathbf{q}) + \mathbf{w}^\top \mathbf{a}}, \tag{1.28}$$

or

$$I(\hat{\sigma}(\mathbf{q}), \hat{u}(\mathbf{q})) + \mathbf{w}^\top(\mathbf{a} - \mathbf{f}(\mathbf{q})) = 0, \tag{1.29}$$

with

$$\frac{\hat{\sigma}(\mathbf{q})}{\hat{u}(\mathbf{q})} = e^{-\mathbf{w}^\top \mathbf{f}(\mathbf{q}) + \mathbf{w}^\top \mathbf{a}}, \tag{1.30}$$

where no partition function is involved in contrast to formula (11) in the introduction given by the principle of maximum entropy with prior information on the averaged quantity. The left and right sides in formulas (1.28) and (1.30) are dimensionless.

Assume the scalar function $f(\mathbf{q})$ is equal to the energy $E(\mathbf{q})$ or Hamiltonian $H(\mathbf{q})$ of the system. The principle of minimum relative information constrained by the condition of zero energy or zero Hamiltonian leads to:

$$\frac{\sigma(\mathbf{q})}{u(\mathbf{q})} = \frac{\hat{\sigma}(\mathbf{q})}{\hat{u}(\mathbf{q})} = e^{-wE(\mathbf{q})} = e^{-wH(\mathbf{q})}. \tag{1.31}$$

If $w = (k_B T)^{-1}$, where k_B is the Boltzmann constant and T denotes the temperature, formula (1.31) is the starting point for the simulated annealing algorithm (Geman and Geman, 1984; Kirkpatrick et al., 1983) and for the Hamiltonian Monte Carlo methods (Fichtner et al., 2019; Neal, 2011).

1.3.2 The theoretical and prior measures exist

As previously, I show what happens if, instead of the principle of minimum relative information, the principle of minimum information is used. The *principle of minimum information* leads to the Tarantola-Vallete setting under a very simple constraint. It is sufficient to assume the existence of the theoretical Θ and prior ρ volumetric measures. The method of Lagrange multipliers gives for any $\mathbf{q} \in \mathbb{Q}$

$$I(\sigma(\mathbf{q})) - w_1 I(\Theta(\mathbf{q})) - w_2 I(\rho(\mathbf{q})) = 0. \tag{1.32}$$

Then

$$\sigma(\mathbf{q}) = \Theta^{w_1}(\mathbf{q}) \rho^{w_2}(\mathbf{q}) \tag{1.33}$$

and

$$\hat{\sigma}(\mathbf{q}) = \frac{\hat{\Theta}^{w_1}(\mathbf{q}) \hat{\rho}^{w_2}(\mathbf{q})}{v^{w_1 + w_2 - 1}(\mathbf{q})}, \tag{1.34}$$

where the choice of weights w_1 and w_2 is required to specify the setting.

The Tarantola-Valette setting is a consequence of additivity of information with the equal weights $w_1 = w_2 = 1$:

$$\sigma(\mathbf{q}) = \Theta(\mathbf{q}) \rho(\mathbf{q}) \tag{1.35}$$

and

$$\hat{\sigma}(\mathbf{q}) = \frac{\hat{\Theta}(\mathbf{q})\hat{\rho}(\mathbf{q})}{v(\mathbf{q})}. \tag{1.36}$$

The *principle of minimum relative information* leads to the slightly different formulas under the same assumptions. The method of Lagrange multipliers gives for any $\mathbf{q} \in \mathbb{Q}$:

$$I(\sigma(\mathbf{q}), u(\mathbf{q})) - w_1 I(\Theta(\mathbf{q}), u(\mathbf{q})) - w_2 I(\rho(\mathbf{q}), u(\mathbf{q})) = 0, \tag{1.37}$$

with

$$\sigma(\mathbf{q}) = \frac{\Theta^{w_1}(\mathbf{q})\rho^{w_2}(\mathbf{q})}{u^{w_1+w_2-1}(\mathbf{q})} \tag{1.38}$$

and

$$\hat{\sigma}(\mathbf{q}) = \frac{\hat{\Theta}^{w_1}(\mathbf{q})\hat{\rho}^{w_2}(\mathbf{q})}{\hat{u}^{w_1+w_2-1}(\mathbf{q})}. \tag{1.39}$$

The Tarantola-Valette setting corresponds to the equal weights $w_1 = w_2 = 1$:

$$\sigma(\mathbf{q}) = \frac{\Theta(\mathbf{q})\rho(\mathbf{q})}{u(\mathbf{q})} \tag{1.40}$$

and

$$\hat{\sigma}(\mathbf{q}) = \frac{\hat{\Theta}(\mathbf{q})\hat{\rho}(\mathbf{q})}{\hat{u}(\mathbf{q})}. \tag{1.41}$$

The original formula (5) in the introduction for the a-posteriori volumetric measure in the Tarantola-Valette setting squares the physical dimension of the volumetric measure. The principle of minimum information with the incorrect logarithms of a dimensional quantity leads to the original formula.

The principle of minimum relative information solves the problem: the physical dimensions of the left and right sides in equation (1.40) are the same. The formulas (1.35) and (1.40) do not differ quantitatively, because $u(\mathbf{q}) = 1$.

The choice of weights $w_1 + w_2 = 1$ provides an *alternative setting* with an interesting interpretation. The identical states of information $\Theta(\mathbf{q}) = \rho(\mathbf{q})$ are combined without the change of the state of information $\sigma(\mathbf{q}) = \Theta(\mathbf{q}) = \rho(\mathbf{q})$. In general, if $w_1 + w_2 = 1$, then

$$\sigma(\mathbf{q}) = \Theta^{w_1}(\mathbf{q})\rho^{w_2}(\mathbf{q}) \tag{1.42}$$

and

$$\hat{\sigma}(\mathbf{q}) = \hat{\Theta}^{w_1}(\mathbf{q})\hat{\rho}^{w_2}(\mathbf{q}). \tag{1.43}$$

In the introduction to this work I proposed to use a square root operation for the idempotent conjunction $\wedge$ as a consequence of the formulas (1.42) and (1.43).

The principle of minimum relative information always leads to the same formulas for the volumetric measure normalized by the uniform volumetric measure and for the measure density normalized by the uniform measure density because

$$\frac{\sigma(\mathbf{q})}{u(\mathbf{q})} = \frac{\hat{\sigma}(\mathbf{q})}{\hat{u}(\mathbf{q})}. \tag{1.44}$$

1.4 Quest for the law

The goal of full-waveform inversion is to find the unique governing law of a physical system under study. I assume the unique governing law exists: there exists the *ideal* description of a physical

system and measuring procedure with the finite set of parameters $\hat{\mathbf{q}} = \{\hat{\mathbf{m}}, \hat{\mathbf{d}}\}$ and the forward theory $\hat{\mathcal{F}}$.

My research hypothesis is the *realizability assumption*: the law $\hat{\mathcal{F}}(\hat{\mathbf{m}})$ explains all subsets of the measured data parameters $\hat{\mathbf{d}}_o$ and predicts all subsets of the unmeasured (in the given experiment) data parameters $\hat{\mathbf{d}}_u$, so that $\hat{\mathcal{F}}(\hat{\mathbf{m}}) = \hat{\mathbf{d}}$, where $\hat{\mathbf{d}} = \{\hat{\mathbf{d}}_o, \hat{\mathbf{d}}_u\}$.

To find the law, I guess a forward theory $\mathcal{F}_p$ and a finite set of parameters $\mathbf{q}_p = \{\mathbf{m}_p, \mathbf{d}_p\}$, measure some of the data parameters $\mathbf{d}_o$ and check the research hypothesis.

To assign the volumetric measure over the model manifold, I use the principle of minimum relative information constrained by the realizability assumption. Maximization of the obtained volumetric measure leads to a multiobjective minimization problem with arbitrary misfits.

1.4.1 Realizability assumption: the law predicts all subsets of the data parameters

If the forward theory $\mathcal{F}_p(\mathbf{m})$ explains *any subset of the measured data parameters* $\mathbf{d}_o$ and predicts *any subset of the unmeasured (in the given experiment) data parameters* $\mathbf{d}_u$ which form $\mathbf{d}_p = \{\mathbf{d}_o, \mathbf{d}_u\}$, then the point $\mathbf{d}_p$ in $\mathbb{D}$ and the point $\mathcal{F}_p(\mathbf{m})$ in $\mathbb{D}$ are identical. *For any choice of a metric* ℓ the identity of indiscernibles holds:

$$\ell(\mathbf{d}_p, \mathcal{F}_p(\mathbf{m})) = 0 \quad \Leftrightarrow \quad \mathbf{d}_p = \mathcal{F}(\mathbf{m}). \tag{1.45}$$

If the data manifold $\mathbb{D}$ has N_d dimensions, there exist $2^{N_d} - 1$ non-empty subsets of the data quantities $\{d_1, ..., d_{N_d}\}$ which specify the coordinates of the $2^{N_d} - 1$ points $\mathbf{d}_p^k$ ($k = 1, ..., 2^{N_d} - 1$) in the corresponding $2^{N_d} - 1$ data sub-manifolds $\mathbb{D}^k$ ($k = 1, ..., 2^{N_d} - 1$).

For any choice of a metric ℓ the $2^{N_d} - 1$ equalities $\ell(\mathbf{d}_p^k, \mathcal{F}_p^k(\mathbf{m})) = 0$ have to be satisfied. The principle of minimum relative information constrained by these equalities leads to

$$\frac{\sigma(\mathbf{m}, \mathbf{w})}{u(\mathbf{m})} = \frac{\hat{\sigma}(\mathbf{m}, \mathbf{w})}{\hat{u}(\mathbf{m})} = e^{-\sum\limits_{k=1}^{N_\alpha} w_k \ell(\mathbf{d}_p^k, \mathcal{F}_p^k(\mathbf{m}))}, \qquad N_\alpha = 2^{N_d} - 1, \tag{1.46}$$

where N_α is the number of non-empty subsets of the data quantities. In contrast to the principle of maximum entropy, where the parameters $\mathbf{w}$ can be obtained by differentiating the logarithm of the partition function (formula (13) in the introduction), the parameters $\mathbf{w}$ in formula (1.46) can be arbitrary.

Instead of using the misfits between all data parameters $\mathbf{d}_p$, formula (1.46) can be reduced to the misfits between the measured data parameters $\mathbf{d}_o = \{d_1, ..., d_{N_o}\}$ and $\mathcal{F}_o(\mathbf{m})$, where N_o is the number of the measured data parameters:

$$\frac{\sigma(\mathbf{m}, \mathbf{w})}{u(\mathbf{m})} = \frac{\hat{\sigma}(\mathbf{m}, \mathbf{w})}{\hat{u}(\mathbf{m})} = e^{-\sum\limits_{k=1}^{N_\beta} w_k \ell(\mathbf{d}_o^k, \mathcal{F}_o^k(\mathbf{m}))}, \qquad N_\beta = 2^{N_o} - 1. \tag{1.47}$$

The realizability assumption leads to the volumetric measure $\sigma(\mathbf{m}, \mathbf{w})$ parametrized with the arbitrary $\mathbf{w} = \{w_1, w_2, ...\}$ and arbitrary misfits ℓ. For any fixed parameters $\mathbf{w}$ and misfits ℓ formula (1.47) can be used to sample the solution in the specified setting.

For example, $\sigma(\mathbf{m}, \mathbf{w})$ can be sampled by the rejection algorithm for the parameters $\sum_{k=1}^{N_\beta} w_k \gg 1$ with an arbitrary misfit ℓ. If the samples of the uniform volumetric measure $u(\mathbf{m})$ are generated, then the rejection algorithm accepts with the probability one as the sample points those points in $\mathbb{M}$ which correspond to the zero misfits. The points in $\mathbb{M}$ corresponding to the non-zero misfit $\ell(\mathbf{m})$ are accepted with the negligible probability $e^{-w\ell(\mathbf{m})}$, where $w \gg 1$ and $\ell(\mathbf{m}) > 0$. If all points are rejected, i.e., if there does not exist even a single point $\mathbf{m}_o$ in the model manifold which satisfies the realizability assumption $\mathcal{F}_o(\mathbf{m}_o) = \mathbf{d}_o$, then the theory, the model parameters and the data parameters are *inconsistent*. In this case the choice of parameters $\mathbf{w}$ can be relaxed $\sum_{k=1}^{N_\beta} w_k = 1$, then the rejection algorithm accepts the points in the model manifold which are *partially consistent* with the theory and observations. In the worse case of zero parameters $w_k = 0$ for $k = 1, ..., N_\beta$ ($\sum_{k=1}^{N_\beta} w_k = 0$), the volumetric measure $\sigma(\mathbf{m}, \mathbf{w})$ is equal to the uniform volumetric measure $u(\mathbf{m})$, i.e., the prior information on the theory $\mathcal{F}_o$ and observations $\mathbf{d}_o$ is ignored.

If only two data parameters d_o^1 and d_o^2 are measured, there exist three non-empty subsets of the data parameters: $\{d_o^1\}$, $\{d_o^2\}$ and $\{d_o^1, d_o^2\}$. Excluding for simplicity the subset $\{d_o^1, d_o^2\}$, the parametric volumetric measure is $\sigma(\mathbf{m}, \mathbf{w}) = u(\mathbf{m}) \exp\big(-w_1 \ell(d_o^1, \mathcal{F}_o(\mathbf{m})) - w_2 \ell(d_o^2, \mathcal{F}_o(\mathbf{m}))\big)$. Figure 1.1 schematically illustrates the 2D parametric volumetric measure $\sigma(\mathbf{m}, \mathbf{w})$ for different values of the weights w_1 and w_2. If the weights are decreasing simultaneously $w_1 = w_2 \to 0$, the parametric volumetric measure $\sigma(\mathbf{m}, \mathbf{w})$ is becoming more and more uniform $u(\mathbf{m})$. If the weights are increasing simultaneously $w_1 = w_2 \to +\infty$, the parametric volumetric measure $\sigma(\mathbf{m}, \mathbf{w})$ is becoming more and more singular. If $w_1 + w_2 = 1$, the parametric volumetric measure $\sigma(\mathbf{m}, \mathbf{w})$ varies from $\sigma(\mathbf{m}, w_1 = 1, w_2 = 0) = u(\mathbf{m}) \exp\big(-\ell(d_o^1, \mathcal{F}_o(\mathbf{m}))\big)$ through $\sigma(\mathbf{m}, w_1 = 0.5, w_2 = 0.5) = u(\mathbf{m}) \exp\big(-0.5\ell(d_o^1, \mathcal{F}_o(\mathbf{m})) - 0.5\ell(d_o^2, \mathcal{F}_o(\mathbf{m}))\big)$ to $\sigma(\mathbf{m}, w_1 = 0, w_2 = 1) = u(\mathbf{m}) \exp\big(-\ell(d_o^2, \mathcal{F}_o(\mathbf{m}))\big)$. Only if $\ell(d_o^1, \mathcal{F}_o(\mathbf{m})) = \ell(d_o^2, \mathcal{F}_o(\mathbf{m}))$ for all $\mathbf{m}$ in $\mathbb{M}$, the parametric volumetric measure $\sigma(\mathbf{m}, \mathbf{w})$ is independent of the choice of parameters w_1 and w_2 satisfying $w_1 + w_2 = 1$.

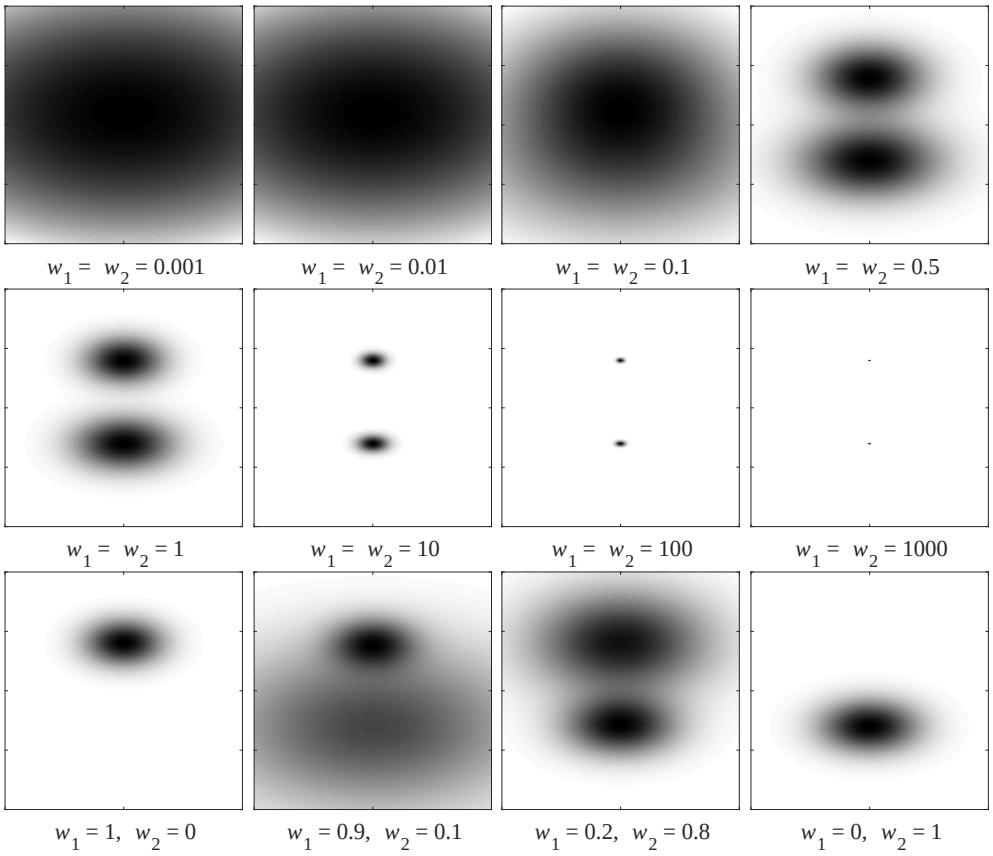


FIGURE 1.1: Schematic illustration of the 2D parametric volumetric measure $\sigma(\mathbf{m}, \mathbf{w}) = u(\mathbf{m}) \exp\big(-w_1 \ell(d_o^1, \mathcal{F}_o(\mathbf{m})) - w_2 \ell(d_o^2, \mathcal{F}_o(\mathbf{m}))\big)$ for different weights w_1 and w_2.

Entropy, risk and empirical risk minimization

The parametrized relative entropy of $\hat{\sigma}$ over the model manifold $\mathbb{M}$ is

$$S(\hat{\sigma}, \hat{u}, \mathbf{w}) = \int_{\mathbb{M}} \hat{u}(\mathbf{m}) \sum_{k=1}^{N_\beta} w_k \ell(\mathbf{d}_o^k, \mathcal{F}_o^k(\mathbf{m})) e^{-\sum_{k=1}^{N_\beta} w_k \ell(\mathbf{d}_o^k, \mathcal{F}_o^k(\mathbf{m}))} d\mathbf{m}. \tag{1.48}$$

Formula (1.48) is nothing but a mathematical expectation of the weighted misfit $\ell(\mathbf{w}, \mathbf{m}) = \sum_{k=1}^{N_\beta} w_k \ell(\mathbf{d}_o^k, \mathcal{F}_o^k(\mathbf{m}))$:

$$R(\mathbf{w}) = \int_{\mathbb{M}} \ell(\mathbf{w}, \mathbf{m}) d\hat{\sigma}(\mathbf{w}, \mathbf{m}), \qquad d\hat{\sigma}(\mathbf{m}, \mathbf{w}) = \hat{u}(\mathbf{m}) e^{-\ell(\mathbf{w}, \mathbf{m})} d\mathbf{m}, \tag{1.49}$$

known as the risk function in statistical learning theory (Shalev-Shwartz and Ben-David, 2014).

The principle of minimum relative information constrained by the realizability assumption leads to the same expression for the relative entropy and the risk function: $S(\hat{\sigma}, \hat{u}, \mathbf{w}) = R(\mathbf{w})$.

The risk function (1.49) can be approximated by the empirical risk:

$$\hat{R}(\mathbf{w}) = \sum_{k=1}^{N_\alpha} w_k \ell(\mathbf{d}_p^k, \mathcal{F}_p^k(\mathbf{m}_k)), \qquad N_\alpha = 2^{N_d} - 1, \tag{1.50}$$

where $\mathbf{m}_k$ are the sample points of $\hat{\sigma}(\mathbf{m})$.

The empirical risk minimization with $w_k = 1/N_\alpha$ ($k = 1, ..., N_\alpha$) is the main inductive principle for learning a theory $\mathcal{F}_p$ in the statistical learning theory (Vapnik, 2006, 1992, 2013).

1.4.2 Deterministic criterion: multiobjective optimization

Maximization of the parametrized volumetric measure (1.46) leads to the scalarized multiobjective minimization problem with the arbitrary misfits

$$\ell(\mathbf{m}, \mathbf{w}) = \sum_{k=1}^{N_\alpha} w_k \ell(\mathbf{d}_p^k, \mathcal{F}_p^k(\mathbf{m})), \qquad N_\alpha = 2^{N_d} - 1. \tag{1.51}$$

Many *solution concepts* exist in non-linear multiobjective optimization (Deb, 2001; Miettinen, 1999). I specify only the non-dominated and strictly-dominated optima.

Non-dominated solutions: $\mathbf{m}^o$ is a Pareto optimum (efficient or non-dominated point) of a multiobjective misfit $\ell = \{\ell_1, ..., \ell_{N_\alpha}\}$ if there does not exist another point $\mathbf{m}^a$ such that $\ell_i(\mathbf{m}^a) \leq \ell_i(\mathbf{m}^o)$ for all $i = 1, ..., N_\alpha$ and $\ell_j(\mathbf{m}^a) < \ell_j(\mathbf{m}^o)$ for at least one index $j = 1, ..., N_\alpha$, i.e., there does not exist another point $\mathbf{m}^a$ which improves at least one misfit without worsening all the other misfits.

Strictly-dominated solutions: $\mathbf{m}_k^o$ ($k = 1, ..., N_\alpha$) is a strictly-dominated optimum of the k-th misfit ℓ_k if there does not exist another model $\mathbf{m}^a$ such that $\ell_k(\mathbf{m}^a) < \ell_k(\mathbf{m}^o)$, i.e., every misfit function is considered individually.

If the definitions are specified in the whole model manifold $\mathbb{M}$, an optimum is global. If the definitions are specified only in a small neighbourhood of $\mathbf{m}^o$, an optimum is local.

The non-dominated points in $\mathbb{M}$ are the points with the highest values of the volumetric measure (1.47). The global (or local) Pareto optima of the multiobjective minimization problem (1.51) are the sample points of $\sigma(\mathbf{m}, \mathbf{w})$ in the regions with the highest (or locally highest) values of the volumetric measure $\sigma(\mathbf{m}, \mathbf{w})$.

If the misfits are *conflicting*, i.e., the theory, the model parameters and the data parameters are *inconsistent*, there *does not exist a point* in the model manifold $\mathbb{M}$, so that all misfits become zero simultaneously. In this case many non-dominated and many strictly-dominated optima exist.

If the misfits are *non-conflicting*, i.e., the theory, the model parameters and the data parameters are *consistent*, all misfits can be zero simultaneously *at least at one point* of the model manifold $\mathbb{M}$.

1.5 Solution and uncertainty

If $\sigma(\mathbf{m}, \mathbf{w})$ is sampled, the sample mean and sample covariance matrix (or standard deviation) are the estimates of the solution and its uncertainty. If N points $\mathbf{m}_k$ ($k = 1, ..., N$) in $\mathbb{M}$ are sampled, their sample mean is

$$\langle \mathbf{m} \rangle = \frac{1}{N} \sum_{k=1}^{N} \mathbf{m}_k, \tag{1.52}$$

and their sample covariance matrix is

$$\langle \mathbf{C} \rangle = \frac{1}{(N-1)} \sum_{k=1}^{N} (\mathbf{m}_k - \langle \mathbf{m} \rangle)(\mathbf{m}_k - \langle \mathbf{m} \rangle)^\top. \tag{1.53}$$

The sample standard deviation is

$$\mathbf{V} = \sqrt{\frac{1}{N-1}\sum_{k=1}^{N}(\mathbf{m}_k - \langle\mathbf{m}\rangle)^2}, \tag{1.54}$$

where the operations are understood in a component-wise sense.

If the governing law of a physical system is deterministic and the sample mean and sample covariance matrix (or standard deviation) do not provide an adequate description of the solution and its uncertainty (for example, if $\sigma(\mathbf{m},\mathbf{w})$ is multimodal), then the measuring procedure has to be improved. For the multimodal $\sigma(\mathbf{m},\mathbf{w})$ the sample points can be divided into the clusters of points and in each cluster the sample mean and sample standard deviation can be estimated.

The regions corresponding to the significant values of the volumetric measure $\sigma(\mathbf{m},\mathbf{w})$ can be sampled by solving the global multiobjective minimization problem (1.47). Unfortunately, it is computationally intractable. Even solving the local multiobjective minimization problem with $10^6 - 10^8$ unknown parameters in FWI is computationally expensive.

In chapters 2 and 3 I show that it is sometimes sufficient to find just two-three arbitrary optima of the local multiobjective minimization problem. Their sample mean and sample standard deviation are the computationally inexpensive estimates of the solution and uncertainty.

1.6 Summary

I proposed a new probabilistic setting for inverse problems and full-waveform inversion based on the principle of minimum relative information. In contrast to the principle of maximum entropy, the prior information is given for non-averaged quantities.

I formulated the quest for the governing law of a physical system using the principle of minimum relative information and the realizability assumption. According to the realizability assumption, the governing law explains all subsets of the measured data parameters and predicts all subsets of the unmeasured (in the given experiment) data parameters. This leads to the parametrized family of the volumetric measures with the arbitrary parameters $\mathbf{w}$ and arbitrary misfits ℓ.

Maximization of the parametric volumetric measure leads to a multiobjective minimization problem with arbitrary misfits. If the misfits are conflicting, they cannot be zero simultaneously at any point of the model manifold. Then the theory, the model parameters and the data parameters are inconsistent and there exist many Pareto optimal solutions which sample the estimate of the solution and uncertainty in the specified setting. If the misfits are non-conflicting, they can be zero simultaneously at least at one point of the model manifold. Then the theory, the model parameters and the data parameters are consistent and the points in the model manifold, corresponding to the zero misfits, sample the estimate of the solution and uncertainty in the specified setting.

In chapter 2 I illustrate these ideas with the 2D acoustic wave equation for the Marmousi-2 model. In chapter 3 I apply these ideas to real-field data acquired in a marine seismic experiment.

Chapter 2

Stochastic algorithm

2.1 Introduction

Under the realizability assumption for any guessed forward theory $\mathcal{F}_p$ and parametrization $\mathbf{q} = \{\mathbf{m}, \mathbf{d}\}$, the solution of an inverse problem is the parametric volumetric measure $\sigma(\mathbf{m}, \mathbf{w})$. Maximization of $\sigma(\mathbf{m}, \mathbf{w})$ leads to a multiobjective minimization problem.

In this chapter I present a simple algorithm of full-waveform inversion in the context of marine exploration experiments to sample at least a few points in the model manifold according to $\sigma(\mathbf{m}, \mathbf{w})$ by partially solving the multiobjective minimization problem.

First, I specify the physical system under study: a marine exploration experiment with ocean bottom cables. Neglecting many phenomena in real Earth, the physical system is described with the 2D time-domain acoustic wave equation solved using a time-domain finite-difference algorithm.

Next, I describe a stochastic single-shot bilevel multiobjective waveform inversion algorithm (BMWI). As the estimates of the solution and uncertainty I use the sample mean and sample standard deviation of the model parameters inverted with the different randomized BMWI algorithms. To estimate the sensitivity of the measuring procedure, I average the initial deviation of the inverted models over the different starting models.

To test the algorithm, I present a few numerical experiments and provide the sample estimates of the mean values, standard deviation and initial deviation of the model parameters inverted by the stochastic single-shot BMWI.

2.2 Setting

2.2.1 Physical system, experimental procedure and continuous description

The main goal of marine seismic exploration experiments is to locate hydrocarbons. The physical system under study is a 3D marine exploration experiment with ocean bottom cables. The region of interest here is a part of the Earth which is less than $15 \times 15 \times 10$ km ($x \times y \times z$) in size. An ocean

bottom cable is around 5-10 km long and contains 4-component receivers every 12.5-50 meters (usually 25 m) which record 6-12 s (usually 9 s) of data with sampling 1-2 ms. 4-component data contain 3 components of the vector of particle velocities $\mathbf{v}(\mathbf{x},t)$ (or its time derivative) and scalar pressure $p(\mathbf{x},t)$, where $\mathbf{x}$ and t are the space and time coordinates. During an experiment 10-50 cables are placed at the bottom of the ocean at depths 100-5000 m with inter-cable spacing around 100-500 m. Both cable deployment and source shooting are usually performed by a seismic vessel of length 50-150 m. The air-gun sources are towed behind the seismic vessel and excite air bubbles in water. A single source contains 3-5 gun strings with 10-20 air-guns in each array with spacing between air-guns 2-10 m. A single source contains many air-guns and can cover the area 100-500 m^2. Each air-gun has approximately cylindrical shape with the length $0.3-1$ m, diameter $0.2-0.3$ m and weight $20-100$ kg. If the spatial scale of air-guns (roughly 30 m) is relatively small in comparison with the wavelengths of sound fields, the set of air-guns can be modelled as a spatial point source. The source directivity function depends entirely on the setup of the gun array: it can be spherical at low frequencies (less than $20-30$ Hz) and ellipsoidal at higher frequencies with the shortest semi-axis in the direction perpendicular to the shot line in the plane of the ocean surface. The source time function is not known and is usually repeatable for different shots. The recorded data at frequencies below 3 Hz and above 100 Hz is often noisy and not reliable. The marine exploration experiment is performed during one or even several months. During that time multi-stage quality control of the data is performed. Many factors can affect the quality of the data: weather conditions, other vessels around (fishing activity or other acquisition), ocean fauna (in particular, mammals), problems with sources and receivers. The raw data contain N_s shot gathers with N_r traces, where N_s is the number of shots and N_r denotes the number of receivers. The shot gather or seismogram contains the direct, reflected, refracted and guided waves caused by the spatial variations of elastic properties of the Earth. To be more precise, the boundaries of the physical system under study are not known: the system is open and interacts with the other systems. The imperfections of instruments (receivers and sources) can be sometimes described mathematically, but usually they are simply neglected.

To describe the experiment above, I use a 2D[1] linear isotropic acoustic equation[2] with the spatially-variable sound velocity $V(\mathbf{x})$ and mass density $\rho(\mathbf{x})$, where $\mathbf{x}\in\mathbb{R}^2$

$$\frac{\partial^2 p(\mathbf{x},t)}{\partial t^2} - V^2(\mathbf{x})\rho(\mathbf{x})\frac{\partial}{\partial x_k}\left(\frac{1}{\rho(\mathbf{x})}\frac{\partial p(\mathbf{x},t)}{\partial x_k}\right) = s(\mathbf{x},t), \tag{2.1}$$

where pressure $p=p(\mathbf{x},t)$ is a space- and time-dependent field, $s=s(\mathbf{x},t)$ is the space-time source of disturbances and the summation over the repeated index $k=1,2$ is assumed.

The forward problem (2.1) is complemented by initial and boundary conditions. The causal initial condition is $p(\mathbf{x},t)=\frac{\partial p(\mathbf{x},t)}{\partial t}=0$ for all $\mathbf{x}$ at $t\le 0$. At the top of a model a free-surface condition is assumed: $p(\mathbf{x}_f,t)=0$ for all times at $\mathbf{x}_f$ corresponding to the free surface. At the left, right and bottom boundaries the non-reflecting conditions are assumed to mimic a semi-infinite medium.

2.2.2 Numerical forward modeling: time-domain finite differences

To solve the problem (2.1) numerically, consider an equivalent system of the three first-order equations:

$$\frac{\partial v_k}{\partial t} = -\frac{1}{\rho(\mathbf{x})}\frac{\partial p}{\partial x_k}, \quad k=1,2, \tag{2.2}$$

$$\frac{\partial p}{\partial t} = -V^2(\mathbf{x})\rho(\mathbf{x})\frac{\partial v_k}{\partial x_k} + \hat{s}(\mathbf{x},t), \tag{2.3}$$

where $\hat{s}(\mathbf{x},t)=\int s(\mathbf{x},t)dt$ with a source term $s(\mathbf{x},t)$ from equation (2.1). The equations (2.2)-(2.3) are solved using a standard explicit staggered-grid finite-difference method (Virieux, 1986, 1984). The time axis t_n is discretized with $n=1:N_t$ and regular spacing dt. The spatial coordinates x_l and z_k are discretized with $l=1:N_x$ and $k=1:N_z$ and spacings $dx=dz=dh$. The second-order in time and eight-order in space finite-difference operators are used with the truncation error $\mathcal{O}(dt^2,dh^8)$.

[1] 3D FWI of OBC-data is also feasible (Borisov and Singh, 2015; Operto et al., 2015), but is beyond the scope of this work.
[2] The corresponding Lagrangian density can be found in Ben-Menahem and Singh (1981).

The accuracy and convergence of the finite-difference algorithm depends strongly on the choice of dt, dh and the Courant number $C = V\frac{dt}{dh}$. To keep the grid dispersion minimal, the following inequality must be satisfied:

$$dh \leq \frac{V_{min}}{5f_{max}}, \tag{2.4}$$

where V_{min} denotes the minimal value of sound velocity in the whole grid and f_{max} denotes the maximal frequency content of the fields propagating in the medium.

To avoid instability of the finite-difference algorithm, the temporal sampling must satisfy the inequality:

$$dt \leq \frac{1680dh}{2161\sqrt{2}V_{max}}, \tag{2.5}$$

where V_{max} denotes the maximal value of sound velocity at the whole grid and the specific factors are related to the choice of the 2×8 order Taylor finite-difference operators (Köhn, 2011). The inequality (2.5) is called the Courant-Friedrichs-Lewy criterion.

To speed the performance of the algorithm, a parallelization by domain decomposition is used (Bohlen, 2002; Köhn, 2011; Kurzmann, 2012). The CPUs communicate using the Message Passing Interface (MPI).

At the top of a computational domain the free-surface condition is implemented using the mirroring technique (Levander, 1988). At the left, right and bottom boundaries the unsplit convolutional perfectly matched layers proposed by Komatitsch and Martin (2007) are used.

As soon as the set of model parameters $\mathbf{m}$ (discretized sound velocity $V(\mathbf{x})$ and mass density $\rho(\mathbf{x})$) and the discretized source term $\mathbf{s}(\mathbf{x}, t)$ are specified, the data parameters $\mathbf{d}$ (discretized pressure field $p(\mathbf{x}, t)$ and particle velocities $v_1(\mathbf{x}, t)$ and $v_2(\mathbf{x}, t)$) can be computed in the whole computational domain corresponding to the discretized space-time region of interest. The discretized source term $\mathbf{s}$ is not measured in the experiment and can be interpreted as either a part of the data parameters $\mathbf{d}$ or a part of the model parameters $\mathbf{m}$.

For any model parameters $\mathbf{m}$ the forward solver $\mathcal{F}$ computes the data parameters $\mathbf{d}$:

$$\mathcal{F}(\mathbf{m}) = \mathbf{d}. \tag{2.6}$$

The modelled data parameters $\mathbf{d}$ can be compared with the observed parameters $\mathbf{d}_o$ using an arbitrary misfit function ℓ. If the misfit is zero, our guess $\mathcal{F}(\mathbf{m})$ is a candidate for the law of the physical system under study.

If the difference between the data parameters is not zero for any non-empty subset of the data, then the guessed law is not the law. To falsify $\mathcal{F}(\mathbf{m})$, it is already sufficient to falsify it at any subset of the data.

2.2.3 Stochastic inversion: bilevel multiobjective waveform inversion

Maximization of the probabilistic parametrized solution (1.46) is equivalent to a multiobjective minimization problem (1.51). If the source time functions $\mathbf{s}$ are considered as model parameters, then the multiobjective problem (1.51) can be reformulated as a bilevel multiobjective problem with the two types of model parameters $\mathbf{s}$ and $\mathbf{m}$: $\mathcal{F}_o(\mathbf{m}, \mathbf{s}) = \mathbf{d}_o$.

I formulate the bilevel multiobjective waveform inversion (BMWI) as a bilevel multiobjective least-squares problem (Bard, 1998; Dempe, 2002; Eichfelder, 2010):

$$\min_{\mathbf{m}, \mathbf{s} \in \mathbf{s}^o} \boldsymbol{\ell}_2^2(\mathbf{d}_o, \mathcal{F}_o(\mathbf{m}, \mathbf{s})), \tag{2.7}$$

$$\mathbf{s}^o = \arg\min \boldsymbol{\ell}_2^2(\mathbf{d}_o, \mathcal{F}_o(\mathbf{m}, \mathbf{s})), \tag{2.8}$$

where the multiobjective misfit function $\boldsymbol{\ell}_2^2(\mathbf{d}_o, \mathcal{F}_o(\mathbf{m}, \mathbf{s})) = \{\ell_2^2(\mathbf{d}_o^1, \mathcal{F}_o(\mathbf{m}, \mathbf{s}^1)), ..., \ell_2^2(\mathbf{d}_o^{N_s}, \mathcal{F}_o(\mathbf{m}, \mathbf{s}^{N_s}))\}$ is formed by the N_s single-shot misfit functions $\ell_2^2(\mathbf{d}_o^k, \mathcal{F}_o(\mathbf{m}, \mathbf{s}^k))$ $(k = 1, ..., N_s)$ and N_s is the number of shot gathers.

The upper-level multiobjective problem (2.7) with respect to $\mathbf{s}$ and $\mathbf{m}$ has to be optimized only for those $\mathbf{s}$ which optimize the lower-level multiobjective problem (2.8). In other words, we have to invert for $\mathbf{m}$ only with the optimal source time functions $\mathbf{s}_o$.

The corresponding scalarized problem with linear weights is

$$\min_{\mathbf{m},\mathbf{s}\in\mathbf{s}^o} \ell(\mathbf{d}_o, \mathcal{F}_o(\mathbf{m},\mathbf{s}), \mathbf{w}) = \sum_{k=1}^{N_s} w_k \ell_2^2(\mathbf{d}_o^k, \mathcal{F}_o(\mathbf{m},\mathbf{s}^k)), \tag{2.9}$$

$$\mathbf{s}^o = \arg\min \left(\ell_2^2(\mathcal{F}_o(\mathbf{m},\mathbf{s}), \mathbf{w})) = \sum_{k=1}^{N_s} w_k \ell_2^2(\mathbf{d}_o^k, \mathbf{m}, \mathbf{s}^k) \right), \tag{2.10}$$

where the weighting parameters $\{w_1, ..., w_{N_s}\}$ are non-negative and $\mathbf{w} \neq 0$. This scalarized problem is still the non-linear multiobjective minimization problem (1.51) with respect to the model parameters $\mathbf{m}$, but considered only for the optimal values of the source time functions.

I use only a single randomly chosen shot gather at each iteration. This reduces the cost of computing a descent direction and the cross-talks between the different shot gathers appearing during the summation of gradients of a misfit function. Choosing a single k-th shot gather per iteration means setting to zero the weights w_j corresponding to all other shot gathers $j \neq k$. I like the following interpretation of the weight w_k: the local parabolic approximation of a step length α is the estimate of the weight w_k.

I use the nested (bilevel) iterative algorithm

$$\mathbf{m}_{k+1} = \mathbf{m}_k + \alpha_k \boldsymbol{\delta}_k(\mathbf{m}_k, \mathbf{s}_{k+1}), \quad \mathbf{s}_{k+1} = \arg\min \boldsymbol{\ell}_2^2(\mathbf{m}_k, \mathbf{s}_k), \tag{2.11}$$

where $\boldsymbol{\delta}_k = \boldsymbol{\delta}_k(\mathbf{m}_k, \mathbf{s}_{k+1})$ is a descent direction at the k-th iteration and α_k is a step length. The initial model $\mathbf{m}_0$ and the initial source time function $\mathbf{s}_0$ are required to start the algorithm. The lower-level linear least-squares problem (2.8) is solved in the frequency domain for $\mathbf{s}_1$ using $\mathbf{m}_0$ and $\mathbf{s}_0$ with a negligible regularization (Pratt, 1999). A solution for the upper-level non-linear least-squares problem (2.7) is updated by $\mathbf{m}_1 = \mathbf{m}_0 + \alpha_0 \boldsymbol{\delta}_0(\mathbf{m}_0, \mathbf{s}_1)$ using the time-domain wavefields filtered with a band-pass filter in the frequency domain.

The descent direction $\boldsymbol{\delta}_k(\mathbf{m}_k, \mathbf{s}_{k+1})$ is a preconditioned gradient of the misfit corresponding to a single randomly chosen shot gather. The gradient is computed using the adjoint-state method (Plessix, 2006) and preconditioned by the inverse of the diagonal approximation of the pseudo-Hessian (Shin et al., 2001) with a negligible regularization. The step length α is estimated using an adaptive parabolic line search algorithm (Kurzmann, 2012). The described algorithm resembles the variable projection algorithm for the separable least-squares problem (Golub and Pereyra, 2003, 1973) introduced into the geophysical context by van Leeuwen and Mulder (2009). The key differences with my work are the single shot gather at each iteration, the adaptive parabolic step length estimation and the multiscale strategy.

To implement the algorithm, I modified the code developed by Köhn (2011) using some parts from the code developed by Kurzmann (2012). I use the multiscale strategy using the following frequency-domain filter: $F(f) = (1 - \cos(\pi(f - f_1)/(f_2 - f_1)))/2$ for $f \in [f_1, f_2]$, $F(f) = 1 - (1 - \cos(\pi(f - f_3)/(f_4 - f_3)))/2)$ for $f \in [f_4, f_3]$, $F(f) = 1$ for $f \in [f_2, f_3]$, $F(f) = 0$ for $f < f_1$ and $f > f_4$, where f_1, f_2, f_3 and f_4 are the corner frequencies of the filter $F(f)$. In all experiments I set $f_1 = 1$ Hz and $f_2 = f_3 = f_4/2$. I convert the time-domain data to the frequency domain using the fast Fourier transform. I generate the (pseudo-)random sequences of descent directions with the standard random number generator in C language. The random sequence can be repeated if the same seed is used.

The computational cost of one nested iteration is five forward modellings only: the source time function inversion requires one modelling, two modellings are needed to compute the gradient of a misfit function and the parabolic step length estimation needs two more modellings.

As a stopping criterion I use the maximum run-time proportional to the number of iterations.

2.2.4 Sample estimates of the solution, uncertainty and sensitivity

Due to limited computational resources finding *all* Pareto optima of the *local* non-linear multiobjective minimization problem is impossible in FWI applications with many millions of model parameters and the today's computational resources. Even the evaluation of the data misfit for

the whole data set and checking the optimality conditions are computationally expensive. I will nowhere in this work check the Pareto optimality. This is expensive and not necessary at all for pragmatic uncertainty analysis.

Instead, I invert just a few "sample points" in $\mathbb{M}$ using the different random sequences of descent directions in BMWI. If the number of iterations is sufficient to reach a local Pareto front[3], then the inverted models are the sample points of the probabilistic solution (1.46) with the locally highest values of the volumetric measure. The sample mean and sample standard deviation of these sample points can be used as the estimates of the solution and uncertainty for the given initial guess and the given computational resources.

For any initial model $\mathbf{m}^i$ the sample mean of the inverted models $\mathbf{m}^i_k$ ($k = 1, ..., N$, where N is the total number of the inverted models) is $\langle \mathbf{m}^i \rangle$

$$\langle \mathbf{m}^i \rangle = \frac{1}{N} \sum_{k=1}^{N} \mathbf{m}^i_k \tag{2.12}$$

and its mean over the different initial models $\mathbf{m}^i$ ($i = 1, ..., N_i$)

$$\langle \mathbf{m} \rangle = \frac{1}{N_i N} \sum_{i=1}^{N_i} \sum_{k=1}^{N} \mathbf{m}^i_k. \tag{2.13}$$

To estimate the uncertainty, I calculate the sample standard deviation of the inverted models with respect to the mean value

$$\mathbf{V}^i = \sqrt{\frac{1}{N-1} \sum_{k=1}^{N} (\mathbf{m}^i_k - \langle \mathbf{m}^i \rangle)^2}, \tag{2.14}$$

and its mean over the different initial models $\mathbf{m}^i$ ($i = 1, ..., N_i$)

$$\langle \mathbf{V} \rangle = \sqrt{\frac{1}{N_i (N-1)} \sum_{i=1}^{N_i} \sum_{k=1}^{N} (\mathbf{m}^i_k - \langle \mathbf{m}^i \rangle)^2}, \tag{2.15}$$

where the operations are understood in a component-wise sense.

To estimate the sensitivity of the measuring procedure, I calculate the sample initial deviation of the inverted models with respect to the initial model $\mathbf{m}^i$

$$\mathbf{S}^i = \sqrt{\frac{1}{N} \sum_{k=1}^{N} (\mathbf{m}^i_k - \mathbf{m}^i)^2}, \tag{2.16}$$

and its mean over the different initial models $\mathbf{m}^i$ ($i = 1, ..., N_i$)

$$\langle \mathbf{S} \rangle = \sqrt{\frac{1}{N_i N} \sum_{i=1}^{N_i} \sum_{k=1}^{N} (\mathbf{m}^i_k - \mathbf{m}^i)^2}, \tag{2.17}$$

where the operations are understood in a component-wise sense. The Bessel's correction $\frac{N}{N-1}$ is not required for the sample estimate of the sensitivity.

The multiparameter aspect of FWI is beyond the scope of this work. I estimate only the variances and ignore the covariances.

2.3 Solution, uncertainty and sensitivity

To estimate the solution, uncertainty and sensitivity with the stochastic single-shot BMWI, I performed a few numerical experiments with the slightly modified Marmousi-2 model (Martin et al.,

[3] Pareto front is a region in the misfit domain corresponding to the Pareto optima.

2006): the sound velocity V of the Marmousi-2 model is set to 1250 m/s if $V < 1250$ m/s as shown at Figure 2.1. The mass density is spatially homogeneous: $\rho = 1000$ kg/m^3. I generated the pseudo-observed data using the following wavelet $s(t) = 0.75 f_c \sin^3(\pi(t+t_d))$, if $t \in [t_d, t_d + 1/f_c]$, and $s(t) = 0$ for all other times, where $f_c = 11$ Hz is the center frequency and $t_d = 0.5$ s is a time shift. As an initial source wavelet in inversion I used $s(t)$ with $f_c = f_3$ (the fourth corner frequency of the band-pass filter) and $t_d = 0$ s. In all experiments I set the corner frequencies of the filter to $f_1 = 1$ Hz and $f_2 = f_3 = f_4/2$. The minimal value of $f_4 = 4$ Hz.

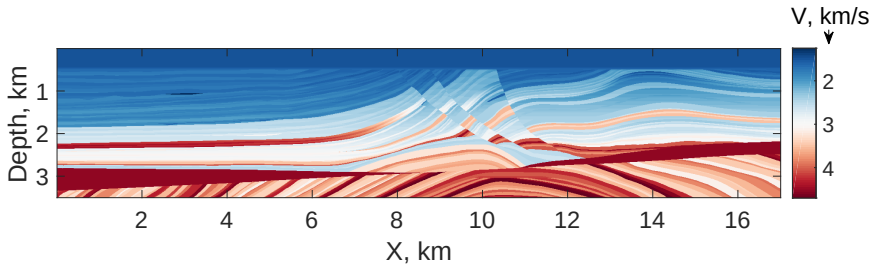


FIGURE 2.1: The model parameters of sound velocity used in the numerical experiments to generate the pseudo-observed data (Martin et al., 2006).

The spatial grid is 350×1700 points with a spatial sampling of 10 m, i.e., 3.5×17 km in size, except of the experiment in the last subsection where the grid is 500×1700 points. The recording time is 9 s. The time sampling is 1 ms. I used 12 points as absorbing boundaries.

In the majority of my experiments I used 26 point pressure sources placed at 20 m depth between $x = 1$ km and $x = 16$ km with spacing $\triangle_s = 600$ m and 231 point receivers located at 450 m depth between $x = 0.75$ km and $x = 16.25$ km with spacing $\triangle_r = 50$ m.

When I studied the effect of a lack of observations, I used 16 and 6 point pressure sources placed at 20 m depth and 121 receivers located at 450 m depth between $x = 5.5$ km and $x = 11.5$ km with the spacing $\triangle_r = 50$ m.

Smoothing and tapers are not applied to the descent directions. Regularization of the preconditioned gradients is negligible with a regularization parameter $\varepsilon = 10^{-10}$.

2.3.1 Inverting the subsets of the model parameters with ℓ_2-norms and cycle skipping

I use the stochastic single-shot BMWI to invert the different subsets of the model parameters with different starting models. I show a few simple examples and highlight the correspondence between the model and misfit domains.

The convergence of a misfit function evaluated for the whole dataset is rarely shown because the misfit is always converging if the initial setting is properly chosen. I suggest to show the misfit values evaluated for the different subsets of the data. Figure 2.2 shows the first 5 iterations of BMWI in two complementary ways. I used the linear gradient starting model of sound velocity $V_1 = 1.25 + 0.57z$ km/s.

The left sub-figure in Figure 2.2 shows the values of the single-shot misfit functions over the iterations at different frequency stages (here only the first stage with a maximum frequency of 4 Hz is shown). The length of the black bold line is equal to the decrease of a misfit function during the parabolic step length estimation: its upper edge corresponds to the value of a misfit function evaluated after the source time function inversion, the white dot corresponds to the misfit value evaluated for the model perturbed over the descent direction times the value of an adaptive step length, and the lower edge (if it does not coincide with the white dot) corresponds to the misfit value evaluated for the model perturbed over the descent direction times twice the value of the adaptive step length. If the lower edge of a black bold line coincides with the white dot, the third value of a misfit is bigger than or equal to the second value of a misfit during the parabolic line search.

The right sub-figure in Figure 2.2 shows the values of the misfits as functions of the shot number. The size of the dot depends on the number of iterations after which the misfit was evaluated: the dot corresponding to the misfit value evaluated at the first iteration has the smallest size and the dot corresponding to the misfit value evaluated at the last iteration has the biggest size.

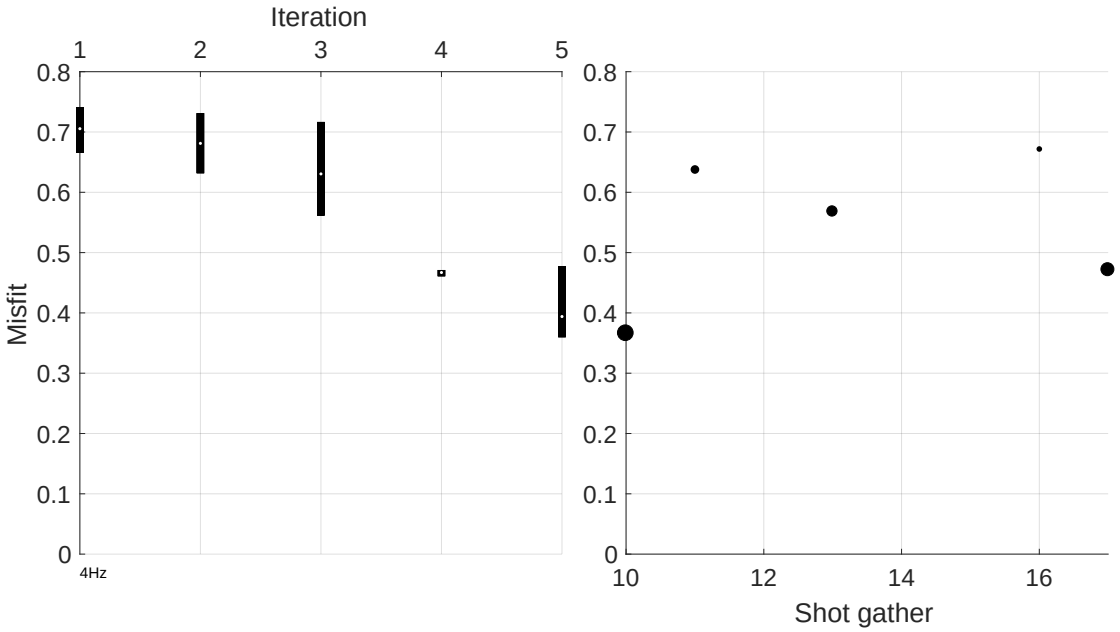


FIGURE 2.2: Single-shot ℓ_2^2-misfits: the first 5 iterations with the initial model $V_1 = 1.25 + 0.57z$ km/s at the lowest frequency stage with a maximum frequency of 4 Hz.

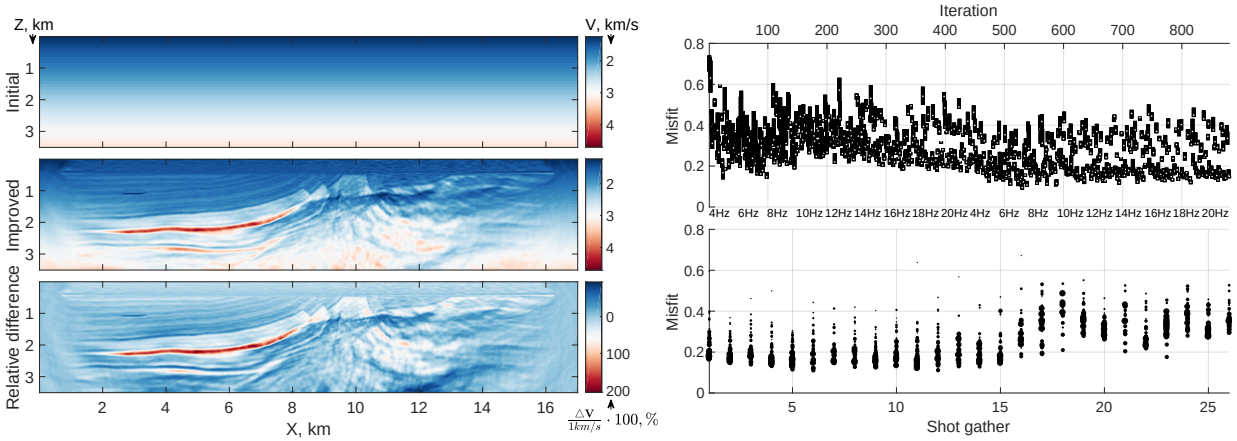
The single-shot misfit plots in Figure 2.2 provide information about the values of the misfit functions evaluated at different frequency stages, for different subsets of the data and at different iterations. These values have to be consistently low if the guessed law $\mathcal{F}(\mathbf{m})$ predicts the observed data $\mathbf{d}_o$. The fit of seismograms is never shown for all shots because it requires a lot of space, but the single-shot misfits (or even the misfits evaluated for smaller subsets of the data) can always be easily shown as illustrated in the following examples.

I run the single-shot BMWI with the two different linear-gradient starting models: $V_1 = 1.25 + 0.57z$ km/s and $V_2 = 1.25 + 0.71z$ km/s. Figures 2.3a and 2.4a show the initial model, the inverted model after 880 iterations with two multiscale cycles, and their relative difference. BMWI does not introduce artefacts *everywhere* in the model domain and recovers the sound velocity of water. Figures 2.3b and 2.4b show the single-shot misfit functions. The initial model with a lower gradient $\alpha = 0.57\ \mathrm{s}^{-1}$ of sound velocity leads to consistently lower values of the misfits in the left part of the computational domain. The initial model with a higher gradient $\alpha = 0.71\ \mathrm{s}^{-1}$ of sound velocity leads to the consistently lower values of the misfits in the right part of the computational domain. The shot gathers with the higher values of the misfits are located in the areas with the artefacts in the water layer.

A new starting model was created using the two previously inverted models, see Figure 2.5a. I cut the left side of the model shown in Figure 2.3a and the right side of the model shown in Figure 2.4a and combined them without any smoothing. The strong contrast between the left and rights parts of the model can be seen in the center of the inverted model. The result of inversion is satisfactory and the single-shot misfits shown in Figure 2.5b have consistently lower values (less than 0.2) than the misfits shown in Figures 2.3b and 2.4b.

The previous test was repeated with a smoothed version of the initial model as shown in Figure 2.6. I applied a mean filter over a square of size 0.1×0.1 km. The results of these two experiments (Figures 2.5 and 2.6), performed with the same random sequences of the descent directions, are comparable. The strong artificial contrast in the center of the inverted model disappeared because of the smoothing of the starting model.

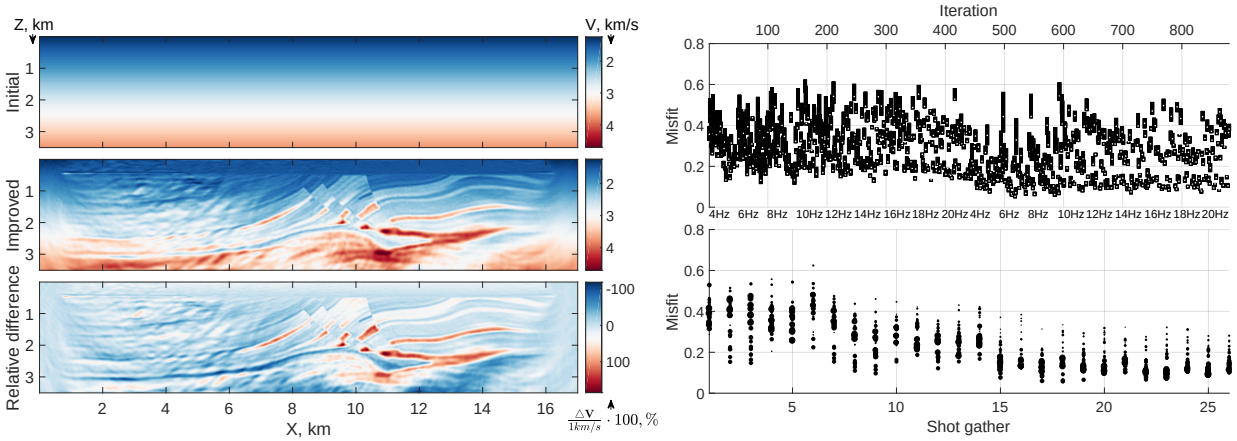
The single-shot misfit plots can be used to compare the guessed laws directly in the misfit domain because they are sensitive even to small variations of the starting model. I ran the single-shot BMWI with three slightly different starting models obtained by smoothing the true model using a mean filter over rectangles of size 0.5×6 km, 0.5×6.5 km and 0.5×7 km, respectively. Visually, the difference between the starting models is negligible (Figures 2.7a, 2.8a and 2.9a). The single-shot misfit plots (Figures 2.7b, 2.8b and 2.9b) provide a better measure of the quality of the inverted



(a) Model parameters.

(b) Single-shot ℓ_2^2-misfits.

FIGURE 2.3: 880 BMWI iterations in the model and misfit domains. The initial model is $V_1 = 1.25 + 0.57z$ km/s. (a) Model domain. (b) Misfit domain.



(a) Model parameters.

(b) Single-shot ℓ_2^2-misfits.

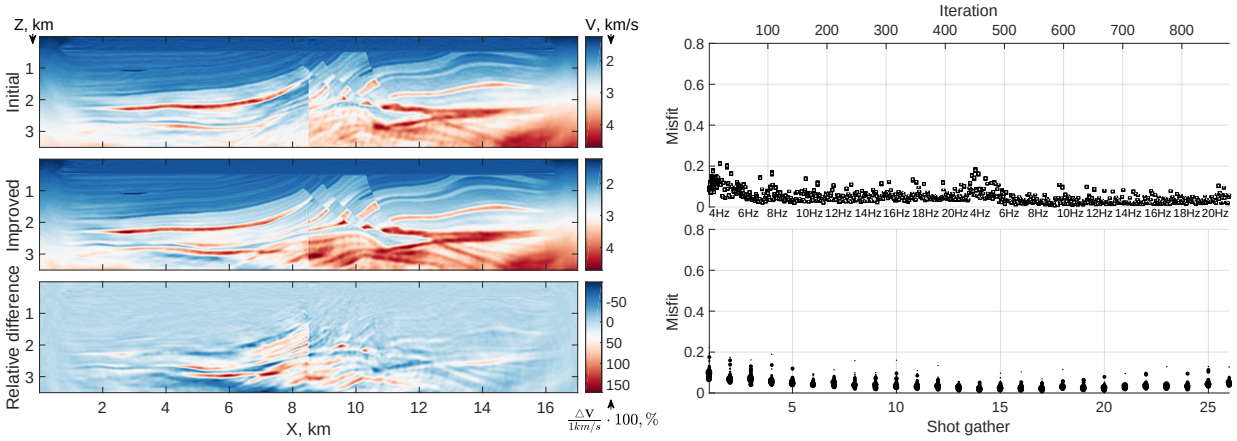
FIGURE 2.4: 880 BMWI iterations in the model and misfit domains. The initial model is $V_2 = 1.25 + 0.71z$ km/s. (a) Model domain. (b) Misfit domain.

results: the misfits shown in Figure 2.7b have *consistently* lower values than those in Figures 2.8b and 2.9b. The artefacts in the water layer (Figure 2.9a) are located in the same area as the shot gathers with comparatively high values of misfits shown in Figure 2.9b.

The stochastic single-shot BMWI does not create artefacts *everywhere* in the model domain. Different subsets of the model parameters can be inverted with the different starting models. The subsets of the model parameters corresponding to the consistently lower values of the misfit functions evaluated for the different subsets of the data during the stochastic BMWI can be combined and smoothed. The obtained model can be used as a new starting model. Comparison of the guessed laws can be performed directly in the misfit domain using the same random sequence of descent directions in BMWI.

2.3.2 Difference between the poor and good initial guesses in the model domain

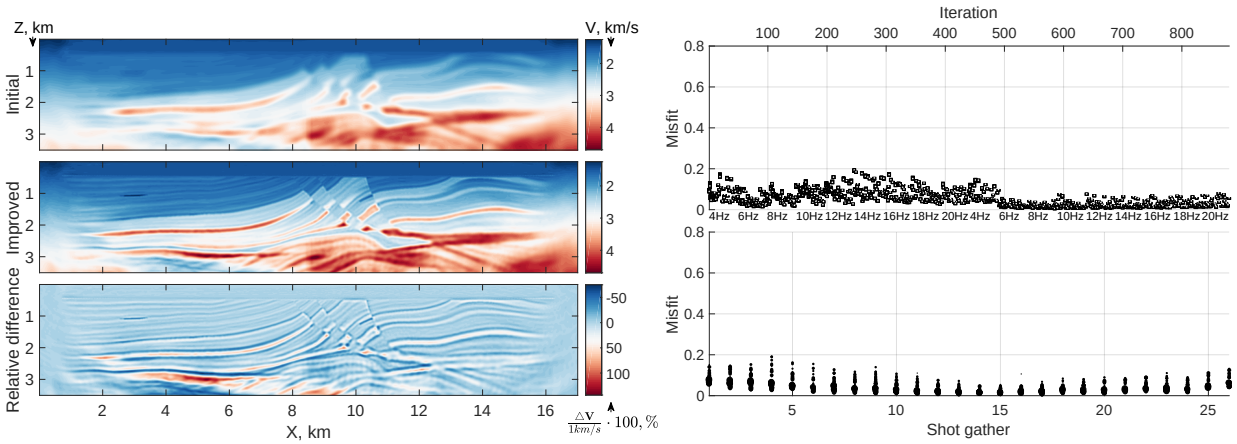
I have shown that the single-shot misfit functions can be used for indirect evaluation of the quality of the starting and inverted models. In this subsection I move the analysis entirely into the model domain without looking at the misfit domain.



(a) Model parameters.

(b) Single-shot ℓ_2^2-misfits.

FIGURE 2.5: 880 BMWI iterations in the model and misfit domains. The left part of the initial model is the left part of the inverted model in Figure 2.3a. The right part of the initial model is the right part of the inverted model in Figure 2.4a. (a) Model domain. (b) Misfit domain.



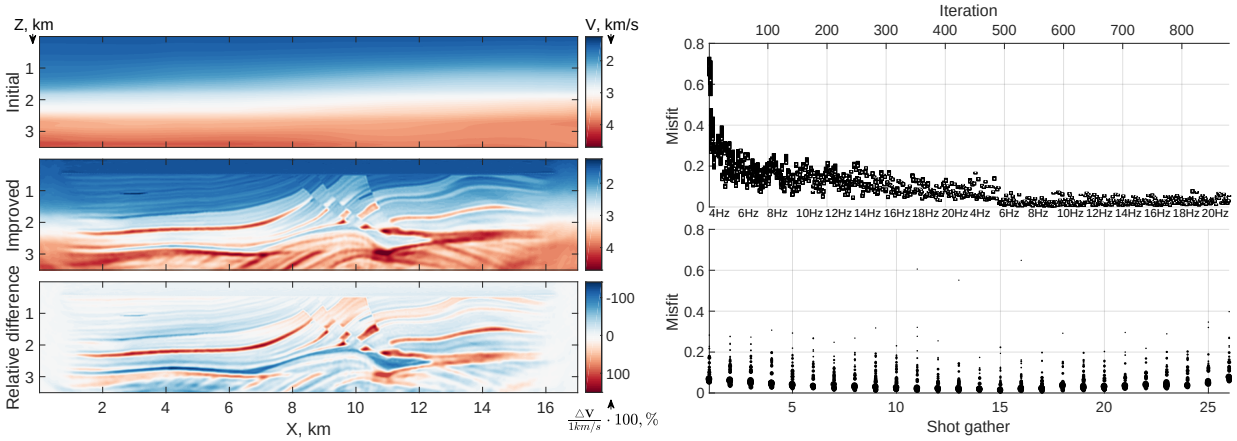
(a) Model parameters.

(b) Single-shot ℓ_2^2-misfits.

FIGURE 2.6: 880 BMWI iterations in the model and misfit domains. The initial model is the initial model shown in Figure 2.5 smoothed with a mean filter over a square of size 0.1×0.1 km. (a) Model domain. (b) Misfit domain.

The different random sequences of the stochastic single-shot BMWI, started from the same initial model $\mathbf{m}^1$, provide a set of the inverted models $\{\mathbf{m}_1^1, \mathbf{m}_2^1, ...\}$. The sample mean and sample standard deviation of $\{\mathbf{m}_1^1, \mathbf{m}_2^1, ...\}$ are the estimates of the solution and uncertainty for the initial guess $\mathcal{F}(\mathbf{m}^1)$. To illustrate this, I performed numerical experiments with smooth and linear gradient starting models.

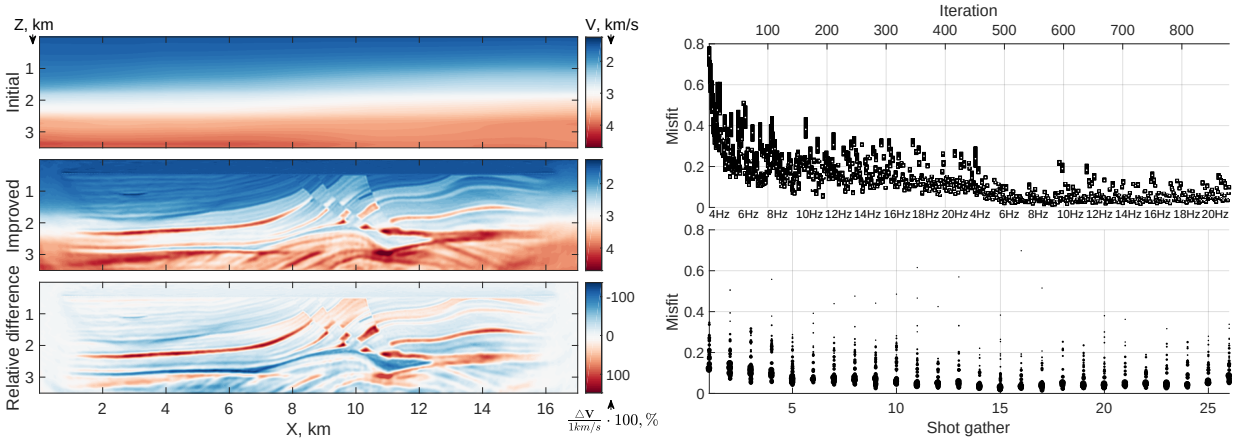
A smooth starting model. The inverted model shown in Figure 2.5a is smoothed with a mean filter over a square of size 0.1×0.1 km and used as a starting model. Figure 2.10 shows the sample estimates of the mean value, standard deviation and initial deviation of the 2, 3, 5 and 10 models inverted by 880 BMWI iterations. Even only two BMWI runs are sufficient to estimate the relative uncertainty of the model parameters. The highest values (roughly 150 m/s) of the standard deviation and the highest values (roughly 1 km/s) of the initial deviation in the left part at the bottom of the model correspond to the worst resolved area. The standard deviation at depths greater than 2 km reaches 100 m/s. Increasing the number of sample points smooths the sample estimates of the uncertainty. The estimate of the solution and uncertainty depends on the number of performed BMWI iterations as shown in Figure 2.11. The more iterations are performed, the lower the uncertainty of the model parameters in the upper part of the computational domain.



(a) Model parameters.

(b) Single-shot ℓ_2^2-misfits.

FIGURE 2.7: 880 BMWI iterations in the model and misfit domains. The initial model is the true model smoothed with a mean filter over a rectangle of size 0.5×6 km. (a) Model domain. (b) Misfit domain.



(a) Model parameters.

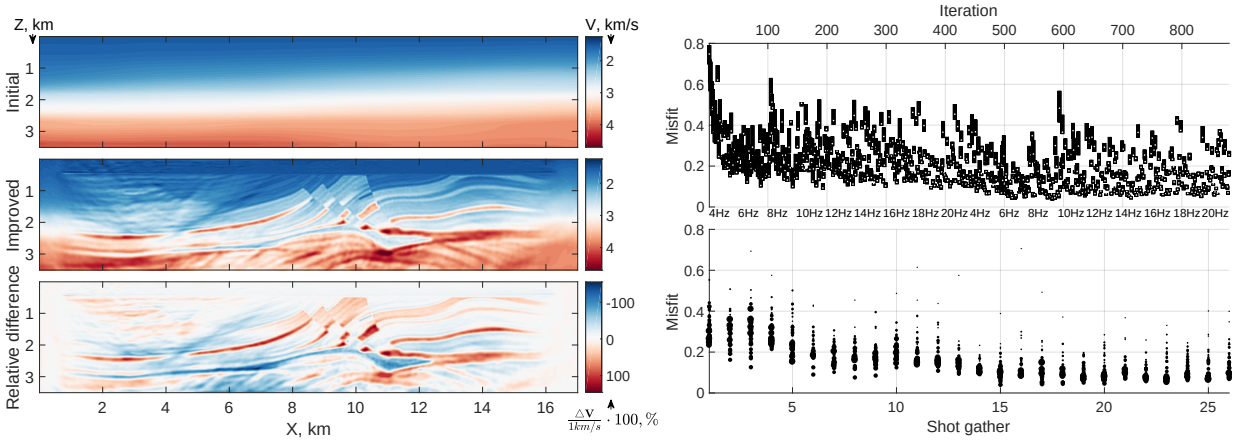
(b) Single-shot ℓ_2^2-misfits.

FIGURE 2.8: 880 BMWI iterations in the model and misfit domains. The initial model is the true model smoothed with a mean filter over a rectangle of size 0.5×6.5 km. (a) Model domain. (b) Misfit domain.

The second multiscale cycle has also decreased the standard deviation in the upper part of the computational domain.

A linear gradient starting model. I used $V_2 = 1.25 + 0.71z$ km/s as a starting model. The sample estimates of the mean value, standard deviation and initial deviation of the 2, 3, 5 and 10 models inverted by 880 BMWI iterations are shown in Figure 2.12. The uncertainties of parameters inverted from a linear gradient starting model are much higher than in the previous case with a smooth starting model. The standard deviation reaches the values 0.5 km/s and the initial deviation is as high as 1-1.5 km/s at depths greater than 2 km. The area with relatively low uncertainty – the upper right corner of the model – is localized. Even two runs of the stochastic BMWI are sufficient to estimate the relative uncertainties of the model parameters. The sample estimates of the solution and uncertainty depend on the number of performed BMWI iterations as shown in Figure 2.13. The second multiscale cycle improved the right part of the mean model and decreased the standard deviation in the upper right part of the model.

The closer the starting model to an optimum, the lower the uncertainty of the model parameters inverted by BMWI. The examples here suggest that even two different runs of the stochastic single-shot BMWI can be sufficient to estimate the uncertainty of the inverted model parameters. Even if the initial model is far away from an optimum, it can be possible to localise at least a subset of the



(a) Model parameters.

(b) Single-shot ℓ_2^2-misfits.

FIGURE 2.9: 880 BMWI iterations in the model and misfit domains. The initial model is the true model smoothed with a mean filter over a rectangle of size 0.5×7 km. (a) Model domain. (b) Misfit domain.

model parameters with relatively low uncertainty as shown in Figures 2.12 and 2.13.

2.3.3 Lack of the measured parameters

The performance of FWI depends on the measuring procedure in a wide sense. The stochastic single-shot BMWI can be used to estimate the uncertainty related to the lack of the observed data. To illustrate this, I performed a few experiments with 26, 16 and 6 sources and with 15.5 km and 6 km long OBCs[4] using three different random sequences of BMWI. The starting model is the true model smoothed with a mean filter over a rectangle of size 0.5×6 km.

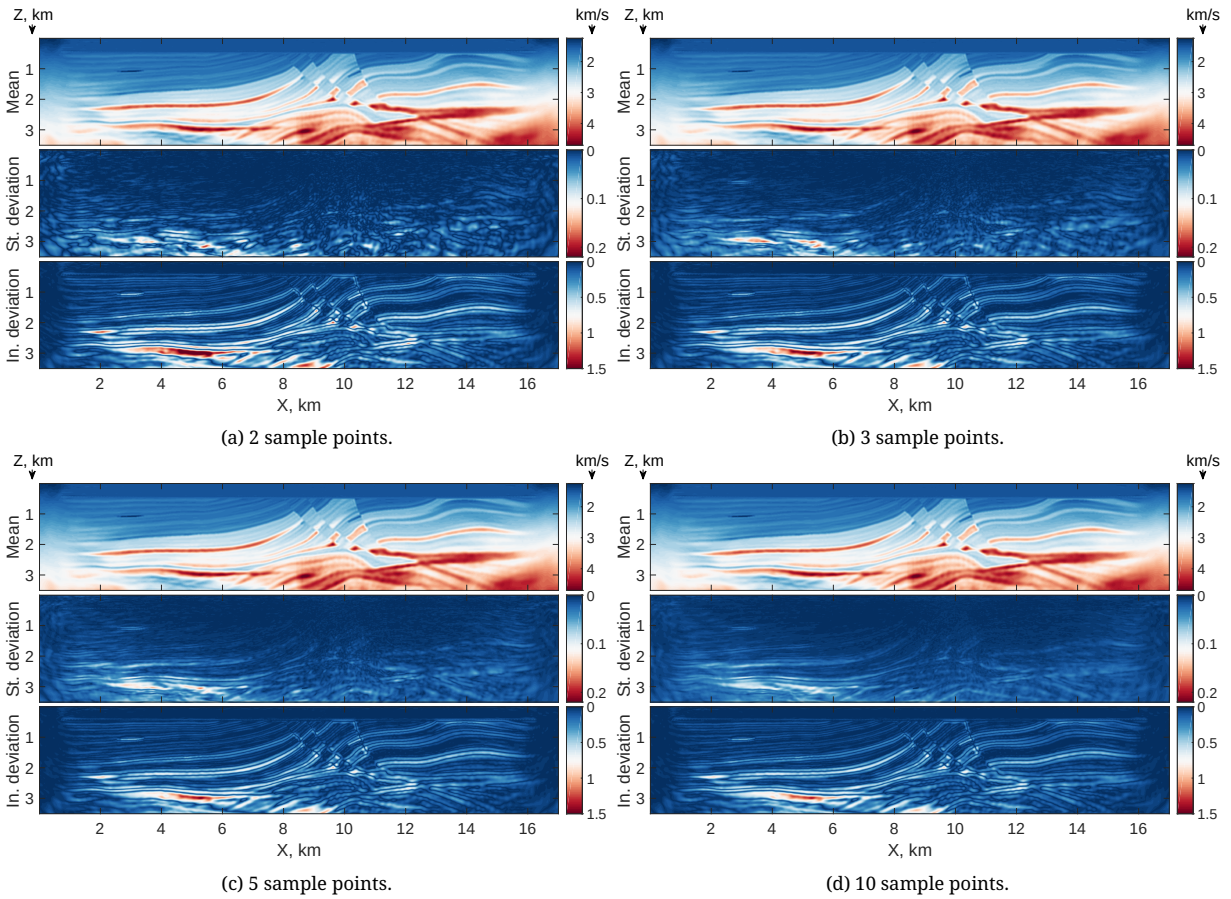
Figure 2.14 shows the sample estimates of the mean value, standard deviation and initial deviation of the three inverted models after 880 BMWI iterations for the different numbers of shots and receivers. The highest quality of the mean value and the lowest uncertainty correspond to the unrealistic acquisition geometry with 26 sources and 231 receivers (Figure 2.14a). The standard deviation reaches 0.3-0.4 km/s at depths greater than 2 km. In the more realistic experiment with 6 km long OBC, 121 receivers and 26 sources the sensitivity of the measuring procedure dramatically decreased in areas which are not covered by the line of receivers (Figure 2.14b). The standard deviation is minimal in the central semicircle with the radius 3 km equal to half length of the OBC representing the footprint of the acquisition geometry. The quality of the mean value of sound velocity in the central semicircle is comparable to the quality of the same parameters in the experiment with 231 receivers.

With 16 sources and 231 receivers the estimate of the mean value of model parameters is comparable with the corresponding estimate in the experiment with 26 sources and 231 receivers. The standard deviation reaches 0.5 km/s in the areas with high contrast of sound velocity at depths greater than 2 km. Reducing the length of the OBC to 6 km decreased the sensitivity of the measuring procedure and increased the uncertainty of the model parameters outside the central semicircle. The quality of the mean value of the model parameters in the central semicircle is slightly reduced in comparison to the experiment with 231 receivers.

The lowest quality of the mean value corresponds to the worst acquisition geometry with 6 sources, shown in Figures 2.14e and 2.14f. Perhaps the case most interesting and most difficult for interpretation is the uncertainty estimate in the experiment with 6 sources and 121 receivers.

First and unexpectedly, the values of the standard deviation in the left part of the model in Figure 2.14f are lower than the corresponding values in the experiment with the 16 sources and 121 receivers shown in Figure 2.14d. The artefacts in the left part of the mean model in Figure 2.14d are stronger than the artefacts in Figure 2.14f.

[4] OBC is an ocean bottom cable.



(a) 2 sample points.

(b) 3 sample points.

(c) 5 sample points.

(d) 10 sample points.

FIGURE 2.10: Sample estimates of the solution, uncertainty and sensitivity for the different number of sample points obtained after 880 BMWI iterations. The starting model is shown in Figure 2.6.

Second and also unexpectedly, the values of the standard deviation in the upper left part of the model in Figure 2.14f are lower than the corresponding values in the experiment with 6 sources and 231 receivers shown in Figure 2.14e. I give an interpretation of these experiments by analysing the decrease of the single-shot misfit functions.

A small change of an arbitrary misfit function ℓ_k ($k = 1, ..., N_s$) due to a small change of the model $\mathbf{m}$ in the direction $\alpha\boldsymbol{\delta}\mathbf{m}$ can be linearly[5] approximated as

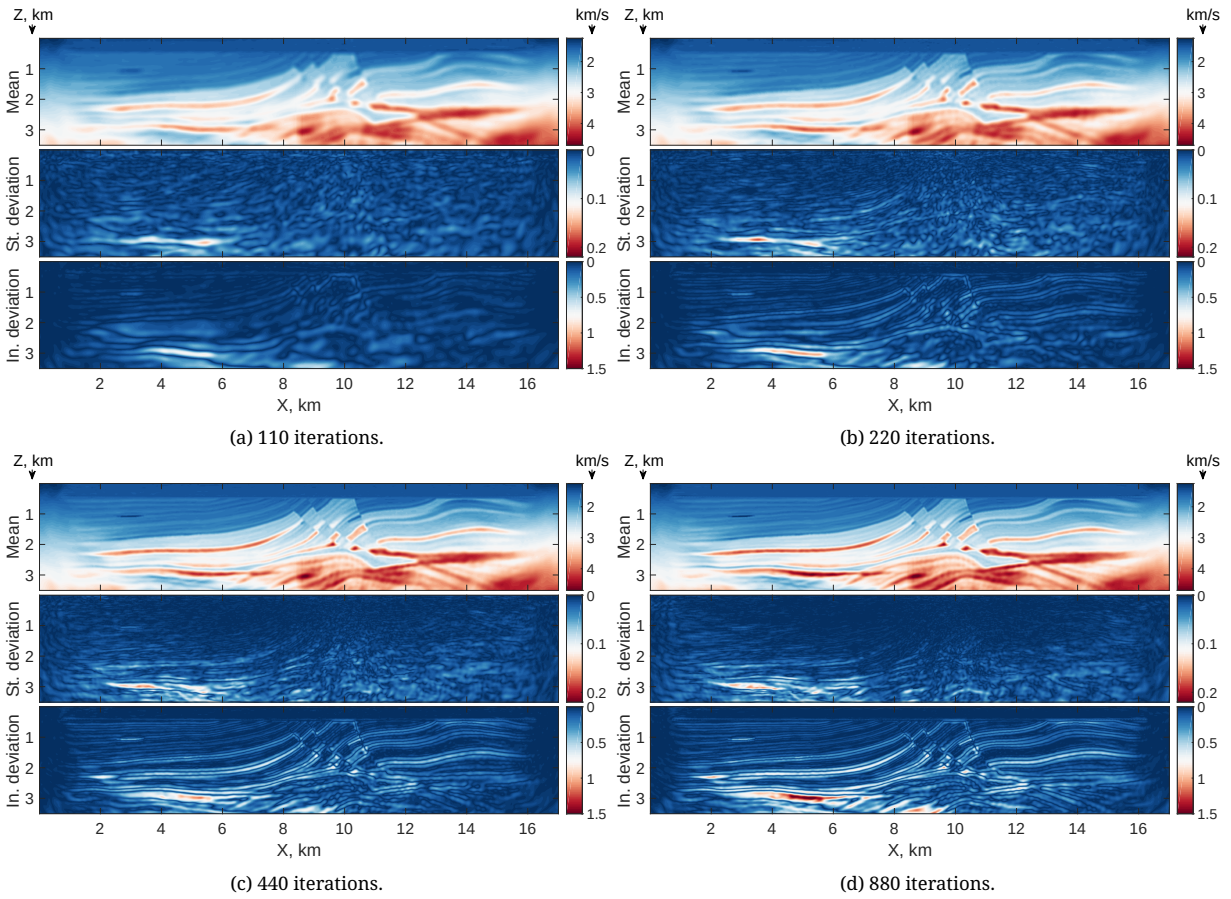
$$\ell_k(\mathbf{m} + \alpha\boldsymbol{\delta}\mathbf{m}) - \ell_k(\mathbf{m}) \approx \alpha \frac{\partial \ell_k}{\partial m_j} \delta m_j, \qquad k = 1, ..., N_s, \tag{2.18}$$

where the summation over $j = 1, ..., N_m$ (N_m is the number of model parameters) is assumed and the descent direction $\boldsymbol{\delta}\mathbf{m} = -\mathbf{P}\frac{\partial \ell_n}{\partial \mathbf{m}}$ is estimated using the n-th shot gather ($n = 1, ..., N_s$), the preconditioning matrix $\mathbf{P}$ is the inverse of the diagonal approximation of the pseudo-Hessian. If $\mathbf{P}$ is a positive semi-definite matrix, then its square root exists: $\mathbf{P} = \mathbf{P}_s\mathbf{P}_s$, where $\mathbf{P}_s$ is a positive semi-definite matrix. Then, the change of the misfit (2.18) is equal to $-\alpha C_{kn}$, where C_{kn} is the kn-th component of the cross-sensitivity matrix $\mathbf{C}$, i.e., the Gram matrix of dot products constructed on the gradients of the single-shot misfit functions preconditioned by $\mathbf{P}_s$:

$$C_{kn} = P_s^{lj} P_s^{li} \frac{\partial \ell_k}{\partial m_j} \frac{\partial \ell_n}{\partial m_i}, \tag{2.19}$$

where the summation over $l, j, i = 1, ..., N_m$ is assumed. If $C_{kn} > 0$, the value of the misfit is decreased. The two preconditioned single-shot gradients for the same set of receivers are positively

[5]The Taylor's expansion up to the quadratic term with the perfect preconditioning matrix $\mathbf{P}$ equal to the inverse Hessian does not change the analysis.



(a) 110 iterations.

(b) 220 iterations.

(c) 440 iterations.

(d) 880 iterations.

FIGURE 2.11: Sample estimates of the solution, uncertainty and sensitivity using three sample points obtained after different number of BMWI iterations. The starting model is shown in Figure 2.6.

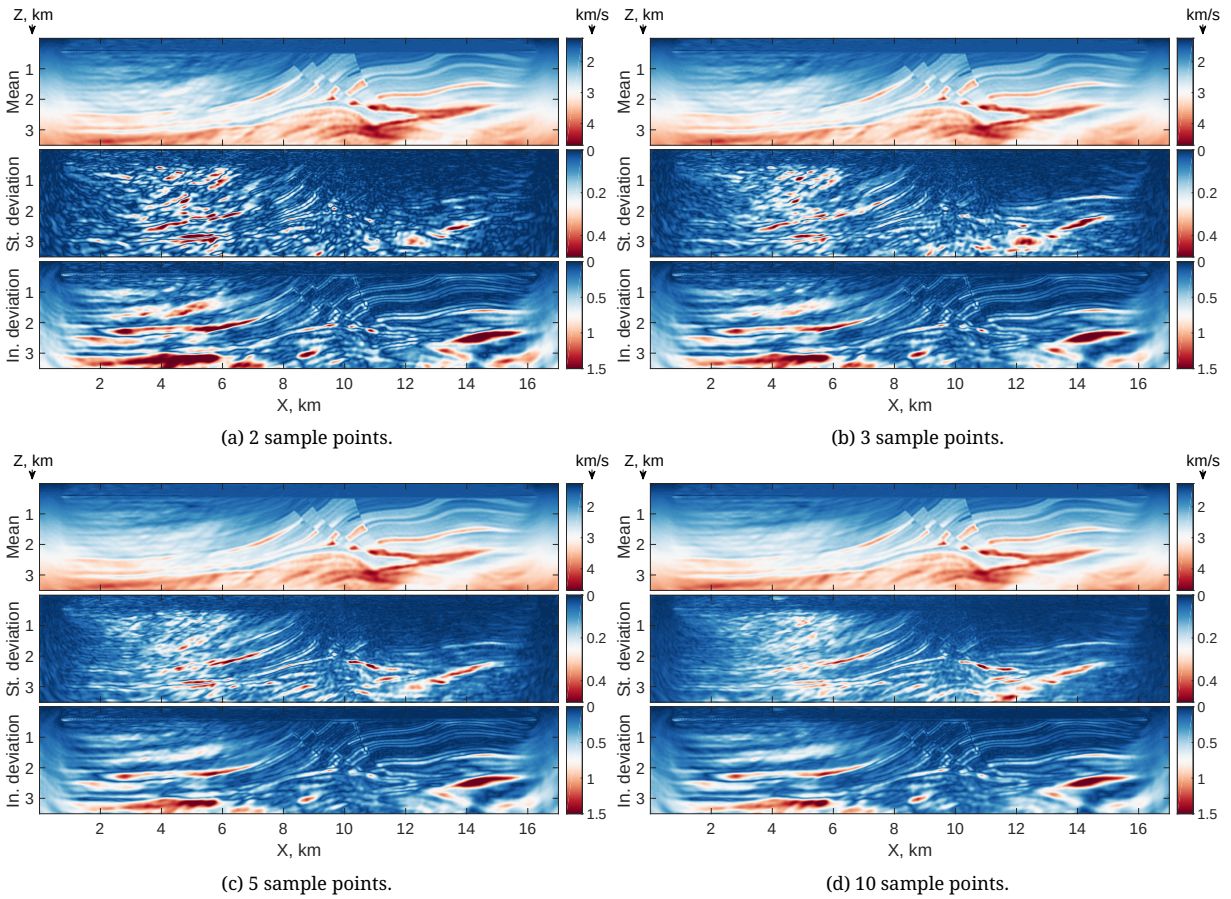
correlated if the distance between the sources is less than half a wavelength. If $C_{kn} = 0$, the misfit is unchanged, which means that the k-th misfit function is insensitive with respect to the gradient estimated using the n-th shot gather. If $C_{kn} < 0$, the value of the misfit is increased. The two preconditioned single-shot gradients for the same set of receivers are negatively correlated if the distance between the sources is less than the wavelength and more than half the wavelength. For larger distances between the sources, the values of C_{kn} can take arbitrary values depending on various factors.

The single-shot BMWI worsens the k-th misfit if $C_{kn} < 0$. If the sources and receivers are sparsely distributed, the cross-sensitivity matrix becomes diagonally dominant and the worsening of the misfits is reduced. In other words, the inversion results in Figure 2.14f are not better, they were just less worsened in comparison with the results in Figures 2.14d and 2.14e. This, perhaps, is at least a partial explanation of the observed results.

2.3.4 Averaging over the starting models

So far I focused on the uncertainties and sensitivities estimated with the same starting model and with different random sequences of BMWI. Here I illustrate that the averaging of the sample points inverted from the different starting models also provides the estimates of the uncertainty and sensitivity even with the same random sequence of descent directions in BMWI. Moreover, the different starting models are even necessary to estimate the sensitivity of the measuring procedure.

In exploration-scale FWI even the boundaries of the physical system under study are unknown. The choice of a computational domain and parametrization is non-unique and is constrained by the available computational resources. If the computational resources are sufficient, I would even



(a) 2 sample points.

(b) 3 sample points.

(c) 5 sample points.

(d) 10 sample points.

FIGURE 2.12: Sample estimates of the solution, uncertainty and sensitivity for the different numbers of sample points obtained after 880 BMWI iterations. The linear gradient starting model is $V_2 = 1.25 + 0.71z$ km/s.

recommend to invert the data in the extended computational domain because in this case the uncertainty and sensitivity plots are almost not affected by the boundaries.

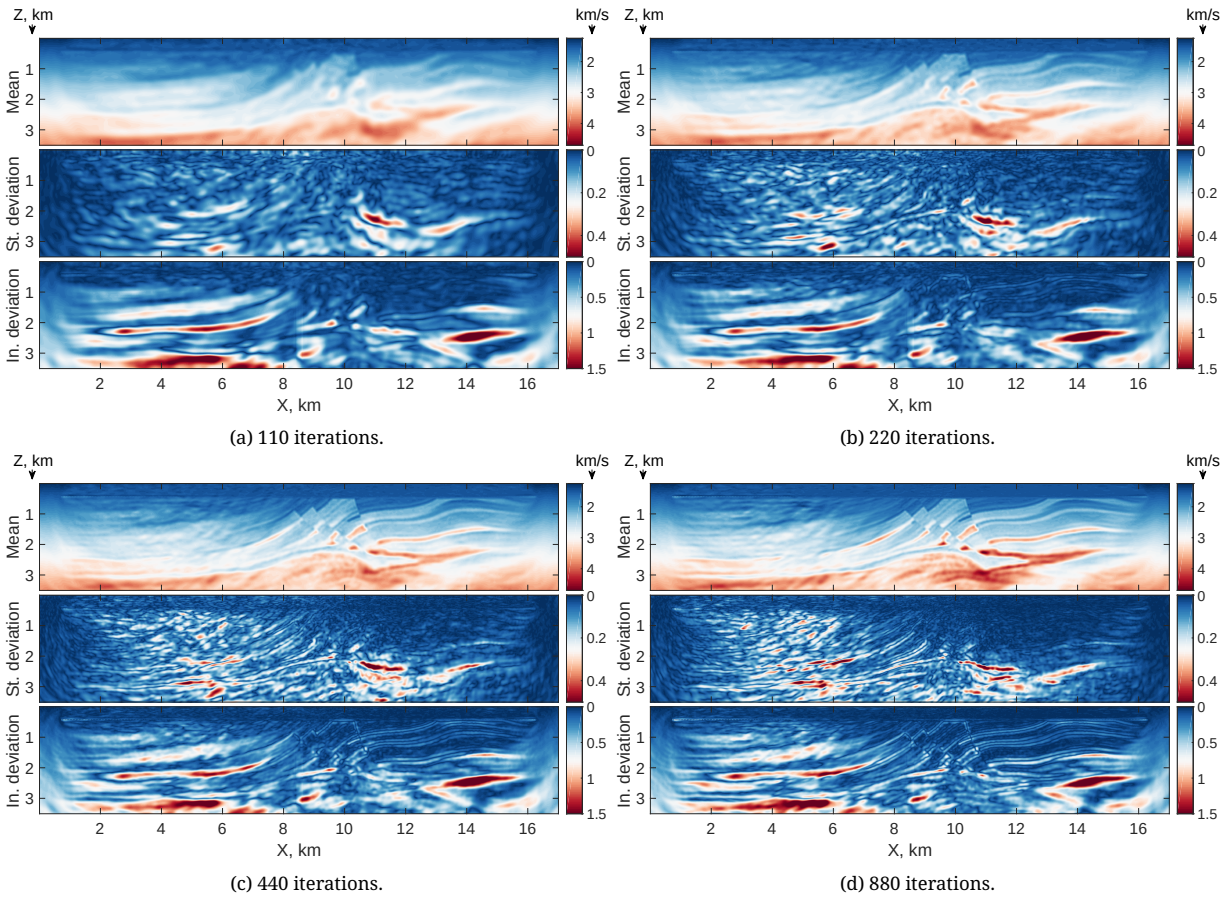
I performed BMWI experiments in a computational domain extended to 5 km depth with three slightly different linear gradient starting models as shown in Figure 2.15a. Figures 2.15b-2.15f show the sample estimates of the mean value, standard deviation and initial deviation of the model parameters inverted by 55, 110, 220, 440 and 880 BMWI iterations. I used the same random sequences of descent directions: the initial models are the only difference between the BMWI runs.

The areas close to the computational boundaries have the lowest sensitivity. Even 55 single-shot BMWI iterations provide the estimates of the uncertainty and sensitivity comparable with those obtained after 880 iterations: the footprint of acquisition geometry is already visible.

The quality of the mean value of the model parameters inverted after 440 and 880 BMWI iterations is excellent in the central semicircle where the corresponding standard deviation is low. The second multiscale cycle (between 440 and 880 BMWI iterations) decreased the uncertainty in the central semicircle, but increased the uncertainty outside the central semicircle.

2.4 Discussion: Higher resolution at lower computational cost?

A descent direction in FWI is a preconditioned gradient of a misfit function. In a homogeneous medium with sound velocity V and in the far-field of the Green's functions the gradient of a misfit $\nabla_{\mathbf{m}}\ell(\mathbf{m}(\mathbf{x}))$ with respect to the model parameters $\mathbf{m}$ is a truncated Fourier series (Sirgue and Pratt,



(a) 110 iterations.

(b) 220 iterations.

(c) 440 iterations.

(d) 880 iterations.

FIGURE 2.13: Sample estimates of the solution, uncertainty and sensitivity using three sample points obtained after different numbers of BMWI iterations. The linear gradient starting model is $V_2 = 1.25 + 0.71z$ km/s.

2004; Virieux and Operto, 2009):

$$\nabla_{\mathbf{m}} \ell(\mathbf{m}(\mathbf{x})) = -\omega^2 \sum_{\omega} \sum_{s} \sum_{r} \Re \left(\exp\left(-\frac{i\omega}{V} \left(\hat{\mathbf{s}} + \hat{\mathbf{r}} \right) \cdot \mathbf{x} \right) \triangle \mathbf{d}(\mathbf{r}, \mathbf{s}) \right), \tag{2.20}$$

where i denotes the imaginary unit, $\omega = 2\pi f$ is the angular frequency, $\hat{\mathbf{s}}$ and $\hat{\mathbf{r}}$ are the unit vectors in the incident propagation direction and in the inverse scattering direction, respectively, and $\triangle \mathbf{d}(\mathbf{r}, \mathbf{s})$ are the data residuals.

The Nyquist-Shannon sampling theorem gives rough estimates for the maximum source $\triangle_s$ and receiver $\triangle_r$ spacing as half the wavelength $\lambda = \frac{V}{f}$. Consequently, to avoid aliasing, at higher frequencies more sources and receivers are required. To suppress aliasing, a descent direction is often smoothed using various filters. Let $N_A = \frac{2\triangle}{\lambda_{min}} = \frac{2\triangle f_{max}}{V_{min}}$ be the sparseness of the data with $\triangle = \max(\triangle_s, \triangle_r)$, then in successful FWI applications the sparseness N_A does not usually exceed 20 if the smoothing of descent directions is used (Bleibinhaus et al., 2009).

Consider again the results inverted by the single-shot stochastic BMWI presented in Figure 2.14. I did not smooth the descent directions. In the experiments with 16 and 6 sources the sparseness is equal to $N_A = 32$ and $N_A = 80$, respectively. These values are computed with $\triangle_s$ because in the experiments with 231 and 121 receivers $\triangle_s > \triangle_r$. On the one side, the standard deviation is different in the experiments with 231 and 121 receivers. Perhaps, the sparseness of the data is not necessarily a proper measure of the resolution limits. On the other side, the high values of sparseness $N_A = 32$ and $N_A = 80$ in the experiments with 16 and 6 sources and 231 receivers look promising given the quality of the results in the experiment with 16 sources.

In contrast to formula (2.20), in the single-shot stochastic BMWI the gradient of a misfit function with respect to the model parameters is a truncated Fourier series without the sum over the shots:

$$\nabla_{\mathbf{m}}\ell(\mathbf{m}(\mathbf{x})) = -\omega^2 \sum_{\omega}\sum_{r} \Re\left(\exp\left(-\frac{i\omega}{V}\left(\hat{\mathbf{s}}+\hat{\mathbf{r}}\right)\cdot\mathbf{x}\right) \triangle\, \mathbf{d}(\mathbf{r},\mathbf{s})\right). \tag{2.21}$$

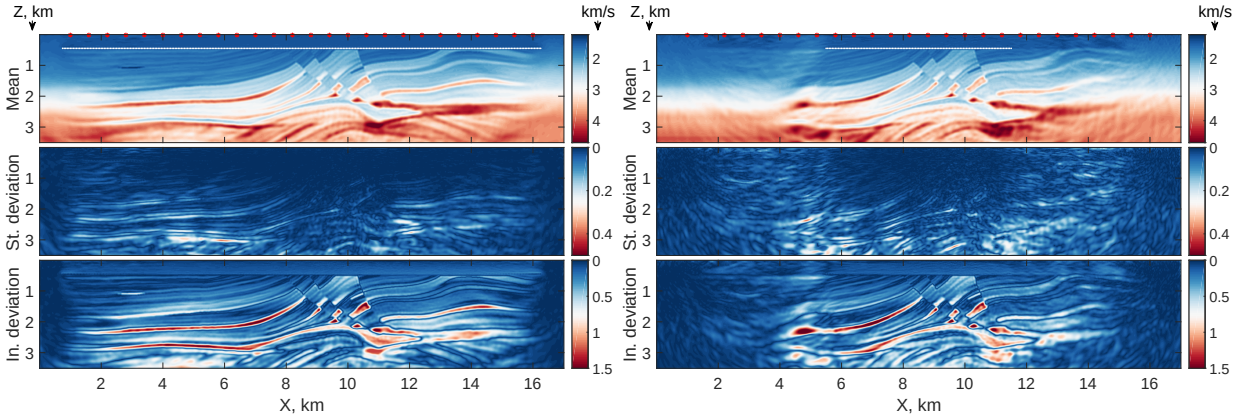
Does it mean that the source sampling can be arbitrary? No, the source-related aliasing is removed from a descent direction, but not excluded from the inverse problem, because the cross-talks appear during the iterations and depend on the choice of a sequence of descent directions. I observed the wave-like artefacts at the bottom of the Marmousi model, when instead of a random sequence of descent directions I used a regular sequence $\{1,...,N_s,1,..,N_s,1...\}$.

2.5 Summary

To make probabilistic waveform inversion computationally attractive, I implemented the stochastic single-shot bilevel multiobjective waveform inversion (BMWI). The main aspects of BMWI are the variable projections, randomly chosen single shot gather at each iteration, adaptive parabolic line search and multiscale strategy.

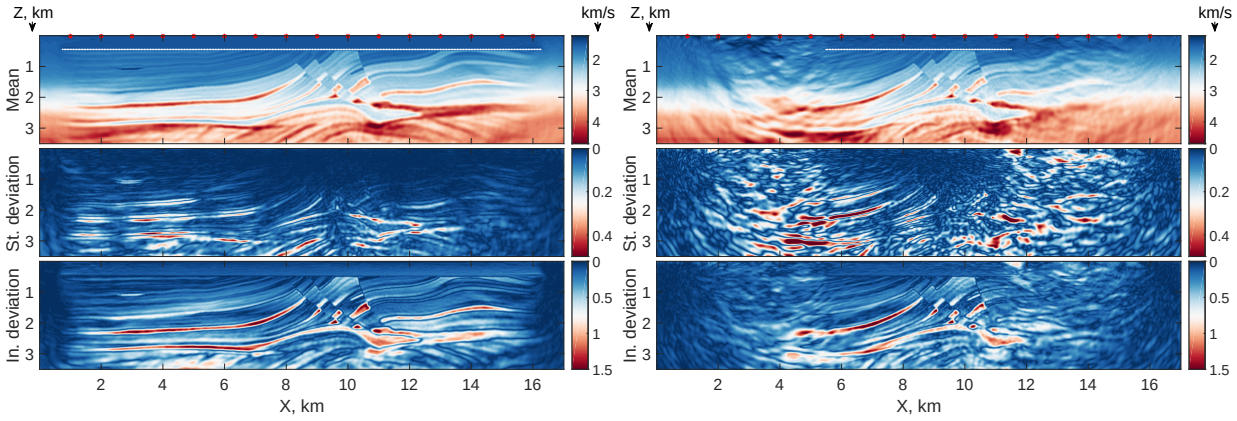
Instead of choosing many samples of the prior uniform distribution over the model manifold, a few simple linear-gradient starting models can be used to reconstruct the subsets of the model parameters with relatively small uncertainty. The subsets of the model parameters with relatively small uncertainty can be combined into a new starting model. Inverting the data using a new starting model with the different random sequences of BMWI leads to different inverted models, i.e., the sample points of the probabilistic solution. The sample mean and sample standard deviation of these models are the estimates of the solution and uncertainty. The initial deviation of the inverted models averaged over the different starting models is the estimate of sensitivity of the measuring procedure.

In chapter 3 I apply the stochastic single-shot BMWI to a real OBC data set acquired in a shallow-water river delta.



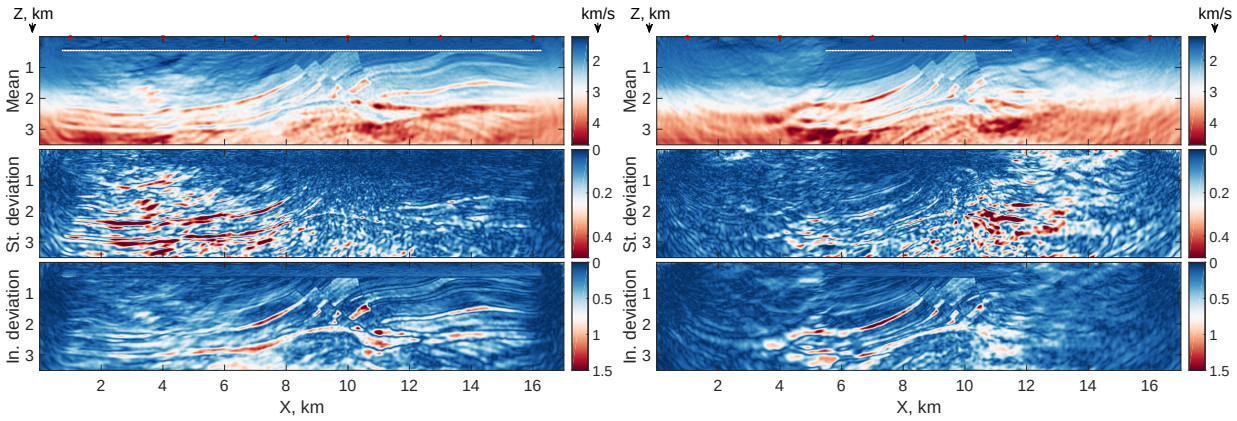
(a) 26 sources and 231 receivers in 15.5 km OBC.

(b) 26 sources and 121 receivers in 6 km OBC.



(c) 16 sources and 231 receivers in 15.5 km OBC.

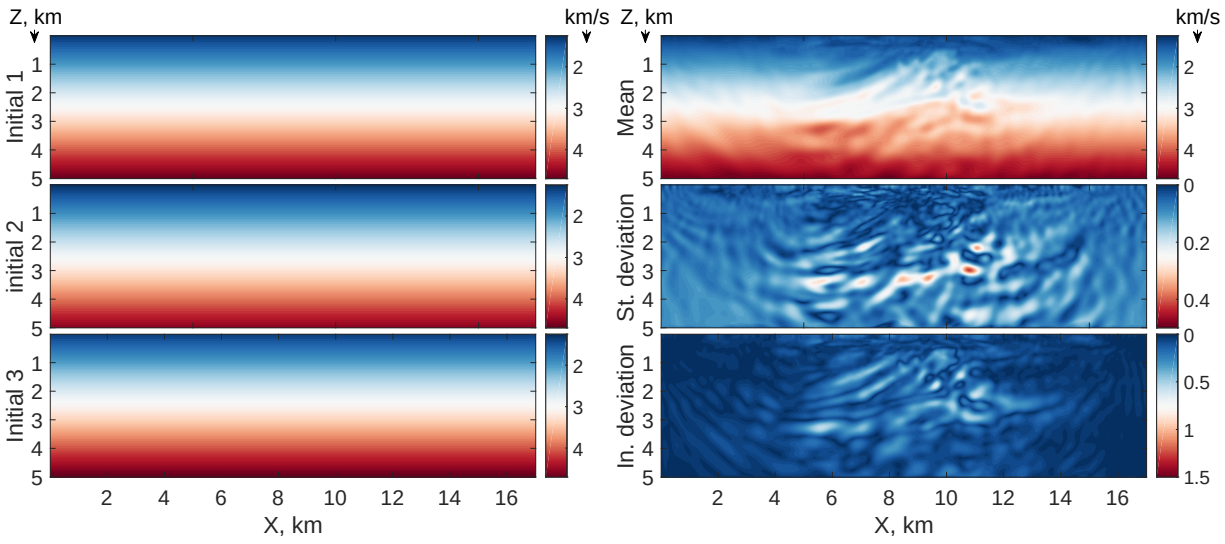
(d) 16 sources and 121 receivers in 6 km OBC.



(e) 6 sources and 231 receivers in 15.5 km OBC.

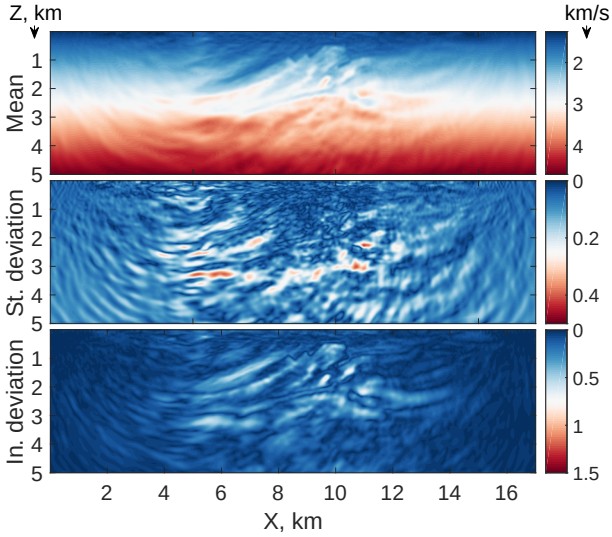
(f) 6 sources and 121 receivers in 6 km OBC.

FIGURE 2.14: Sample estimates of the solution, uncertainty and sensitivity using three sample points obtained after 880 BMWI iterations for different number of sources and receivers. The initial model is the true model smoothed with the mean filter over the rectangle 0.5×6 km. The red stars denote the sources. The white dots denote the receivers.

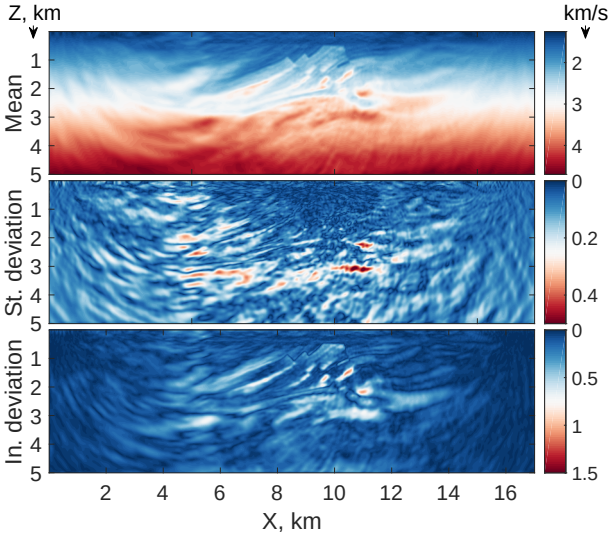


(a) Initial sound velocities.

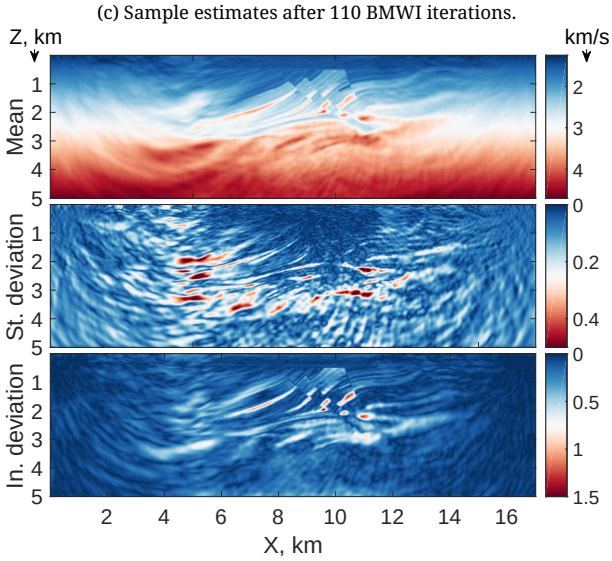
(b) Sample estimates after 55 BMWI iterations.



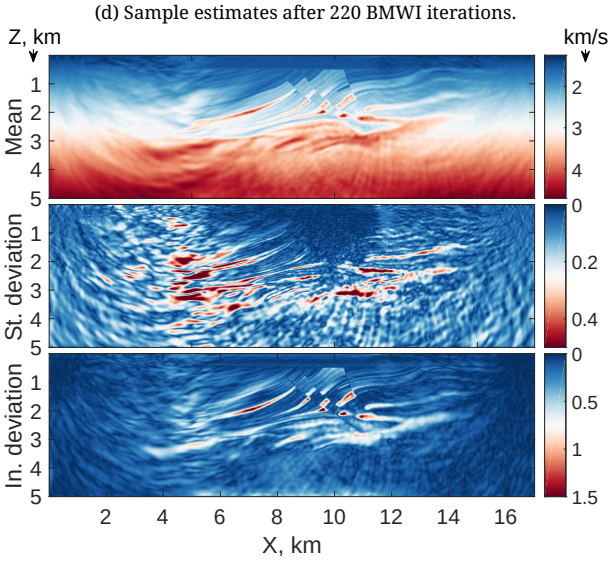
(c) Sample estimates after 110 BMWI iterations.



(d) Sample estimates after 220 BMWI iterations.



(e) Sample estimates after 440 BMWI iterations.



(f) Sample estimates after 880 BMWI iterations.

FIGURE 2.15: Sample estimates of the solution, uncertainty and sensitivity using three sample points inverted by the different number of BMWI iterations from the three different linear gradient starting models. The computational domain is extended up to 5 km depth. 26 sources and 121 receivers in the 6 km long OBC are used.

Chapter 3

Real data example

3.1 Introduction

In this chapter I estimate the governing law of a physical system using field data acquired during a marine seismic experiment with ocean bottom cables. The experiment was performed in a shallow-water river delta by an unspecified company.

I use the stochastic single-shot bilevel multiobjective waveform inversion (BMWI) in a 2D acoustic isotropic finite-difference approximation as described in chapter 2. I perform minimal data preprocessing (only a new 3D-to-2D transform) and use minimal prior information (the linear gradient starting models).

With a reasonable computational cost I reconstruct subsurface models which are partially consistent with the observed data. The inverted models sample the probabilistic solution in the specified setting. As the estimate of the solution, uncertainty and sensitivity, I compute the sample estimates of the mean value, standard deviation and initial deviation of the model parameters inverted by the stochastic single-shot BMWI.

The structure of this chapter is as follows. First, I specify the setting and present a new 3D-to-2D transform strictly valid for a linear gradient acoustic medium. Then, I estimate the theoretical uncertainties due to the neglected 3D effects by inverting the data 3D-to-2D transformed with different values of the velocity gradient. Next, I estimate the uncertainties using the different random sequences of descent directions in BMWI. My final estimates of the solution, uncertainty and sensitivity are the averages over the different starting models, different random sequences and different 3D-to-2D transforms.

I do not give any geological interpretation of the results. This is beyond the scope of my work.

3.2 Setting

The system under study is a physical system including a measuring procedure (experimental, numerical, technical, computational, algorithmic and formal aspects). Many aspects of the physical system and the measuring procedure are either unknown or have not been provided by the acquisition company.

3.2.1 OBC experiment in a shallow-water river delta

The physical system under study is a seismic marine experiment with ocean bottom cables performed in a shallow-water river delta. The physical system is open and the boundaries of the system are unknown. Only relative coordinates of the sources and receivers are known as illustrated in Figure 3.1.

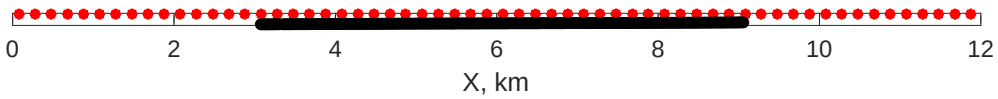


FIGURE 3.1: The acquisition geometry. The red circles denote the sources. The black line denotes the ocean bottom cable. The boundaries of the physical system are unknown.

I had only a 2D line (6 s long) of the full 3D OBC data set (9 s long) available with a time sampling of 2 ms. The raw data set contains 61 shot gathers acquired with air-guns at around 5-6 m depth and 240 hydrophones at a tilted sea floor. The receiver depths are 120-140 m. The receiver spacing is 25 m. The sources covered a range of 12 km. The length of the OBC is 6 km. The known parameters of acquisition are summarized in Table 3.1.

Parameter	Value
Number of shots	61
Shot depth	5-6 m
Shot point interval	200 m
Number of hydrophones	240
Hydrophone depth	122 m - 146 m
Hydrophone interval	25 m
Profile length	12 km
Offsets	118 m - 8993 m
Record length	6 s
Sample interval	2 ms
Usable frequency range	3 Hz - 230 Hz

TABLE 3.1: Parameters of the acquisition geometry.

The physical system is modelled by the 2D acoustic isotropic wave equation discretized by the time-domain finite-difference algorithm as described in chapter 2. Many physical processes in the real Earth are neglected, but for the given limited amount of observed data – a 2D line of the pressure seismograms – it is computationally unreasonable to use a 3D forward solver.

3.2.2 Minimal data preprocessing: a 3D-to-2D transform only

Figure 3.2 shows a few examples of the raw far-offset shot gathers. The shot gather 55 is, perhaps, the worst one but I use it in the inversion as well. Figure 3.3 shows a few examples of the raw near-offset shot gathers. The near-offset data have higher signal-to-noise ratio in comparison with the far-offset data. The one trace in each shot gather is dead.

Muting, denoising, offset- and time-windowing are not used. I use only spline interpolation to change the data sampling when it is necessary for the finite-difference approximation of the wave equation. The single dead trace in each shot gather is not used.

The conventional FWI applied to this data set fails with a linear gradient starting model. The previous attempts to estimate a subsurface model with this data set were performed using a conventional FWI algorithm with a very smooth starting model obtained via a traveltime tomography and the mass density model estimated via Gardner's empirical relation (Habelitz, 2017; Kunert et al., 2016; Kunert, 2015).

To take into account 3D effects in a computationally efficient way, I apply a 3D-to-2D transform to field data. Under the assumption of a translational symmetry of the physical system along the

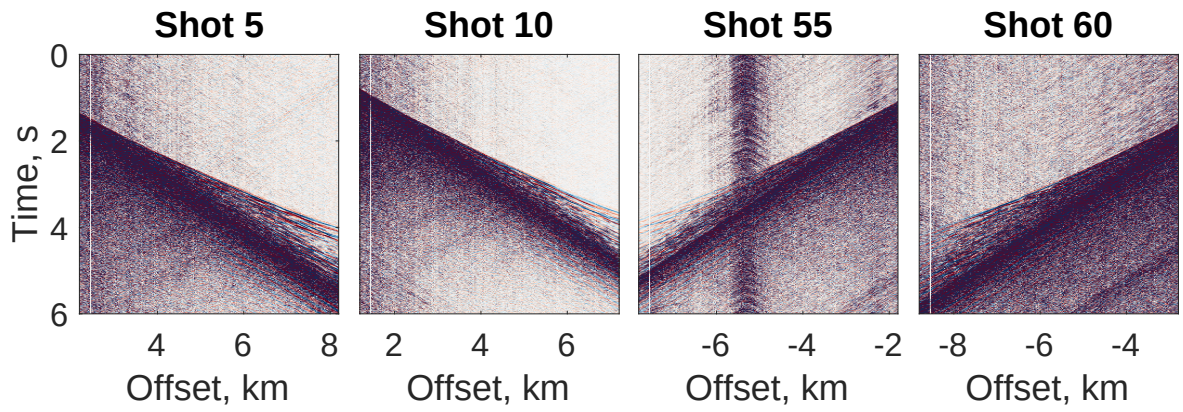


FIGURE 3.2: The raw far-offset shot gathers of pressure data. One trace is dead.

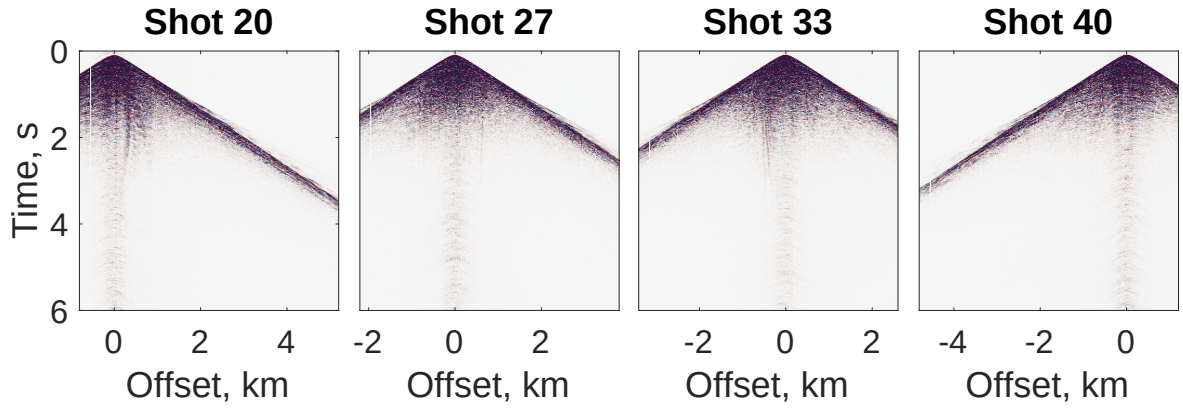


FIGURE 3.3: The raw near-offset shot gathers of pressure data. One trace is dead.

y-axis, a 3D-to-2D transform of the 3D pressure field $p_3(x,0,z,t)$ into the 2D pressure field $p_2(x,z,t)$ is the time-domain convolution integral $p_2(x,z,t) = \int p_3(x,0,z,t-\tau)F(\tau)d\tau$ with a kernel $F(\tau)$ that depends on the model parameters of a medium.

The exact *time-domain* kernel in a homogeneous acoustic medium is given by (Forbriger et al., 2014):

$$F_h = \frac{2V\tau_0 H(\tau)}{\sqrt{\tau\,(\tau+2\tau_0)}}, \tag{3.1}$$

where $\tau_0 = \frac{r}{V}$, V is the sound speed, r is the source-receiver distance and H denotes the Heaviside function. The standard time-domain 3D-to-2D transform as "a multiplication with $\sqrt{t}$ and convolution with $\sqrt{\frac{1}{t}}$" (Crase et al., 1990; Pica et al., 1990) is a special case of the exact kernel (3.1) for $\tau \ll 2\tau_0$.

I derived the kernel F_g for a 3D-to-2D transform in a linear gradient medium $V(z) = V_0 + \alpha z$ with the sound velocity V_0 at zero depth $z = 0$ and velocity gradient α:

$$F_g = \frac{\alpha\sqrt{\hat{z}\hat{z}_0}\sinh(\alpha\tau_0)H(\tau)}{\sqrt{\sinh(\frac{\alpha\tau}{2})\sinh\left(\frac{\alpha(\tau+\tau_0)}{2}\right)}}, \tag{3.2}$$

where $\tau_0 = \frac{1}{\alpha}\operatorname{acosh}\left(\frac{x^2+\hat{z}^2+\hat{z}_0^2}{2\hat{z}\hat{z}_0}\right)$, $\hat{z} = z + \frac{V_0}{\alpha}$, x and z are the coordinates of a receiver, $x_0 = 0$ and z_0 are the coordinates of the source, $\sinh$ denotes the hyperbolic sine and acosh denotes the inverse hyperbolic cosine. Formula (3.2) is obtained by integrating the well-known analytic solution for a 3D linear-gradient medium (Kuvshinov and Mulder, 2006; Pekeris, 1946) over the y-axis and by changing the variable of integration from y to τ. The kernel F_g reduces to F_h, if α goes to zero.

3.2.3 Bilevel multiobjective waveform inversion

The stochastic single-shot BMWI is applied to the data set. Figure 3.4 shows the data fit at the initial stage of inversion using the linear gradient starting models for sound velocity $V = 1500 + 0.6z$ m/s and mass density $\rho = 1000 + 0.6z$ kg/m^3: at iteration 13 for shot 52 and at iteration 15 for shot 8. The field data below 3 Hz are noisy and not reliable.

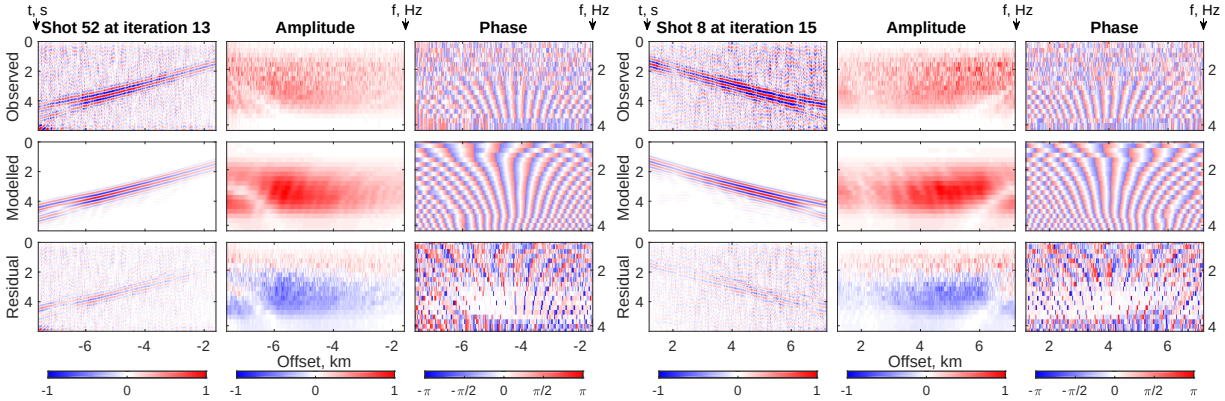


FIGURE 3.4: The data fit at the lowest frequency stage of BMWI. The field data below 3 Hz are noisy and not reliable. Note, the same colorbar in the range [-1,1] is used for the frequency-domain amplitudes and frequency-domain residual amplitudes, although the frequency-domain amplitudes vary only in the range [0,1].

The initial source wavelet is $s(t) = 0.75 f_c \sin^3(\pi(t + t_d))$ if $t \in [t_d, t_d + 1/f_c]$, and $s(t) = 0$ for all other times, where f_c is the center frequency and t_d is a time shift. I used a free-surface boundary condition at the top of the computational domain and absorbing layers at the left, right and bottom of the computational domain simulated by 11 grid points.

The forward modelled pressure fields are saved in memory only at every 10-th time sample and at every 2-nd point in x- and z-direction.

3.3 Solution, uncertainty and sensitivity

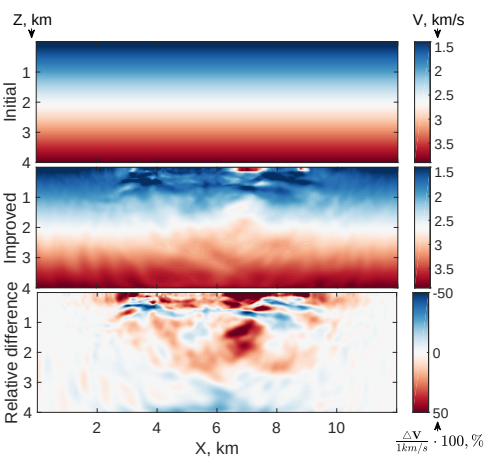
3.3.1 Updating the linear gradient starting models

In the following examples I used only 57 shot gathers (the shot gathers 1, 2, 60 and 61 are not used), a computational grid of 800×2400 points corresponding to 4×12 km, spatial interval of 5 m and the time sampling 0.0007 s. To avoid numerical dispersion and instability of the finite-difference algorithm, the upper and lower limits for sound velocity are set to $V_{max} = 3900$ m/s and $V_{min} = 1000$ m/s.

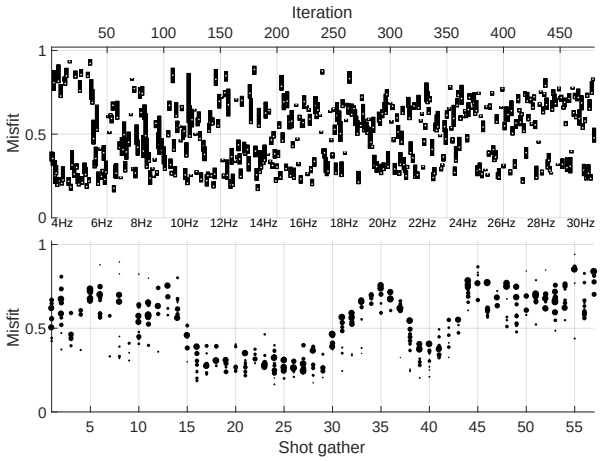
I performed four experiments with the linear gradient starting models: $V_1 = 1300 + 0.65z$ m/s, $V_2 = 1500 + 0.6z$ m/s, $V_3 = 1700 + 0.55z$ m/s and $V_4 = 1800 + 0.525z$ m/s, where z is in m. I keep the same mass density $\rho = 1000 + 0.6z$ kg/m^3 in each experiment. In all examples the maximum value of sound velocity is equal to the upper limit of sound velocity in the computational domain $V_{max} = 3900$ m/s.

Figures 3.5a, 3.6a, 3.7a and 3.8a show the inverted models after 100 BMWI iterations. If unrealistic high- or low-velocity artefacts arise at the top of the model during the initial stages of inversion as shown at Figures 3.5a and 3.8a, then the initial estimate of the governing law is falsified and it does not make any sense to continue inversion.

Figures 3.5b, 3.6b, 3.7b and 3.8b show the corresponding single-shot misfit plots for 450 iterations. In Figure 3.5b the misfit values evaluated at different frequency stages and different shot gathers vary in the range 0.25-0.9, i.e., the misfits are conflicting. The misfit values for the shot gathers 16-29 have consistently lower values than the other misfit functions. Figures 3.6b and 3.7b are the examples of a good choice of a starting model. The misfit values of the data in the frequency range between 3 and 14 Hz are consistently low. At lower and higher frequencies the values of the misfits are higher which means that the estimate of the governing law is inconsistent with the

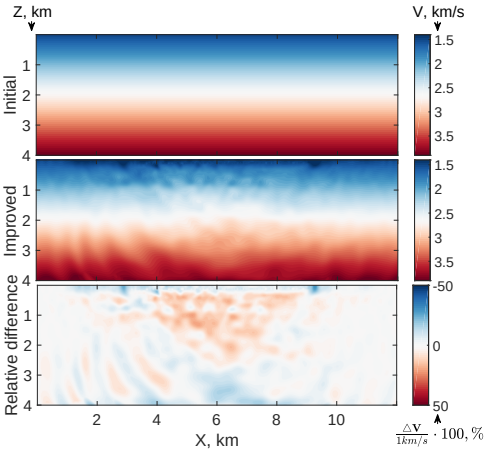


(a) After 100 BMWI iterations.

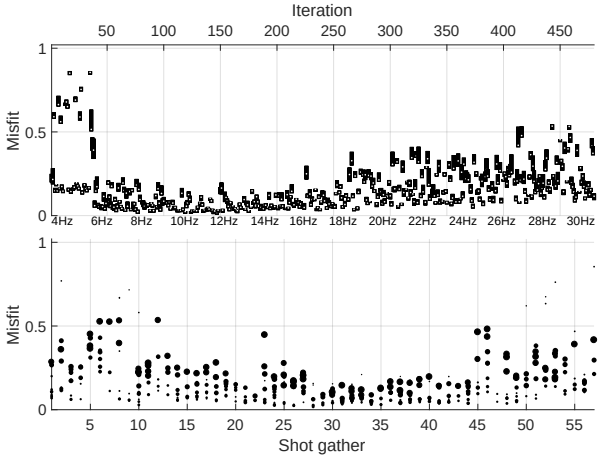


(b) Single-shot ℓ_2^2-misfits.

FIGURE 3.5: The model and misfit domains. The initial model is $V_1 = 1300 + 0.65z$ m/s.

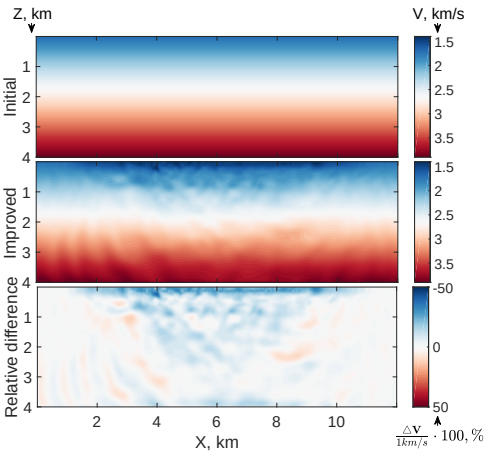


(a) After 100 BMWI iterations.

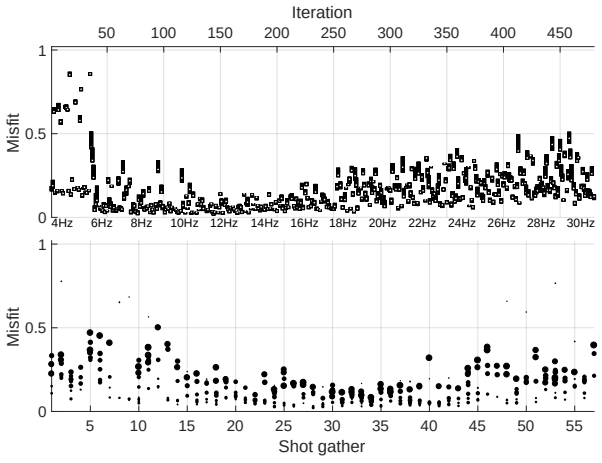


(b) Single-shot ℓ_2^2-misfits.

FIGURE 3.6: The model and misfit domains. The initial model is $V_2 = 1500 + 0.6z$ m/s.

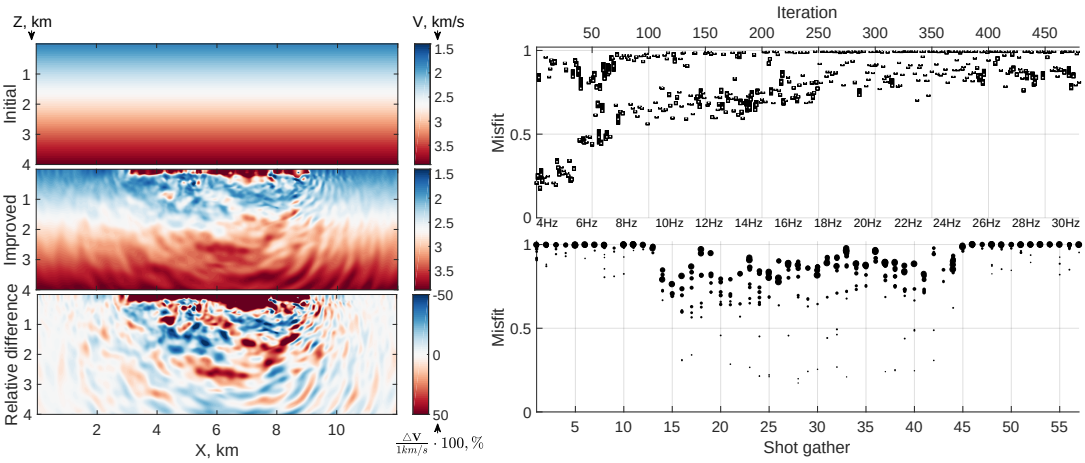


(a) After 100 BMWI iterations.



(b) Single-shot ℓ_2^2-misfits.

FIGURE 3.7: The model and misfit domains. The initial model is $V_3 = 1700 + 0.55z$ m/s.



(a) After 100 BMWI iterations.

(b) Single-shot ℓ_2^2-misfits.

FIGURE 3.8: The model and misfit domains. The initial model is $V_4 = 1800 + 0.525z$ m/s.

observed data. Figure 3.8b shows a result of a poor choice of the starting model. The misfits are highly conflicting and the algorithm diverges.

The plots in the model and misfit domains are complementary. The large misfit values for the shot gathers 30-37 in Figure 3.5b correspond to the artefacts in the model domain between 6 and 8 km in Figure 3.5a. The large misfit values for the shot gathers 12 and 46 in Figure 3.6b correspond to the artefacts at the top of the model at 3 and 9 km (the edges of the OBC).

The stochastic single-shot BMWI, started with the sound velocities $V_2 = 1500 + 0.6z$ m/s and $V_3 = 1700 + 0.55z$ m/s, improved the data fit in the frequency range 3-14 Hz. I use these starting models for further uncertainty analysis.

3.3.2 The neglected 3D effects estimated by the 3D-to-2D transforms

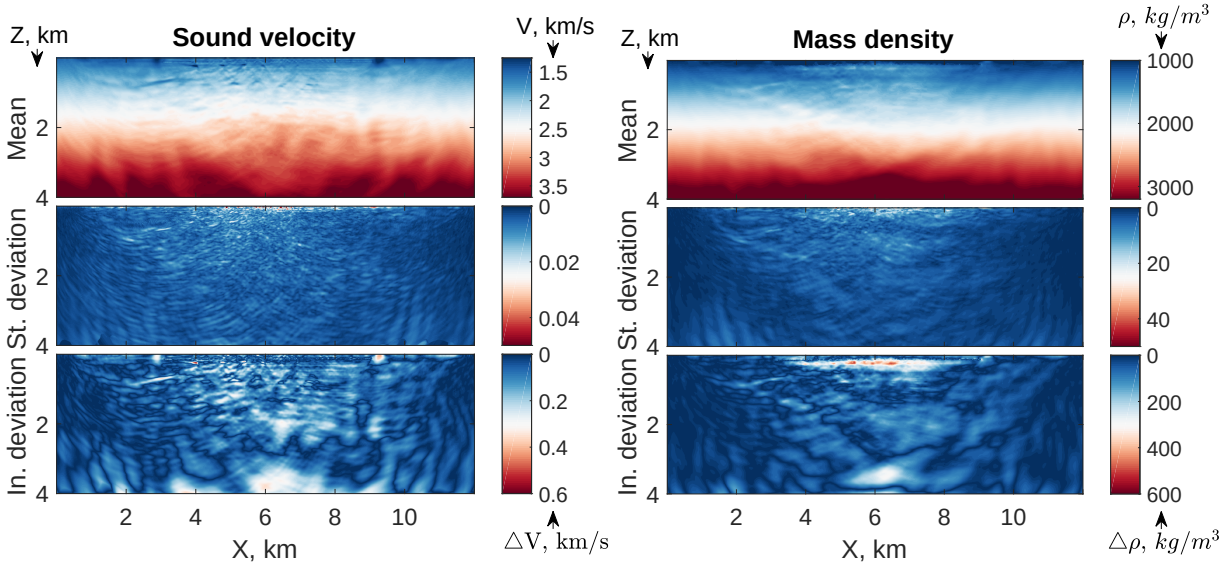
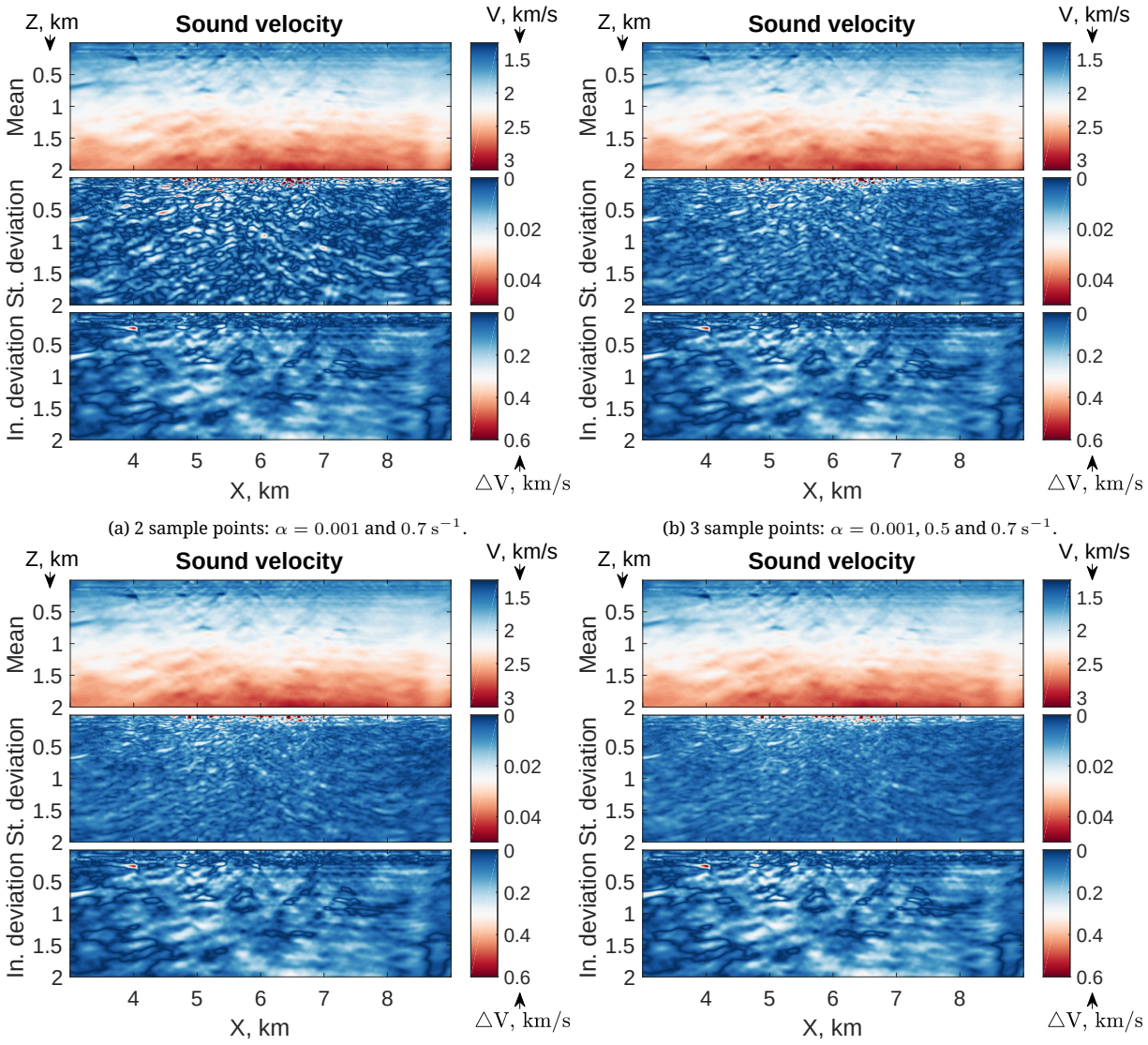


FIGURE 3.9: Sample estimates of the solution, uncertainty and sensitivity after 980 BMWI iterations of the data 3D-to-2D-converted with $\alpha = 0.001, 0.1, 0.3, 0.5, 0.7, 0.9$ and 1.1 s^{-1}. The full computational domain is shown.

To estimate the theoretical uncertainties due to neglected 3D effects in a 2D forward solver, I applied the stochastic single-shot BMWI to the data sets converted with different 3D-to-2D transforms using formula (3.2) with various values of the velocity gradient α.

I ran the same random sequence of descent directions in the stochastic single-shot BMWI for the seven data sets 3D-to-2D-transformed with $\alpha = 0.001$, 0.1, 0.3, 0.5, 0.7, 0.9 and $1.1\ \mathrm{s}^{-1}$ using the linear gradient starting models $V = 1500 + 0.6z\ \mathrm{m/s}$ and $\rho = 1000 + 0.6z\ \mathrm{kg/m^3}$.



(a) 2 sample points: $\alpha = 0.001$ and $0.7\ \mathrm{s}^{-1}$.

(b) 3 sample points: $\alpha = 0.001$, 0.5 and $0.7\ \mathrm{s}^{-1}$.

(c) 5 sample points: $\alpha = 0.001$, 0.3, 0.5, 0.7 and $0.9\ \mathrm{s}^{-1}$.

(d) 7 sample points: $\alpha = 0.001$, 0.1, 0.3, 0.5, 0.7, 0.9 and $1.1\ \mathrm{s}^{-1}$.

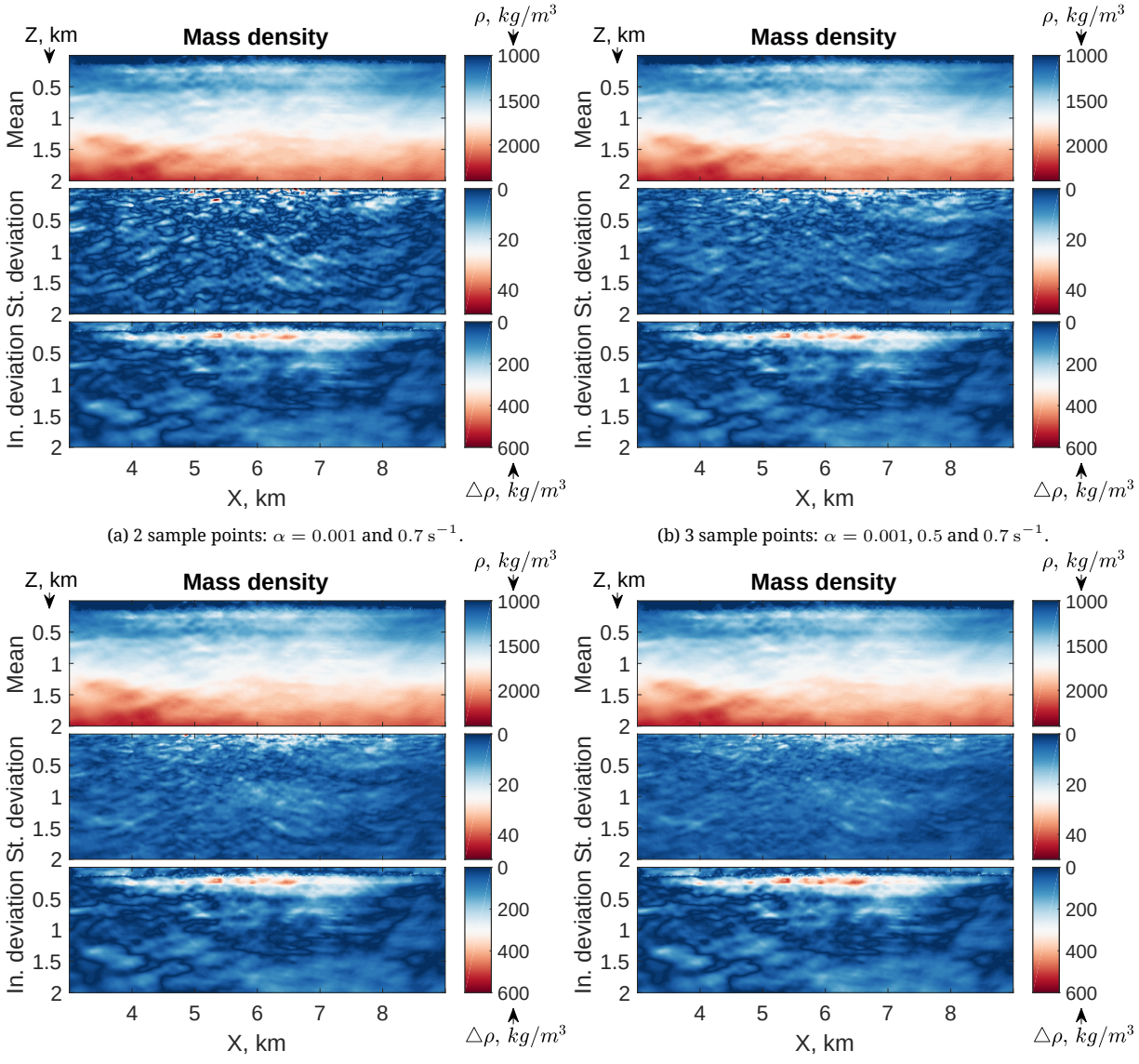
FIGURE 3.10: Sample estimates of the solution, uncertainty and sensitivity using the sound velocities inverted by 980 BMWI iterations of the data 3D-to-2D-converted with different α. A part of the computational domain is shown.

Figure 3.9 shows the sample estimates of the solution, uncertainty and sensitivity for the sound velocity and mass density in the full computational domain. I used 980 single-shot BMWI iterations and all seven data sets. In the water layer the parameters have the highest uncertainty with the values of the standard deviation up to 40-110 $\mathrm{m/s}$ for sound velocity (roughly 3-7 $\%$ of the sound velocity in water) and up to 30-50 $\mathrm{kg/m^3}$ for mass density (roughly 5 $\%$ of the mass density of water). In the rest of the computational domain the standard deviation does not exceed 10-20 $\mathrm{m/s}$ for sound velocity and 10-20 $\mathrm{kg/m^3}$ for mass density.

The large values of the initial deviation at the bottom of the model in Figure 3.9 is an indication that the chosen computational domain is too small. The large values of the initial deviation of sound velocity and the artefacts in the water layer at the edges of the OBC in Figure 3.9 are related

to the 3D-to-2D transform of the data. This becomes clear later in the experiments in a larger computational domain with the two data sets 3D-to-2D transformed differently.

Figures 3.10 and 3.11 show the sample estimates of the solution, uncertainty and sensitivity of the sound velocity and mass density for different numbers of the sample points obtained after 980 BMWI iterations of the data converted with the different 3D-to-2D transforms. To improve visibility, I show only a part of the computational domain. The estimates of uncertainty are smoother for the larger number of sample points. The parameters in the water layer have the highest uncertainty independent of the number of the sample points. The estimates of the solution and sensitivity are repeatable for any number of sample points.



(a) 2 sample points: $\alpha = 0.001$ and $0.7\ \mathrm{s}^{-1}$.

(b) 3 sample points: $\alpha = 0.001, 0.5$ and $0.7\ \mathrm{s}^{-1}$.

(c) 5 sample points: $\alpha = 0.001, 0.3, 0.5, 0.7$ and $0.9\ \mathrm{s}^{-1}$.

(d) 7 sample points: $\alpha = 0.001, 0.1, 0.3, 0.5, 0.7, 0.9$ and $1.1\ \mathrm{s}^{-1}$.

FIGURE 3.11: Sample estimates of the solution, uncertainty and sensitivity using the mass densities inverted by 980 BMWI iterations of the data 3D-to-2D-converted with different α. A part of the computational domain is shown.

The spatially localised perturbations of sound velocity at depths 0.2-0.3 km have values as low as 1.1 $\mathrm{km/s}$. A few spatially localised layers of sound velocity at depths 0.6-0.7 km have values less than 1.5 $\mathrm{km/s}$. The strong contrast of mass density is reconstructed at the bottom of the water layer. At 0.6 km depth the layer with relatively low values ($\approx$ 1300 $\mathrm{kg/m^3}$) of mass density is inverted.

To get a rough estimate of the uncertainty related to the neglected 3D effects, the two sample points inverted by the stochastic single-shot BMWI are sufficient. For example, the data set can

be 3D-to-2D-transformed with $\alpha \to 0\ \mathrm{s}^{-1}$, which is applicable for the direct arrivals in the water, and with $\alpha \approx 0.55 - 0.6\ \mathrm{s}^{-1}$ corresponding to the estimated velocity gradient of the medium (see the previous subsection), which is appropriate for the diving waves. In the last subsection I perform these experiments in a larger computational domain with different starting models.

3.3.3 Random sequences of descent directions

To estimate the solution, uncertainty and sensitivity using the sample points inverted by different random sequences of descent directions in the stochastic single-shot BMWI, I used the field data 3D-to-2D-converted with $\alpha = 0.55\ \mathrm{s}^{-1}$ and the linear-gradient starting models $V = 1500 + 0.6z\ \mathrm{m/s}$ and $\rho = 1000 + 0.6z\ \mathrm{kg/m^3}$.

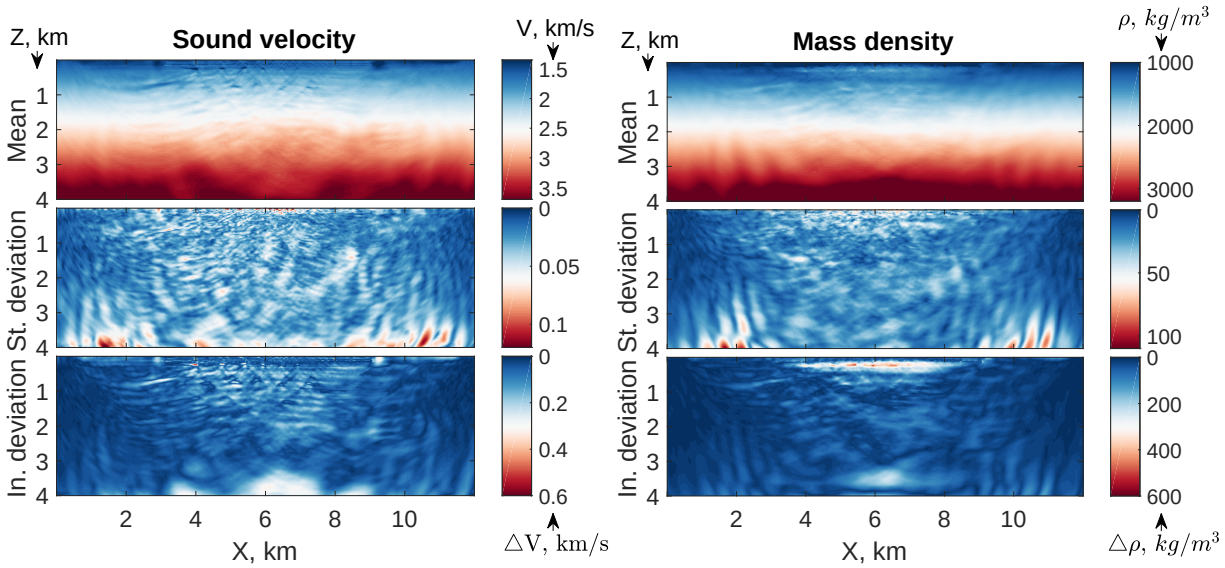
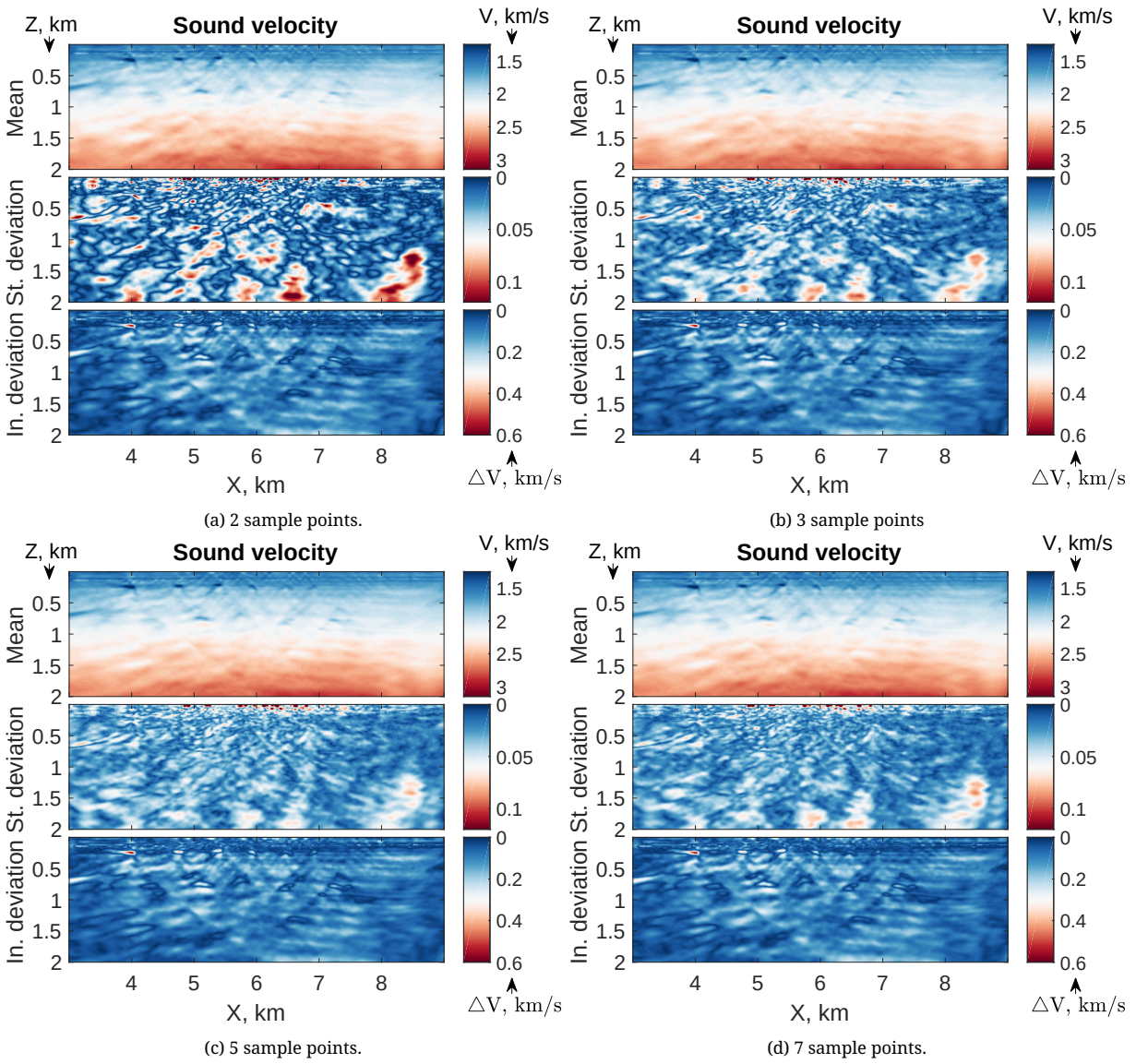


FIGURE 3.12: The sample estimates of the solution, uncertainty and sensitivity after 980 BMWI iterations with seven different random sequences of descent directions. The full computational domain is shown.

Figure 3.12 shows the sample estimates of the solution, uncertainty and sensitivity of the sound velocity and mass density inverted by 980 single-shot BMWI iterations with seven different random sequences of descent directions. The standard deviation (roughly 50-60 $\mathrm{m/s}$ for sound velocity and 50-60 $\mathrm{kg/m^3}$ for mass density) estimated using the different random sequences of descent directions is higher than the standard deviation (roughly 10-20 $\mathrm{m/s}$ for sound velocity and 10-20 $\mathrm{kg/m^3}$ for mass density) estimated using the different 3D-to-2D transformed field data sets computed with the same stochastic algorithm. The parameters with the highest uncertainty correspond to the water layer and to the bottom of the computational domain. The high uncertainty and high sensitivity of the parameters at the bottom of the model in Figure 3.12 is an indication of the small size of the computational domain.

The dependence of the sample estimates of the solution, uncertainty and sensitivity on the number of sample points is shown in Figures 3.13 and 3.14 for sound velocity and mass density, respectively. The larger the number of sample points, the lower the value of the strong perturbations in the standard deviation and the higher the value of the background standard deviation for both sound velocity and mass density. Increasing the number of sample points decreases the standard deviation in the lower parts of the computational domain. The values of the standard deviation in the upper part of the computational domain are relatively high independent of the number of sample points, for both sound velocity and mass density. The standard deviation in the water layer exceeds 100 $\mathrm{m/s}$ for sound velocity and 70 $\mathrm{kg/m^3}$ for mass density.

Due to the relatively small uncertainties, the estimates of the solution and sensitivity are repeatable for any number of sample points and similar to those obtained previously with the data sets 3D-to-2D transformed with various values of the velocity gradient α.



(a) 2 sample points.

(b) 3 sample points

(c) 5 sample points.

(d) 7 sample points.

FIGURE 3.13: Sample estimates of the solution, uncertainty and sensitivity of sound velocity using different numbers of sample points inverted by 980 BMWI iterations. A part of the computational domain is shown.

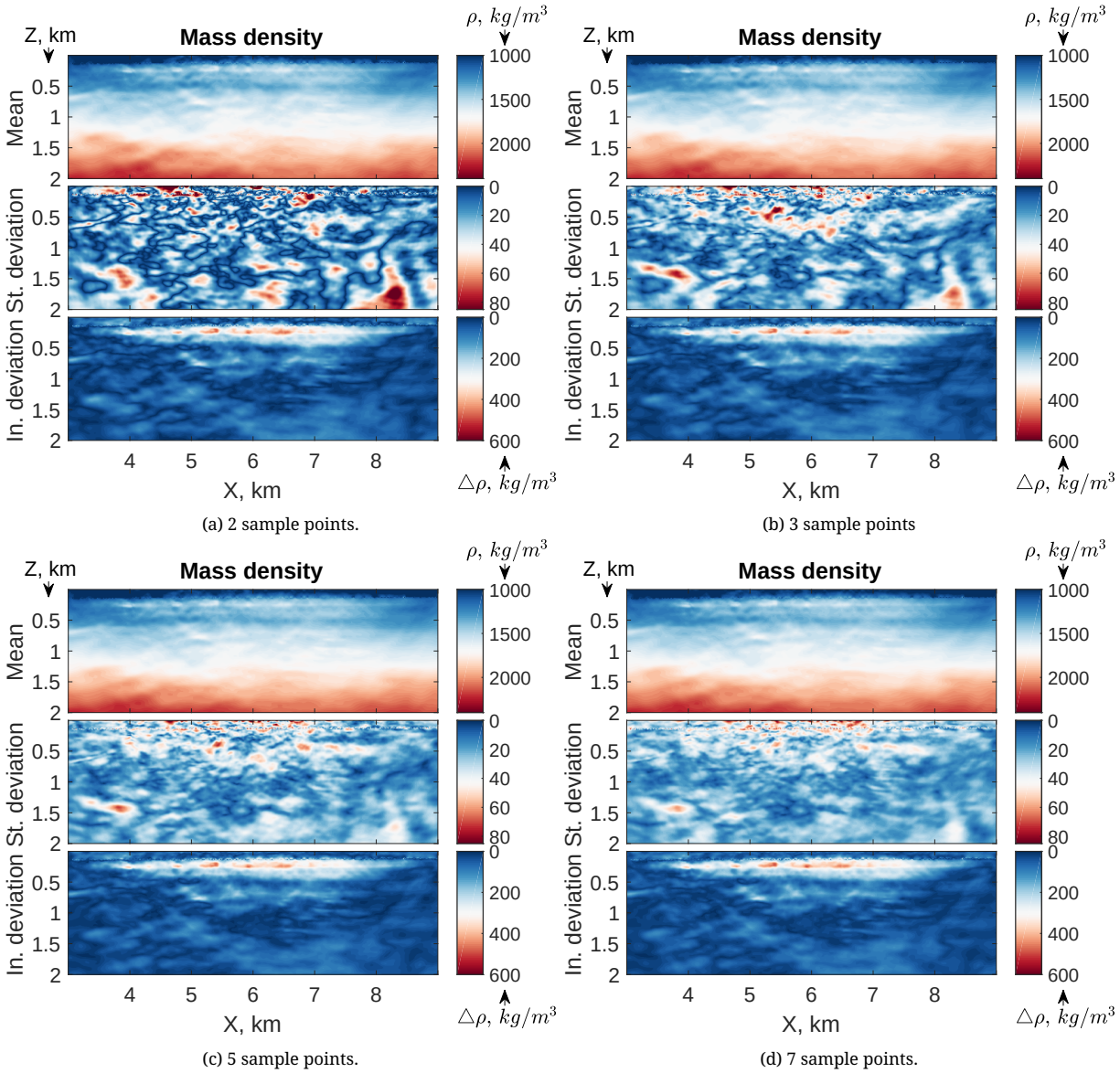
The initial deviation of sound velocity clearly shows the high-contrast areas and faults. The sensitivity estimate, given by the initial deviation of the estimated models with respect to the starting model, is a non-linear image of the subsurface in contrast to a linear image provided by a seismic reverse-time migration.

Three sample points are sufficient to get an estimate of the uncertainty using different runs of the stochastic single-shot BMWI. If the computational resources are very limited, just two different runs of BMWI can be used to get a rough estimate of the uncertainty.

The uncertainties estimated by the different random runs of BMWI are higher than the uncertainties estimated by the same random sequence of BMWI iterations applied to the data sets 3D-to-2D-transformed with different values of the velocity gradient α.

3.3.4 Averaging over the starting models

To obtain more reliable estimates of the solution, uncertainty and sensitivity, I averaged the results over the different starting models, the different 3D-to-2D transforms and the different random sequences of descent directions in BMWI.



(a) 2 sample points.

(b) 3 sample points

(c) 5 sample points.

(d) 7 sample points.

FIGURE 3.14: Sample estimates of the solution, uncertainty and sensitivity of mass density using the different number of sample points inverted by 980 BMWI iterations. A part of the computational domain is shown.

I used all 61 shot gathers, an increased computational grid of 1200×3000 points corresponding to 6×15 km, a spatial interval of 5 m and the time sampling 0.0005 s. To avoid numerical dispersion and instability of the finite-difference algorithm, the upper and lower limits for sound velocity are set to $V_{max} = 5300$ m/s and $V_{min} = 1000$ m/s, respectively.

Seven different linear gradient starting models are given in Table 3.2. All starting models have the same value of the gradient equal to 0.575 s^{-1} for sound velocity and to 0.575 kg/m^{-4} for mass density. The sound velocity models V_k and V_{k+1} ($k = 1, ..., 6$) differ by 50 m/s at any fixed depth. The mass density models ρ_k and ρ_{k+1} ($k = 1, ..., 6$) differ by 150 kg/m^3 at any fixed depth.

The raw pressure seismograms were 3D-to-2D-transformed with the kernel 1) $\mathcal{F}_h$ strictly valid for a homogeneous medium with the sound velocity 1500 m/s, and 2) $\mathcal{F}_g$ strictly valid for the linear gradient medium with the sound velocity 1500 m/s at zero depth and the gradient $\alpha = 0.575$ s^{-1}.

Figures 3.15 and 3.16 show the inverted sample points and the sample estimates of the solution, uncertainty and sensitivity using the two data sets 3D-to-2D-transformed with the kernels $\mathcal{F}_h$ and $\mathcal{F}_g$, respectively. The sample points of sound velocity and mass density are inverted by different random sequences of 980 single-shot BMWI iterations using the different starting models described

Sound velocity, $\mathrm{m/s}$	Mass density, $\mathrm{kg/m^3}$
$V_1 = V_8 = 1450 + 0.575z$	$\rho_1 = \rho_8 = 850 + 0.575z$
$V_2 = V_9 = 1500 + 0.575z$	$\rho_2 = \rho_9 = 1000 + 0.575z$
$V_3 = V_{10} = 1550 + 0.575z$	$\rho_3 = \rho_{10} = 1150 + 0.575z$
$V_4 = V_{11} = 1600 + 0.575z$	$\rho_4 = \rho_{11} = 1300 + 0.575z$
$V_5 = V_{12} = 1650 + 0.575z$	$\rho_5 = \rho_{12} = 1450 + 0.575z$
$V_6 = V_{13} = 1700 + 0.575z$	$\rho_6 = \rho_{13} = 1600 + 0.575z$
$V_7 = V_{14} = 1750 + 0.575z$	$\rho_7 = \rho_{14} = 1750 + 0.575z$

TABLE 3.2: The linear gradient starting models of sound velocity and mass density.

in Table 3.2. The sample points of sound velocity are repeatable in a central semicircle with a radius of 3 km. The sample points of mass density are very different. The estimates of mass density are poor. The standard deviation reaches 400-600 $\mathrm{kg/m^3}$ even in the central semicircle where the sensitivity of the measuring procedure with respect to mass density is the highest.

The footprint of the acquisition geometry – the central semicircle with a radius of 3 km equals half the length of the OBC – is sharply visible in the plots of the standard and initial deviations in Figures 3.15 and 3.16. We have seen a similar pattern in chapter 2 with the Marmousi-2 model for the 6 km-long OBC. The uncertainty of sound velocity is the lowest in the central semicircle. The uncertainty of mass density is the highest in the central semicircle. The sensitivity of the measuring procedure with respect to both sound velocity and mass density is the highest in the central semicircle.

Inversion of the data 3D-to-2D transformed with the kernel $\mathcal{F}_g$ strictly valid in a linear-gradient medium leads to higher values of the standard deviation in the central semicircle in comparison to the values of the standard deviation corresponding to the models inverted from the data 3D-to-2D transformed with the kernel $\mathcal{F}_h$ strictly valid in a homogeneous medium. The sample points of sound velocity in the case of $\mathcal{F}_g$ contain many artefacts in the water layer.

Figure 3.17 shows the sample estimates of the solution, uncertainty and sensitivity using 14, 7 and 3 sample points from Figures 3.15 and 3.16. In all three cases the sample estimates are comparable.

Figure 3.18 shows the sample estimates of the solution, uncertainty and sensitivity using the two sample points from Figures 3.15 and 3.16. The sample estimates, based on two sample points, depend strongly on the choice of the sample points. But the footprint of the acquisition geometry is still visible.

Figures 3.19 and 3.20 are the zoomed versions of Figures 3.15, 3.16, 3.17 and 3.18. The mean values of sound velocity are repeatable for any number of sample points. The mean values of mass density are repeatable if the number of sample points is more than two.

Figure 3.21 shows the averaged normalized and non-normalized relative differences between the inverted and the initial sound velocities using all 14 sample points from Figures 3.15 and 3.16. These averaged relative differences are a kind of non-linear structural image of the subsurface in contrast to the linear images obtained by reverse-time migration algorithms. The faults and the low-velocity anomalies are clearly visible.

In this subsection I obtained the estimates of the solution, uncertainty and sensitivity averaged over seven linear-gradient starting models, different random sequences of descent directions in BMWI and using the field data 3D-to-2D-transformed with two different kernels $\mathcal{F}_h$ and $\mathcal{F}_g$. Using the 2D acoustic finite-difference forward solver, I have shown that the measuring procedure is sensitive to the parameters of sound velocity and mass density in the central semicircle of radius 3 km. The parameters outside the central semicircle are not falsifiable within the specified setting. The uncertainty of sound velocity is as low as 50 $\mathrm{m/s}$ in the central semicircle. The uncertainty of mass density is as high as 300-600 $\mathrm{kg/m^3}$ in the central semicircle.

3.4 Summary

I have estimated the governing law of a physical system using a field OBC data set acquired in a shallow-water river delta by an unspecified company. I used the minimal prior information (the linear gradient starting models) and minimal data preprocessing (only a 3D-to-2D transform). I used a 2D acoustic finite-difference forward solver, 980 stochastic single-shot BMWI iterations, the field data set 3D-to-2D-transformed with different kernels, different linear-gradient starting models and different random sequences of descent directions.

The inverted model parameters are interpreted as the sample points of the probabilistic solution to the problem. I estimated the sample mean, standard deviation and initial deviation of the sample points. These sample estimates of the model parameters for the chosen forward solver are the estimates of the solution, uncertainty and sensitivity.

I have shown that the measuring procedure is sensitive to the parameters of sound velocity and mass density in the central semicircle of radius 3 km. The parameters outside the central semicircle are not falsifiable in the specified setting. The standard deviation of sound velocities in the central semicircle is, on average, 50 $\mathrm{m/s}$.

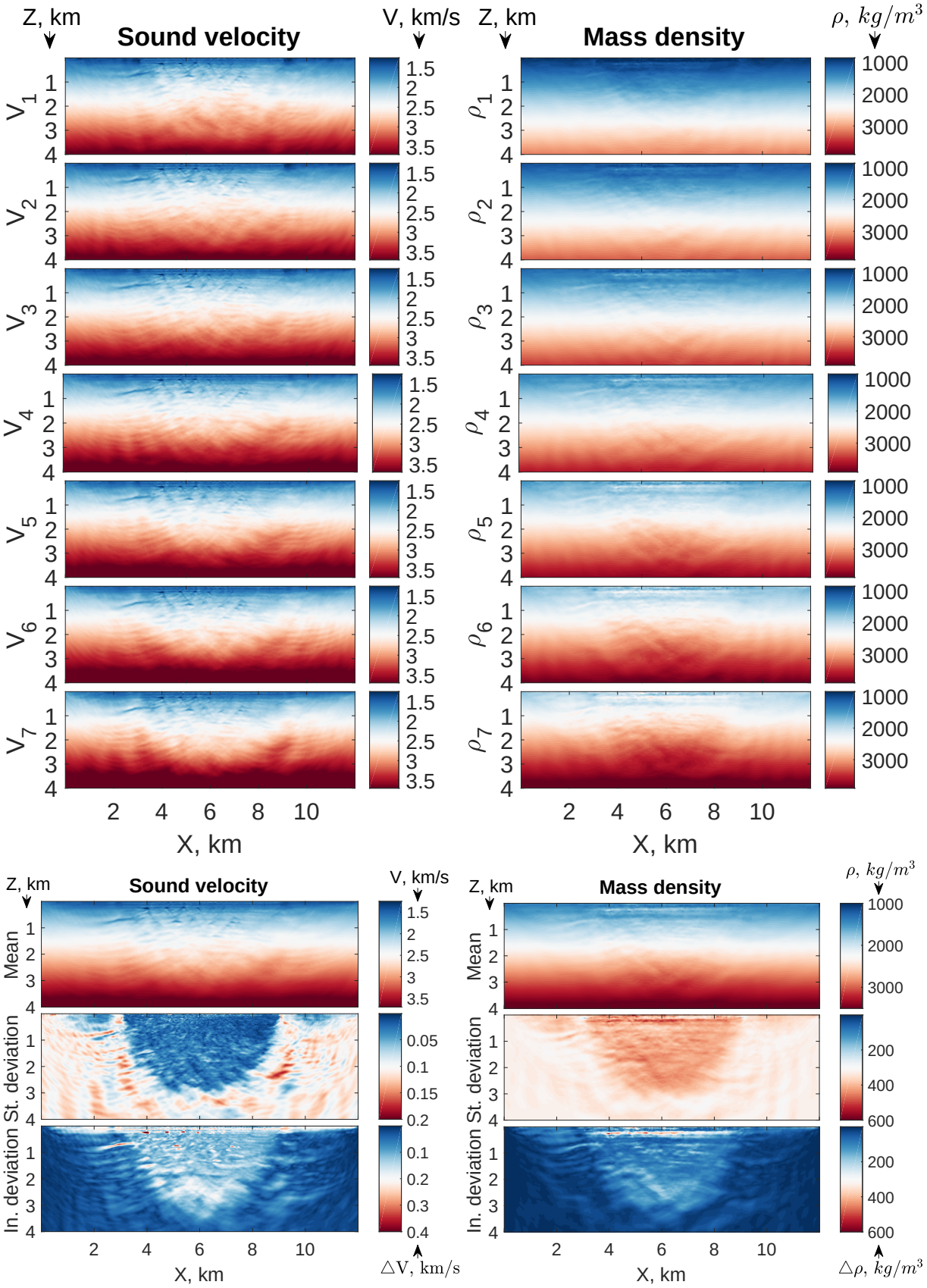


FIGURE 3.15: Sample estimates of the solution, uncertainty and sensitivity using the sample points 1-7 inverted by different sequences of 980 BMWI iterations applied to the data 3D-to-2D transformed with the kernel $\mathcal{F}_h$.

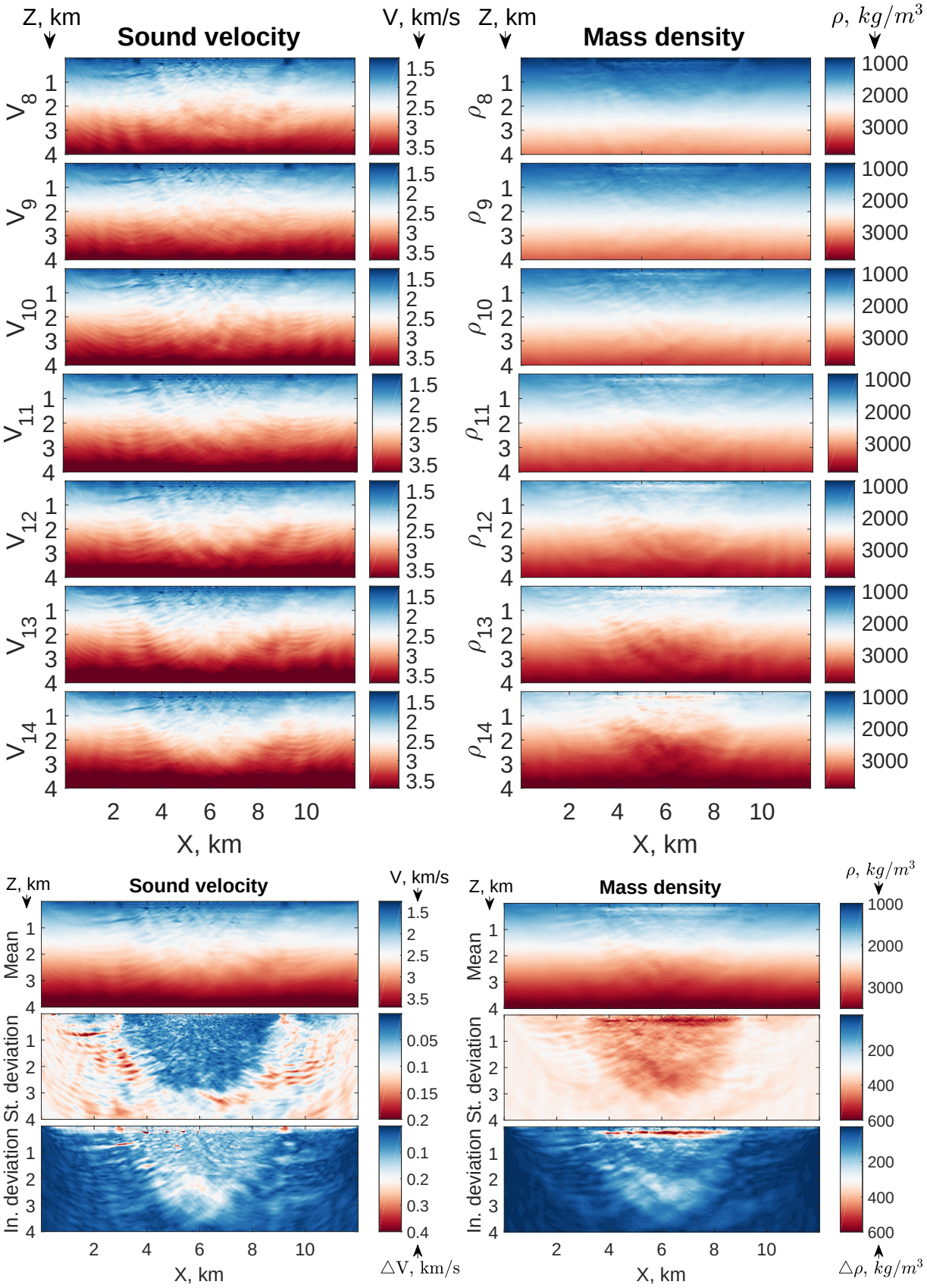
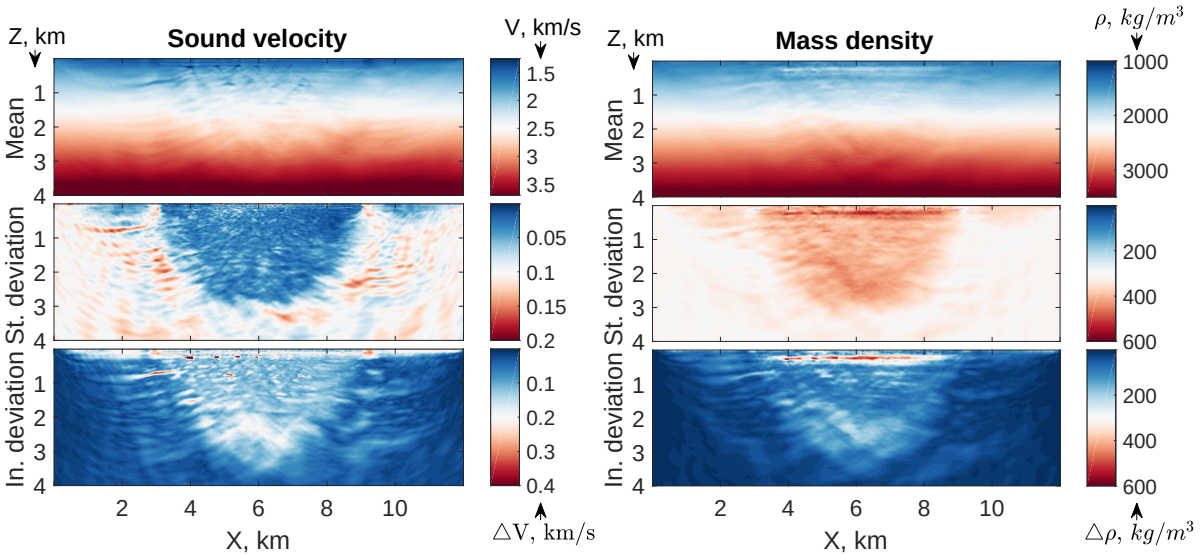
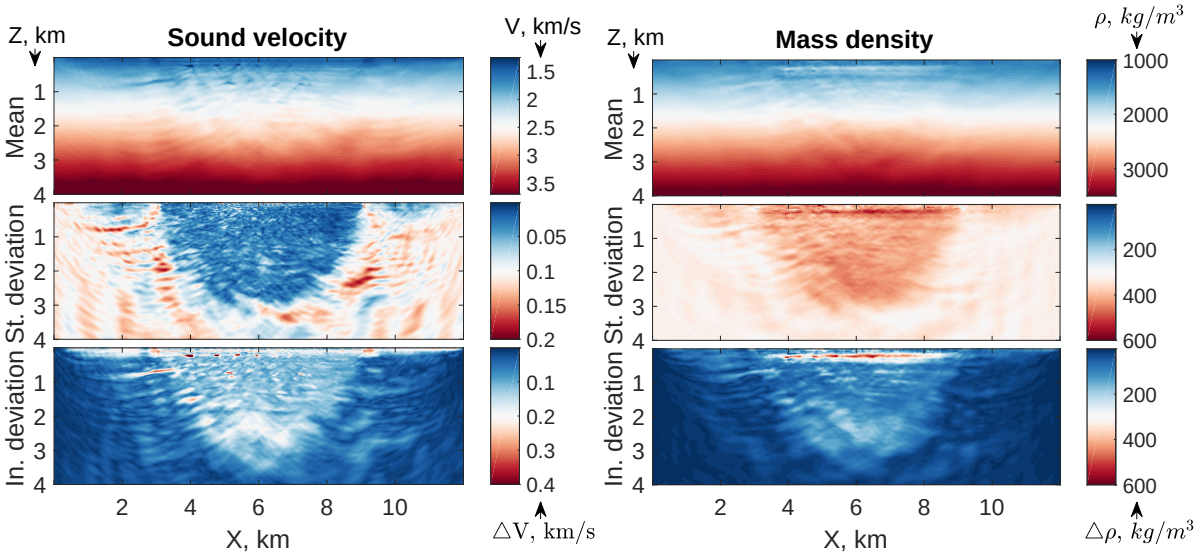


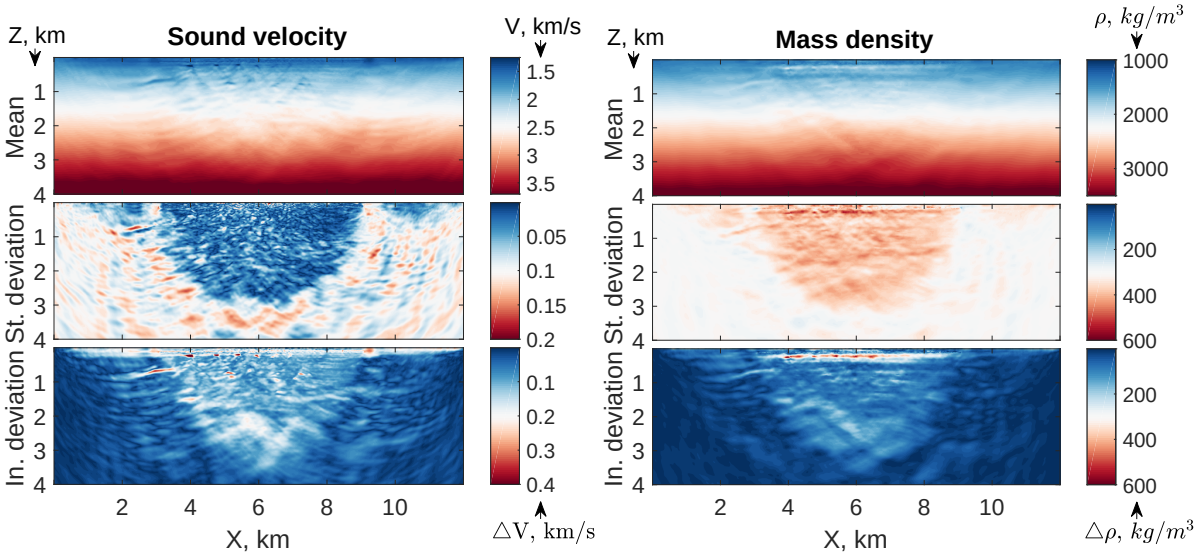
FIGURE 3.16: Sample estimates of the solution, uncertainty and sensitivity using the sample points 8-14 inverted by different sequences of 980 BMWI iterations applied to the data 3D-to-2D transformed with the kernel $\mathcal{F}_g$ and $\alpha = 0.575\ \mathrm{s}^{-1}$.



(a) 14 sample points: 1-st to 7-th from Figure 3.15 and 8-th to 14-th from Figure 3.16.

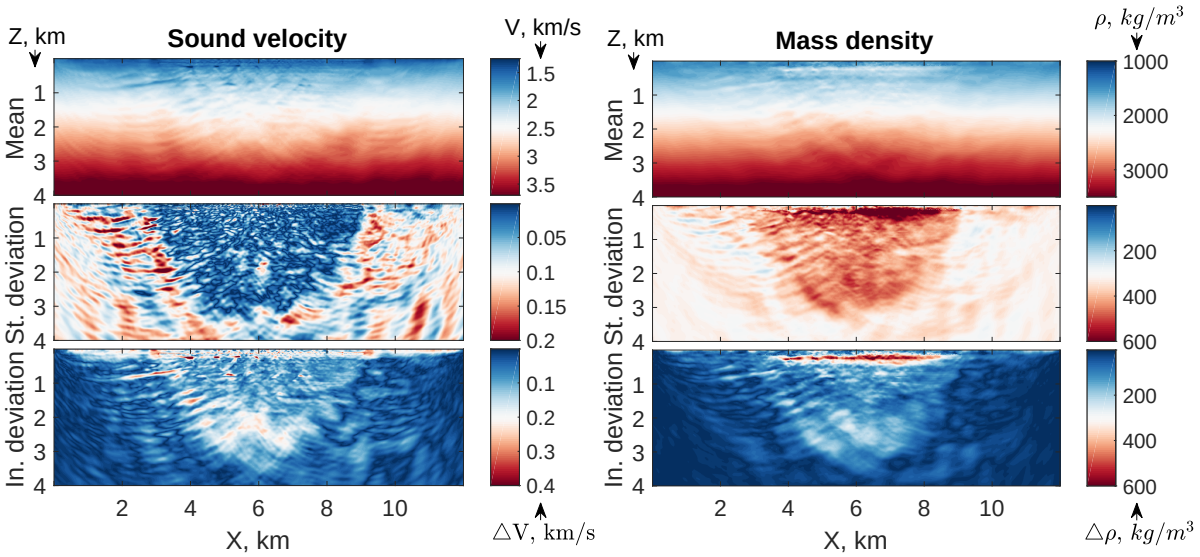


(b) 7 sample points: 1-st, 3-rd, 5-th and 7-th from Figure 3.15 and 9-th, 11-th and 13-th from Figure 3.16.

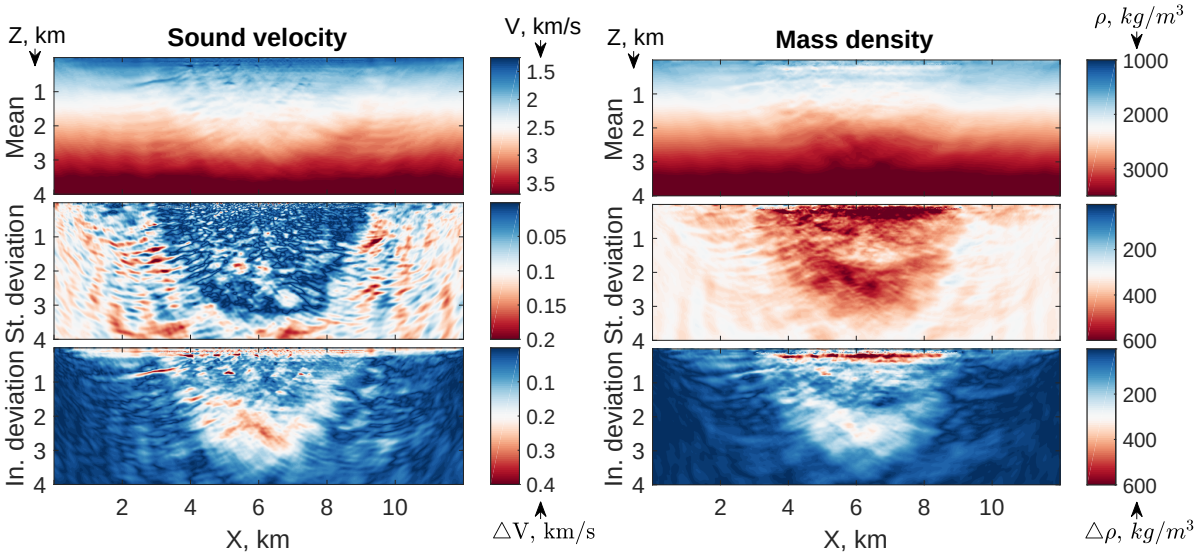


(c) 3 sample points: 2-nd and 6-th from Figure 3.15 and 11-th from Figure 3.16.

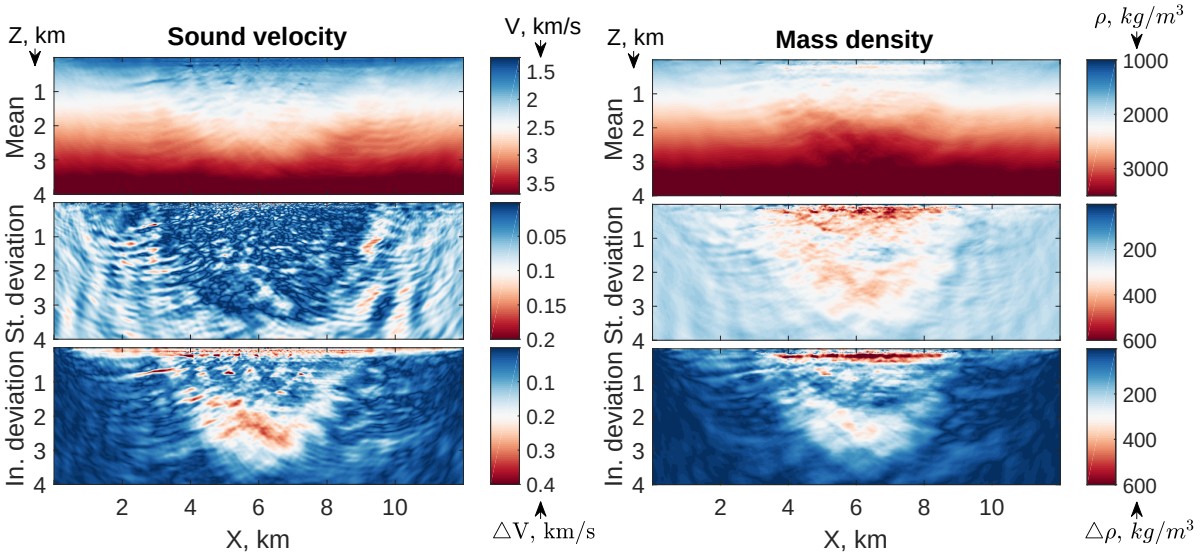
FIGURE 3.17: Sample estimates of the solution, uncertainty and sensitivity using different numbers of sample points from Figures 3.15 and 3.16.



(a) 2 sample points: 3-rd from Figure 3.15 and 13-th from Figure 3.16.

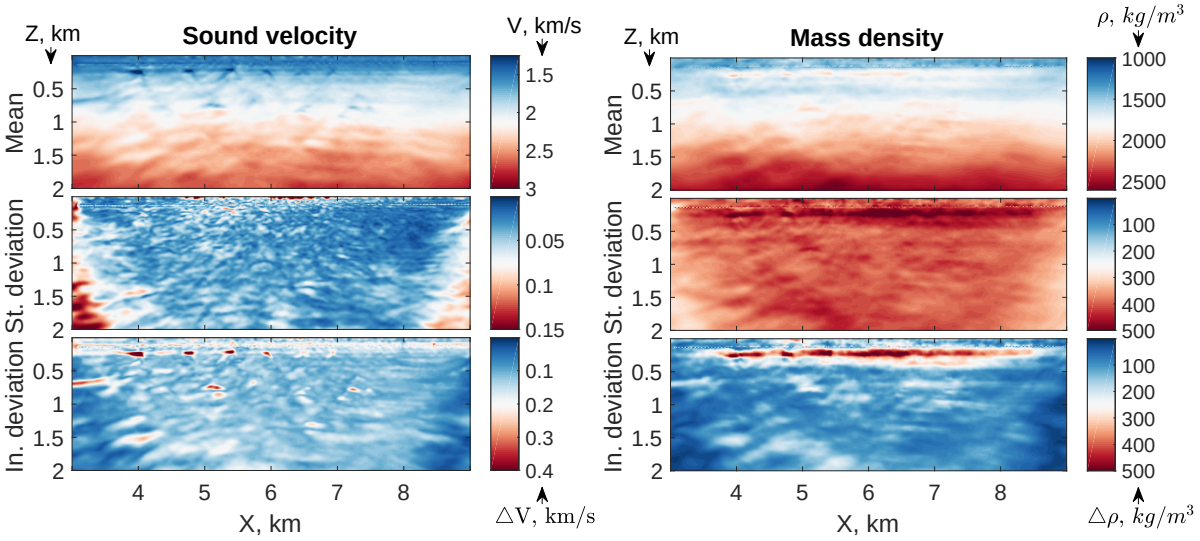


(b) 2 sample points: 4-th from Figure 3.15 and 14-th from Figure 3.16.

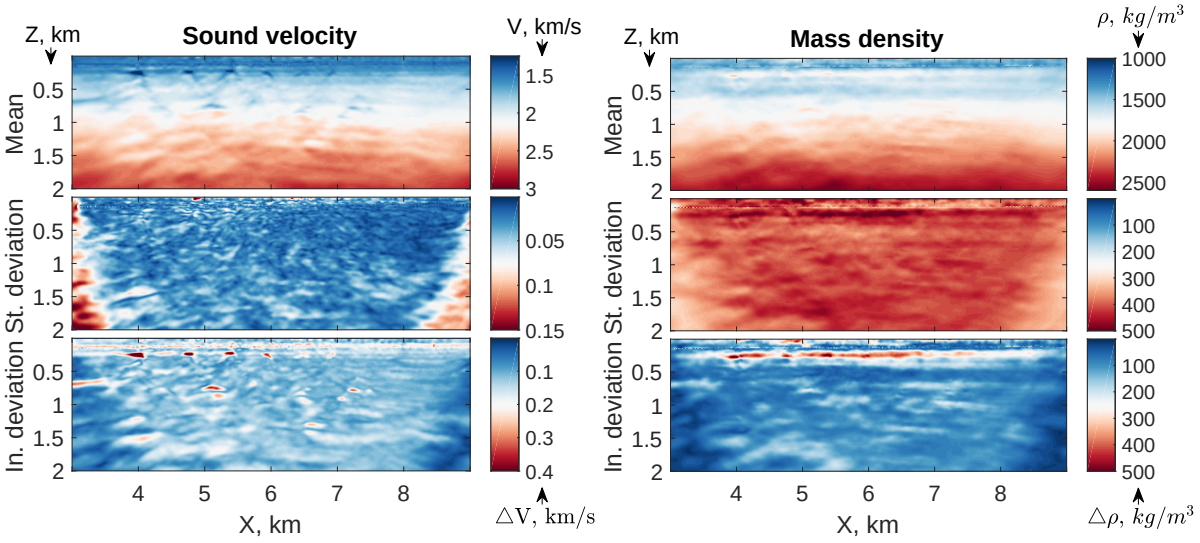


(c) 2 sample points: 5-th from Figure 3.15 and 14-th from Figure 3.16.

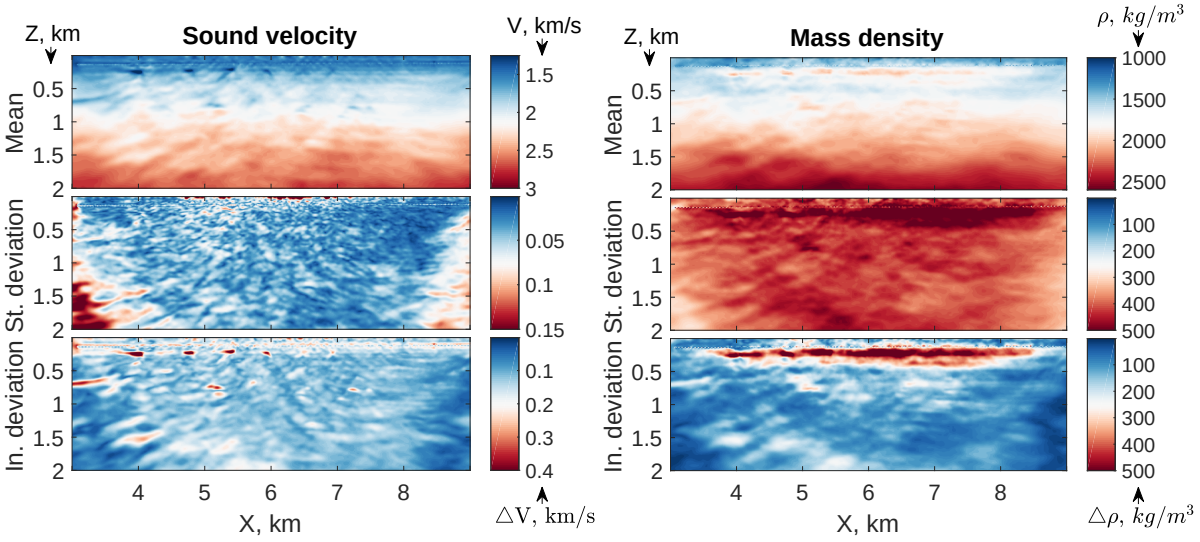
FIGURE 3.18: Sample estimates of the solution, uncertainty and sensitivity using two sample points from Figures 3.15 and 3.16.



(a) 14 sample points: 1-st to 7-th from Figure 3.15 and 8-th to 14-th from Figure 3.16.

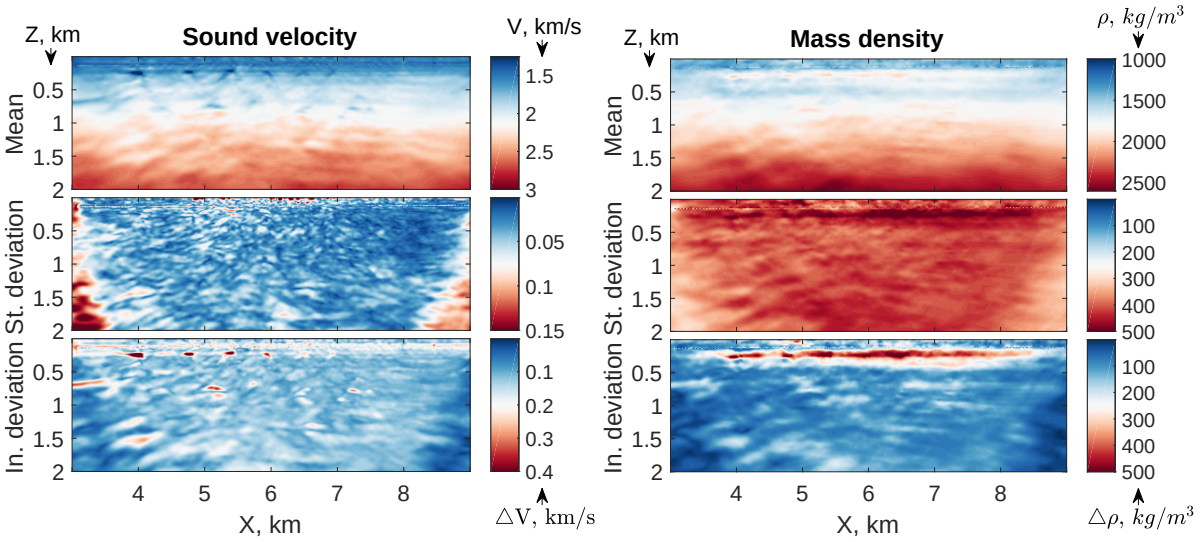


(b) 7 sample points: 1-st to 7-th from Figure 3.15.

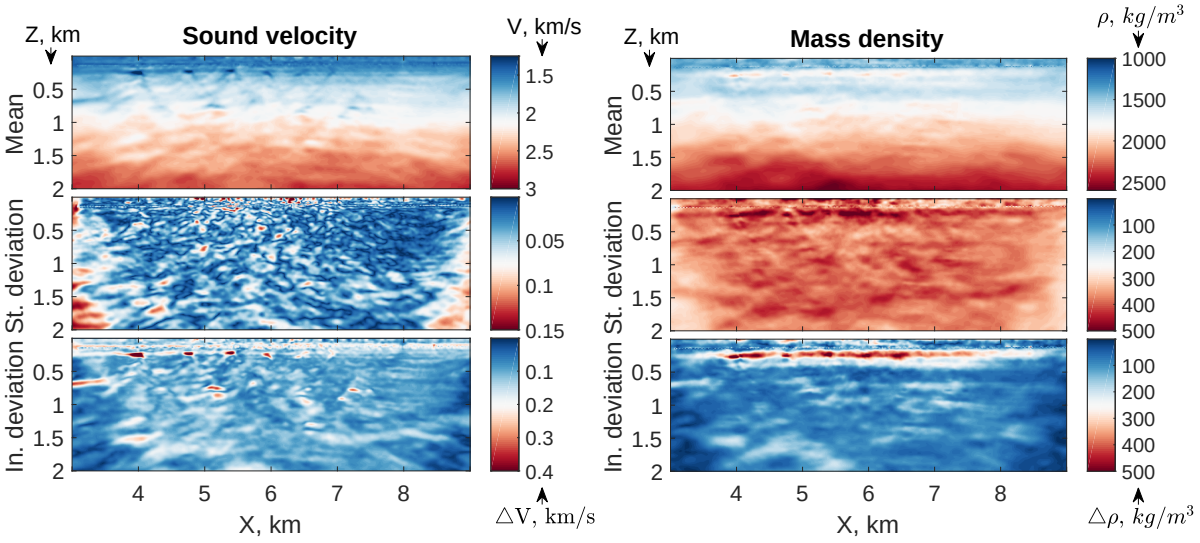


(c) 7 sample points: 8-th to 14-th from Figure 3.16.

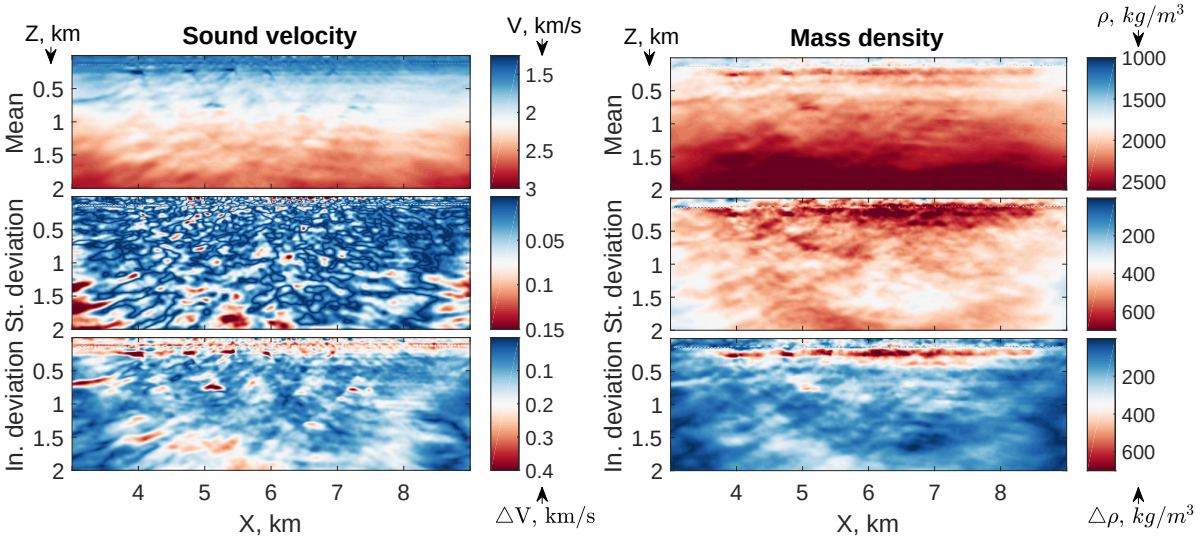
FIGURE 3.19: Sample estimates of the solution, uncertainty and sensitivity. A part of the computational domain is shown.



(a) 7 sample points: 1-st, 3-rd, 5-th and 7-th from Figure 3.15 and 9-th, 11-th and 13-th from Figure 3.16.



(b) 3 sample points: 2-nd and 6-th from Figure 3.15 and 11-th from Figure 3.16.



(c) 2 sample points: 4-th from Figure 3.15 and 14-th from Figure 3.16.

FIGURE 3.20: Sample estimates of the solution, uncertainty and sensitivity. A part of the computational domain is shown.

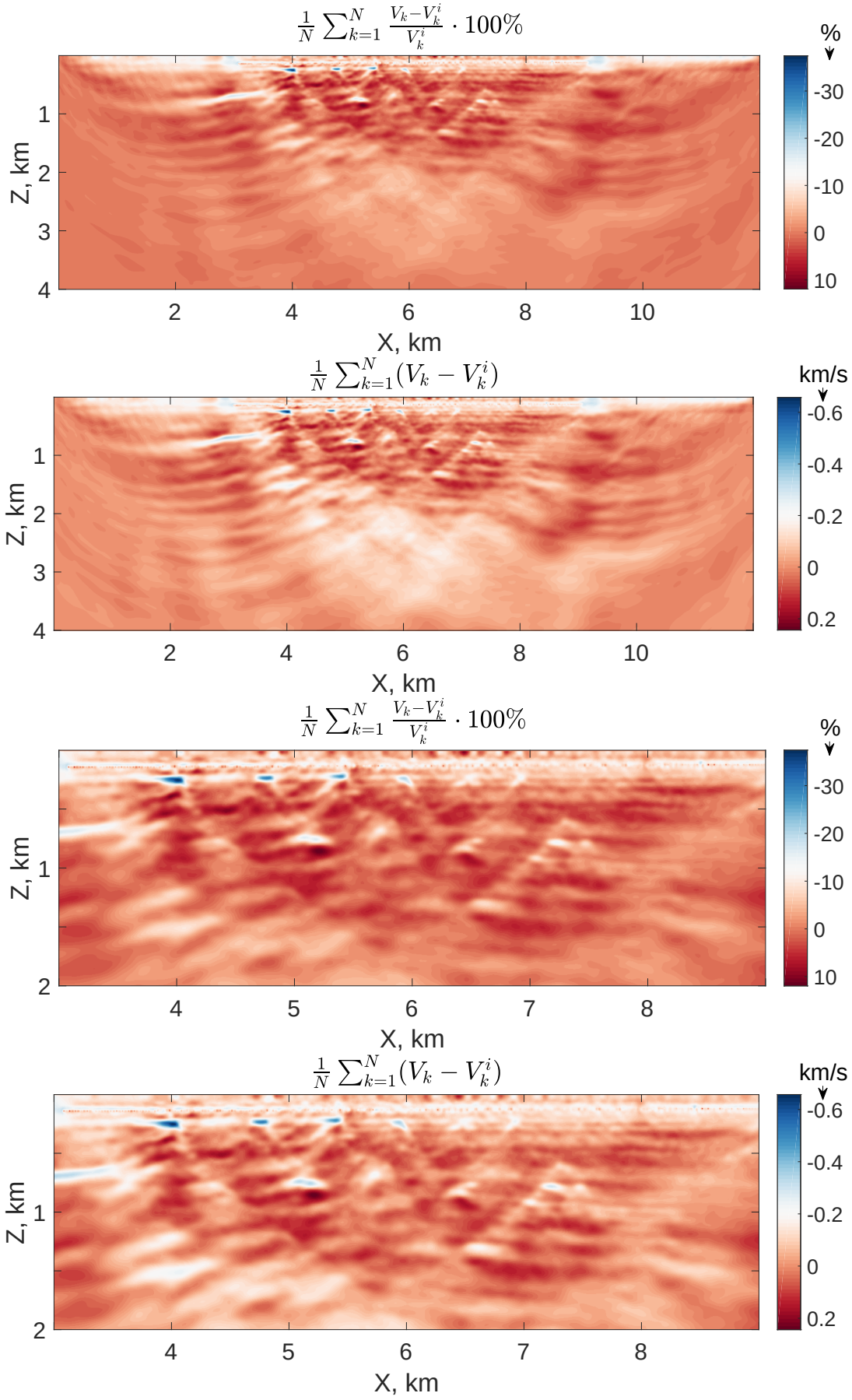


FIGURE 3.21: The averaged normalized and non-normalized relative differences between the inverted and the initial sound velocities. All $N = 14$ sample points of sound velocity were used. V_k denotes the k-th inverted model. V_k^i denotes the k-th starting model.

Conclusions

I proposed a new definition of full-waveform inversion as an algorithm (and a part of the measuring procedure in a wide sense) with the aim to find the governing law of a physical system using the partially measured physical fields with limited computational resources. This definition clearly sets the goal and outlines the key problem: the limitations of the measuring procedure. The measuring procedure in a wide sense includes the experimental procedure, the computational and formal systems.

The thesis statement was: "To find the governing law, guess a law (a forward solver and parametrization), measure some data parameters and check the realizability assumption." The realizability assumption is: the law explains all subsets of the measured data parameters and predicts all subsets of the unmeasured (in the given experiment) data parameters.

The Bayes' theorem, the Tarantola-Valette setting and the principle of maximum entropy using the prior information for the averaged quantities do not allow to formalize the realizability assumption. I proposed a very simple formal system: the principle of minimum relative information using the prior information for non-averaged quantities.

The principle of minimum relative information constrained by the realizability assumption leads to the parametric probabilistic solution (volumetric measure)

$$\sigma(\mathbf{m}, \mathbf{w}) = u(\mathbf{m}) \exp\left(- \sum_{k=1}^{N_\beta} w_k \ell(\mathbf{d}_o^k, \mathcal{F}_o^k(\mathbf{m})) \right)$$

where $\mathbf{w}$ are the arbitrary parameters, ℓ is an arbitrary misfit function and N_β is the number of non-empty subsets of the measured data parameters. In contrast to the Bayes' theorem and the Tarantola-Valette setting, the parametric solution here is not well-posed: the different values of parameters and different misfits lead to different solutions. However, at the points $\mathbf{m}$ of the model manifold $\mathbb{M}$ where the realizability assumption holds, the solution $\sigma(\mathbf{m}, \mathbf{w})$ does not depend on the choice of parameters $\mathbf{w}$ and misfits ℓ.

For any fixed parameters $\mathbf{w}$ and misfits ℓ, guessed theory $\mathcal{F}$ and chosen parametrization $\mathbf{q}$ the volumetric measure $\sigma(\mathbf{m}, \mathbf{w})$ is the solution of the Tarantolian Bayes-Popper problem with minimal prior information (Tarantola, 2007, 2006b). All sample points generated from the uniform prior measure $u(\mathbf{m})$ which have not been accepted when considering the exponential term coming from the realizability assumption, have been falsified.

Maximization of the parametric probabilistic solution $\sigma(\mathbf{m}, \mathbf{w})$ leads to a multiobjective minimization problem

$$\ell(\mathbf{m}, \mathbf{w}) = \sum_{k=1}^{N_\beta} w_k \ell(\mathbf{d}_o^k, \mathcal{F}_o^k(\mathbf{m})),$$

where the weighting parameters $\mathbf{w}$ and misfits ℓ are arbitrary. At the points $\mathbf{m}$ of the model manifold $\mathbb{M}$, where the realizability assumption holds, the misfit $\ell(\mathbf{m}, \mathbf{w})$ is zero independently of the choice of the parameters $\mathbf{w}$ and misfits ℓ. At those points the theory, the model parameters and the measured data parameters are consistent.

If the misfits are conflicting, i.e., the theory, the model parameters and the measured data parameters are inconsistent, the solution of the multiobjective problem is the Pareto optimal set formed by the Pareto optima for which there does not exist another point in the model manifold improving at least one misfit without worsening all the other misfits. The Pareto optima are the points in the model manifold corresponding to the highest values of the parametric probabilistic solution $\sigma(\mathbf{m}, \mathbf{w})$.

The research question of this work was: "What are the solution, uncertainty and sensitivity in full-waveform inversion?" To set the problem of FWI, we guess a theory $\mathcal{F}$, parametrization $\mathbf{q}$, misfit

ℓ, initial model parameters $\mathbf{m}^i$ and measure some data parameters $\mathbf{d}_o$. In this setting I suggest to solve at least partially a local multiobjective minimization problem. The local Pareto optima are the sample points of the parametric volumetric measure $\sigma(\mathbf{m}, \mathbf{w})$. The answer to the research question is: The estimates of the "solution", "uncertainty" and "sensitivity" in FWI are the sample mean, standard deviation and initial deviation of the sample points inverted by the local multiobjective FWI. I would perhaps be criticized for the mean values by Albert Tarantola (a proponent of "the movie strategy"), but I think that if our goal is to find the unique governing law of a physical system under study, then the averaged simplified description of the solution should be sufficient. Because if it is an insufficient description for the deterministic governing law, then the measuring procedure does not allow to find the unique governing law and has to be improved.

In accordance with the no-free-lunch theorems an FWI algorithm has to be adapted to the problem under study, i.e., to a physical system equipped with a measuring procedure. At least in the context of marine exploration experiments the stochastic single-shot bilevel multiobjective waveform inversion (BMWI) performs well with the simplest ℓ_2-norm and linear gradient starting models and does not create artefacts everywhere in the model domain even under cycle-skipped conditions. This allows to invert different subsets of the model parameters using different starting models. The algorithm is easily implementable and easily parallelizable over the different random runs and by domain decomposition. The main ingredients of the algorithm are the single randomly chosen shot gather per iteration, variable projections, an adaptive parabolic line search, a multiscale strategy and diagonal preconditioning of the gradient by the pseudo-Hessian with a negligible regularization.

My experience with BMWI applied to a field data set – a 2D line of pressure data acquired in a shallow-water river delta with ocean bottom cables – convinced me that the chosen definitions for the estimates of the solution, uncertainty and sensitivity are reasonable. These estimates of the uncertainty and sensitivity are non-linear and non-quadratic and sharply highlight a footprint of the acquisition geometry related to the 6 km long ocean bottom cable. Although a very simple idea, the averaging over the different starting models provides indeed more reliable estimates of the solution, uncertainty and sensitivity. The repeatable values of sound velocity in the central semicircle inverted using seven different linear gradient starting models, two differently 3D-to-2D-transformed field data sets and fourteen different random sequences of descent directions, demonstrate the potential of BMWI in a simplified setting. The further studies on BMWI will clarify its domain of applicability.

Bibliography

[1] Afanasiev, M., Peter, D., Sager, K., Simutė, S., Ermert, L., Krischer, L., and Fichtner, A. "Foundations for a multiscale collaborative Earth model". In: *Geophysical Journal International* 204.1 (2016), pp. 39–58. DOI: 10.1093/gji/ggv439.

[2] Agudo, O. C., Guasch, L., Huthwaite, P., and Warner, M. "3D imaging of the breast using full-waveform inversion". In: *Proceedings of the International Workshop on Medical Ultrasound Tomography: 1.-3. Nov. 2017, Speyer, Germany*. KIT Scientific Publishing. 2018, p. 99. DOI: 10.5445/KSP/1000071328.

[3] Alkhalifah, T. *Full Waveform Inversion in an Anisotropic World: Where are the Parameters Hiding?* Education tour series. https://bookshop.eage.org/product/full-waveform-inversion. EAGE Publications, 2014. ISBN: 9789073834835.

[4] Alkhalifah, T. and Plessix, R.-É. "A recipe for practical full-waveform inversion in anisotropic media: An analytical parameter resolution study". In: *Geophysics* 79.3 (2014), R91–R101. DOI: 10.1190/geo2013-0366.1.

[5] Askan, A., Akçelik, V., Bielak, J., and Ghattas, O. "Full Waveform Inversion for Seismic Velocity and Anelastic Losses in Heterogeneous Structures". In: *Bulletin of the Seismological Society of America* 97.6 (2007). DOI: 10.1785/0120070079.

[6] Asnaashari, A., Brossier, R., Garambois, S., Audebert, F., Thore, P., and Virieux, J. "Regularized seismic full waveform inversion with prior model information". In: *Geophysics* 78.2 (2013), R25–R36. DOI: 10.1190/geo2012-0104.1.

[7] Aster, R. C., Borchers, B., and Thurber, C. H. *Parameter Estimation and Inverse Problems*. International Geophysics Series v. 1. Elsevier Academic Press, 2005. ISBN: 9780120656042. DOI: 10.1016/s0074-6142(05)x8014-7.

[8] Backus, G. and Gilbert, F. "The Resolving Power of Gross Earth Data". In: *Geophysical Journal of the Royal Astronomical Society* 16.2 (1968), pp. 169–205. DOI: 10.1111/j.1365-246X.1968.tb00216.x.

[9] Bamberger, A., Chavent, G., Hemon, C., and Lailly, P. "Inversion of normal incidence seismograms". In: *Geophysics* 47.5 (1982), pp. 757–770. DOI: 10.1190/1.1441345.

[10] Bard, J. F. *Practical Bilevel Optimization: Algorithms and Applications (Nonconvex Optimization and Its Applications)*. Springer-Verlag New York, Inc., 1998. ISBN: 0792354583. DOI: 10.1007/978-1-4757-2836-1.

[11] Ben-Hadj-Ali, H., Operto, S., and Virieux, J. "Velocity model building by 3D frequency-domain, full-waveform inversion of wide-aperture seismic data". In: *Geophysics* 73.5 (2008), VE101–VE117. DOI: 10.1190/1.2957948.

[12] Ben-Menahem, A. *Historical Encyclopedia of Natural and Mathematical Sciences*. Springer-Verlag Berlin Heidelberg, 2009. DOI: 10.1007/978-3-540-68832-7.

[13] Ben-Menahem, A. and Singh, S. J. *Seismic waves and sources*. Springer Science & Business Media, 1981. DOI: 10.1007/978-1-4612-5856-8.

[14] Biondi, B. and Almomin, A. "Simultaneous inversion of full data bandwidth by tomographic full-waveform inversion". In: *Geophysics* 79.3 (2014), WA129–WA140. DOI: 10.1190/geo2013-0340.1.

[15] Biondi, B. and Almomin, A. "Tomographic full-waveform inversion (TFWI) by combining FWI and wave-equation migration velocity analysis". In: *The Leading Edge* 32.9 (2013), pp. 1074–1080. DOI: 10.1190/tle32091074.1.

[16] BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP, and OIML. *Evaluation of measurement data – An introduction to the "Guide to the expression of uncertainty in measurement" and related documents*. Tech. rep. Available at https://www.bipm.org/utils/common/documents/jcgm/JCGM_104_2009_E.pdf. Joint Committee for Guides in Metrology, 2009.

[17] BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP, and OIML. *Evaluation of Measurement Data - Guide to the Expression of Uncertainty in Measurement*. Tech. rep. Available at https://www.bipm.org/utils/common/documents/jcgm/JCGM_100_2008_E.pdf. Joint Committee for Guides in Metrology, 2008.

[18] Bleibinhaus, F., Lester, R. W., and Hole, J. A. "Applying waveform inversion to wide-angle seismic surveys". In: *Tectonophysics* 472.1–4 (2009). Deep seismic profiling of the continents and their margins, pp. 238 –248. DOI: 10.1016/j.tecto.2008.08.023.

[19] Boehm, C., Hanzich, M., de la Puente, J., and Fichtner, A. "Wavefield compression for adjoint methods in full-waveform inversion". In: *Geophysics* 81.6 (2016), R385–R397. DOI: 10.1190/geo2015-0653.1.

[20] Bohlen, T. "Parallel 3-D viscoelastic finite difference seismic modelling". In: *Computers & Geosciences* 28.8 (2002), pp. 887 –899. DOI: 10.1016/S0098-3004(02)00006-7.

[21] Borisov, D. and Singh, S. C. "Three-dimensional elastic full waveform inversion in a marine environment using multicomponent ocean-bottom cables: a synthetic study". In: *Geophysical Journal International* 201.3 (2015), pp. 1215–1234. DOI: 10.1093/gji/ggv048.

[22] Bozdağ, E., Trampert, J., and Tromp, J. "Misfit functions for full waveform inversion based on instantaneous phase and envelope measurements". In: *Geophysical Journal International* 185.2 (2011), pp. 845–870. DOI: 10.1111/j.1365-246X.2011.04970.x.

[23] Brenders, A. J. and Pratt, R. G. "Efficient waveform tomography for lithospheric imaging: implications for realistic, two-dimensional acquisition geometries and low-frequency data". In: *Geophysical Journal International* 168.1 (2007), pp. 152–170. DOI: 10.1111/j.1365-246X.2006.03096.x.

[24] Brenders, A. J. and Pratt, R. G. "Full waveform tomography for lithospheric imaging: results from a blind test in a realistic crustal model". In: *Geophysical Journal International* 168.1 (2007), p. 133. DOI: 10.1111/j.1365-246X.2006.03156.x.

[25] Brossier, R. "Two-dimensional frequency-domain visco-elastic full waveform inversion: Parallel algorithms, optimization and performance". In: *Computers & Geosciences* 37.4 (2011), pp. 444 –455. DOI: 10.1016/j.cageo.2010.09.013.

[26] Brossier, R., Operto, S., and Virieux, J. "Seismic imaging of complex onshore structures by 2D elastic frequency-domain full-waveform inversion". In: *Geophysics* 74.6 (2009), WCC105–WCC118. DOI: 10.1190/1.3215771.

[27] Bui-Thanh, T., Ghattas, O., Martin, J., and Stadler, G. "A Computational Framework for Infinite-Dimensional Bayesian Inverse Problems Part I: The Linearized Case, with Application to Global Seismic Inversion". In: *SIAM Journal on Scientific Computing* 35.6 (2013), A2494–A2523. DOI: 10.1137/12089586X.

[28] Bunks, C., Saleck, F. M., Zaleski, S., and Chavent, G. "Multiscale seismic waveform inversion". In: *Geophysics* 60.5 (1995), pp. 1457–1473. DOI: 10.1190/1.1443880.

[29] Calvetti, D. and Somersalo, E. "Inverse problems: From regularization to Bayesian inference". In: *Wiley Interdisciplinary Reviews: Computational Statistics* 10.3 (2018), e1427. DOI: 10.1002/wics.1427.

[30] Castellanos, C., Métivier, L., Operto, S., Brossier, R., and Virieux, J. "Fast full waveform inversion with source encoding and second-order optimization methods". In: *Geophysical Journal International* 200.2 (2015), pp. 720–744. DOI: 10.1093/gji/ggu427.

[31] Charara, M., Barnes, C., and Tarantola, A. "Full waveform inversion of seismic data for a viscoelastic medium". In: *Methods and Applications of Inversion*. Ed. by P. C. Hansen, B. H. Jacobsen, and K. Mosegaard. Springer Berlin Heidelberg, 2000, pp. 68–81. ISBN: 978-3-540-48866-8. DOI: 10.1007/BFb0010284.

[32] Chavent, G. "Identification of functional parameters in partial differential equations". In: *Joint Automatic Control Conference*. 12. https://www.infona.pl/resource/bwmeta1.element.ieee-art-000004170030. 1974, pp. 155–156.

[33] Clapp, R. G. “Reverse time migration: saving the boundaries”. In: *Stanford Exploration Project* 137 (2008). Available at http://sepwww.stanford.edu/data/media/public/docs/sep136/bob4/paper.pdf.

[34] Crase, E., Pica, A., Noble, M., McDonald, J., and Tarantola, A. “Robust elastic nonlinear waveform inversion: Application to real data”. In: *Geophysics* 55.5 (1990), pp. 527–538. DOI: 10.1190/1.1442864.

[35] Crase, E., Wideman, C., Noble, M., and Tarantola, A. “Nonlinear elastic waveform inversion of land seismic reflection data”. In: *Journal of Geophysical Research: Solid Earth* 97.B4 (1992), pp. 4685–4703. DOI: 10.1029/90JB00832.

[36] Dashti, M. and Stuart, A. M. “The Bayesian approach to inverse problems”. In: *Handbook of Uncertainty Quantification* (2017), pp. 311–428. DOI: 10.1007/978-3-319-12385-1_7.

[37] Deb, K. *Multi-Objective Optimization Using Evolutionary Algorithms*. New York, NY, USA: John Wiley & Sons, Inc., 2001. ISBN: 047187339X. DOI: \url{https://books.google.de/books?id=0STn4GSy2uQC}..

[38] Dempe, S. *Foundations of Bilevel Programming*. Nonconvex Optimization and Its Applications. Springer US, 2002. ISBN: 9781402006319. DOI: 10.1007/b101970.

[39] Duane, S., Kennedy, A., Pendleton, B. J., and Roweth, D. “Hybrid Monte Carlo”. In: *Physics Letters B* 195.2 (1987), pp. 216 –222. DOI: 10.1016/0370-2693(87)91197-X.

[40] Dussaud, E., Symes, W. W., Williamson, P., Lemaistre, L., Singer, P., Denel, B., and Cherrett, A. “Computational strategies for reverse-time migration”. In: *SEG Technical Program Expanded Abstracts*. 2008, pp. 2267–2271. DOI: 10.1190/1.3059336.

[41] Eichfelder, G. “Multiobjective bilevel optimization”. In: *Mathematical Programming* 123.2 (2010), pp. 419–449. DOI: 10.1007/s10107-008-0259-0.

[42] Engquist, B. and Froese, B. “Application of the Wasserstein metric to seismic signals”. In: *Communications in Mathematical Sciences* 12.5 (Mar. 2014), pp. 979–988. DOI: 10.4310/CMS.2014.v12.n5.a7.

[43] Fang, Z. “Source estimation and uncertainty quantification for wave-equation based seismic imaging and inversion”. PhD thesis. University of British Columbia, 2018. DOI: 10.14288/1.0366129.

[44] Fang, Z., Silva, C. D., Kuske, R., and Herrmann, F. J. “Uncertainty quantification for inverse problems with weak partial-differential-equation constraints”. In: *Geophysics* 83.6 (2018), R629–R647. DOI: 10.1190/geo2017-0824.1.

[45] Fichtner, A. *Full Seismic Waveform Modelling and Inversion*. Advances in Geophysical and Environmental Mechanics and Mathematics. Springer Berlin Heidelberg, 2011. ISBN: 9783642158070. DOI: 10.1007/978-3-642-15807-0.

[46] Fichtner, A. and Trampert, J. “Hessian kernels of seismic data functionals based upon adjoint techniques”. In: *Geophysical Journal International* 185.2 (2011), pp. 775–798. DOI: 10.1111/j.1365-246X.2011.04966.x.

[47] Fichtner, A. and Trampert, J. “Resolution analysis in full waveform inversion”. In: *Geophysical Journal International* 187.3 (2011), pp. 1604–1624. DOI: 10.1111/j.1365-246X.2011.05218.x.

[48] Fichtner, A. and van Leeuwen, T. “Resolution analysis by random probing”. In: *Journal of Geophysical Research: Solid Earth* 120.8 (2015), pp. 5549–5573. DOI: 10.1002/2015JB012106.

[49] Fichtner, A., Kennett, B. L. N., Igel, H., and Bunge, H.-P. “Full seismic waveform tomography for upper-mantle structure in the Australasian region using adjoint methods”. In: *Geophysical Journal International* 179.3 (2009), pp. 1703–1725. DOI: 10.1111/j.1365-246X.2009.04368.x.

[50] Fichtner, A., Zunino, A., and Gebraad, L. “Hamiltonian Monte Carlo solution of tomographic inverse problems”. In: *Geophysical Journal International* 216.2 (2019), pp. 1344–1363. DOI: 10.1093/gji/ggy496.

[51] Fichtner, A., Trampert, J., Cupillard, P., Saygin, E., Taymaz, T., Capdeville, Y., and Villaseñor, A. “Multiscale full waveform inversion”. In: *Geophysical Journal International* 194.1 (2013), pp. 534–556. DOI: 10.1093/gji/ggt118.

[52] Fichtner, A., van Herwaarden, D.-P., Afanasiev, M., Simutė, S., Krischer, L., Çubuk-Sabuncu, Y., Taymaz, T., Colli, L., Saygin, E., Villaseñor, A., Trampert, J., Cupillard, P., Bunge, H.-P., and Igel, H. "The Collaborative Seismic Earth Model: Generation 1". In: *Geophysical Research Letters* 45.9 (2018), pp. 4007–4016. DOI: 10.1029/2018GL077338.

[53] Fichtner, A., Kennett, B. L. N., Igel, H., and Bunge, H.-P. "Theoretical background for continental- and global-scale full-waveform inversion in the time–frequency domain". In: *Geophysical Journal International* 175.2 (2008), pp. 665–685. DOI: 10.1111/j.1365-246X.2008.03923.x.

[54] Forbriger, T., Groos, L., and Schäfer, M. "Line-source simulation for shallow-seismic data. Part 1: Theoretical background". In: *Geophysical Journal International* 198.3 (2014), pp. 1387–1404. DOI: 10.1093/gji/ggu199.

[55] Friedlander, M. P. and Schmidt, M. "Hybrid Deterministic-Stochastic Methods for Data Fitting". In: *SIAM Journal on Scientific Computing* 34.3 (2012), A1380–A1405. DOI: 10.1137/110830629.

[56] Fu, L. and Symes, W. W. "A discrepancy-based penalty method for extended waveform inversion". In: *Geophysics* 82.5 (2017), R287–R298. DOI: 10.1190/geo2016-0326.1.

[57] Fu, L. and Symes, W. W. "An adaptive multiscale algorithm for efficient extended waveform inversion". In: *Geophysics* 82.3 (2017), R183–R197. DOI: 10.1190/geo2016-0426.1.

[58] Gabbay, D. M., Pelletier, F. J., and Woods, J., eds. *Logic: A History of its Central Concepts*. Vol. 11. Handbook of the History of Logic. Elsevier, 2012. ISBN: 978-0-444-52937-4. DOI: 10.1016/c2009-0-16929-5.

[59] Gao, F., Levander, A., Pratt, R. G., Zelt, C. A., and Fradelizio, G.-L. "Waveform tomography at a groundwater contamination site: Surface reflection data". In: *Geophysics* 72.5 (2007), G45–G55. DOI: 10.1190/1.2752744.

[60] Gauthier, O., Virieux, J., and Tarantola, A. "Two-dimensional nonlinear inversion of seismic waveforms: Numerical results". In: *Geophysics* 51.7 (1986), pp. 1387–1403. DOI: 10.1190/1.1442188.

[61] Geman, S. and Geman, D. "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6 (1984), pp. 721–741. DOI: 10.1109/TPAMI.1984.4767596.

[62] Gholami, Y., Brossier, R., Operto, S., Ribodetti, A., and Virieux, J. "Which parameterization is suitable for acoustic vertical transverse isotropic full waveform inversion? Part 1: Sensitivity and trade-off analysis". In: *Geophysics* 78.2 (2013), R81–R105. DOI: 10.1190/geo2012-0204.1.

[63] Golub, G. and Pereyra, V. "Separable nonlinear least squares: the variable projection method and its applications". In: *Inverse Problems* 19.2 (2003), R1. DOI: 10.1088/0266-5611/19/2/201.

[64] Golub, G. and Pereyra, V. "The Differentiation of Pseudo-Inverses and Nonlinear Least Squares Problems Whose Variables Separate". In: *SIAM Journal on Numerical Analysis* 10.2 (1973), pp. 413–432. DOI: 10.1137/0710036.

[65] Goncharsky, A., Romanov, S., and Seryozhnikov, S. "Inverse problems of 3D ultrasonic tomography with complete and incomplete range data". In: *Wave Motion* 51.3 (2014), pp. 389–404. DOI: 10.1016/j.wavemoti.2013.10.001.

[66] Guitton, A. "Blocky regularization schemes for Full-Waveform Inversion". In: *Geophysical Prospecting* 60.5 (2012), pp. 870–884. DOI: 10.1111/j.1365-2478.2012.01025.x.

[67] Guitton, A. and Diaz, E. "Attenuating crosstalk noise with simultaneous source full waveform inversion". In: *Geophysical Prospecting* 60.4 (2012), pp. 759–768. DOI: 10.1111/j.1365-2478.2011.01023.x.

[68] Habelitz, P. M. "2D akustische Wellenforminversion geführter Wellen im Flachwasser". German. MSc. thesis. Karlsruher Institut für Technologie (KIT), 2017. DOI: 10.5445/IR/1000080198.

[69] Hak, B. and Mulder, W. A. "Migration for velocity and attenuation perturbations". In: *Geophysical Prospecting* 58.6 (2010), pp. 939–952. DOI: 10.1111/j.1365-2478.2010.00866.x.

[70] Hak, B. and Mulder, W. A. “Seismic attenuation imaging with causality”. In: *Geophysical Journal International* 184.1 (2011), pp. 439–451. DOI: 10.1111/j.1365-246X.2010.04848.x.

[71] Hastings, W. K. “Monte Carlo Sampling Methods Using Markov Chains and Their Applications”. In: *Biometrika* 57.1 (1970), pp. 97–109. DOI: 10.2307/2334940.

[72] Huang, G., Nammour, R., and Symes, W. W. “Full-waveform inversion via source-receiver extension”. In: *Geophysics* 82.3 (2017), R153–R171. DOI: 10.1190/geo2016-0301.1.

[73] Huang, G., Nammour, R., and Symes, W. W. “Source-independent extended waveform inversion based on space-time source extension: Frequency-domain implementation”. In: *Geophysics* 83.5 (2018), R449–R461. DOI: 10.1190/geo2017-0333.1.

[74] Huang, G., Nammour, R., and Symes, W. W. “Volume source-based extended waveform inversion”. In: *Geophysics* 83.5 (2018), R369–R387. DOI: 10.1190/geo2017-0330.1.

[75] Jaynes, E. T. “Information Theory and Statistical Mechanics”. In: *Phys. Rev.* 106 (4 1957), pp. 620–630. DOI: 10.1103/PhysRev.106.620.

[76] Jaynes, E. T. “Information Theory and Statistical Mechanics. II”. In: *Phys. Rev.* 108 (2 1957), pp. 171–190. DOI: 10.1103/PhysRev.108.171.

[77] Jaynes, E. T. “Information theory and statistical mechanics (notes by the lecturer)”. In: *Statistical Physics*. Available at https://bayes.wustl.edu/etj/articles/brandeis.pdf. 1963, p. 181.

[78] Jaynes, E. T. “Prior information and ambiguity in inverse problems”. In: *Inverse Problems* 14 (1984). Available at https://bayes.wustl.edu/etj/articles/ambiguity.pdf., pp. 151–166.

[79] Jaynes, E. T. *Probability theory: The logic of science*. Cambridge University Press, 2003. DOI: 10.1017/cbo9780511790423.

[80] Kaipio, J. and Somersalo, E. *Statistical and Computational Inverse Problems*. Applied Mathematical Sciences. Springer New York, 2006. ISBN: 9780387271323. DOI: 10.1007/b138659.

[81] Kamei, R. and Pratt, R. G. “Inversion strategies for visco-acoustic waveform inversion”. In: *Geophysical Journal International* 194.2 (2013), pp. 859–884. DOI: 10.1093/gji/ggt109.

[82] Kazei, V. and Alkhalifah, T. “Waveform inversion for orthorhombic anisotropy with P waves: feasibility and resolution”. In: *Geophysical Journal International* 213.2 (2018), pp. 963–982. DOI: 10.1093/gji/ggy034.

[83] Keilis-Borok, V. I. and Yanovskaja, T. B. “Inverse Problems of Seismology (Structural Review)”. In: *Geophysical Journal of the Royal Astronomical Society* 13 (1967), pp. 223–234. DOI: 10.1111/j.1365-246X.1967.tb02156.x.

[84] Kirkpatrick, S., Gelatt, C. D., and Vecchi, M. P. “Optimization by simulated annealing”. In: *Science* 220.4598 (1983), pp. 671–680. DOI: 10.1126/science.220.4598.671.

[85] Kirsch, A. *An Introduction to the Mathematical Theory of Inverse Problems*. Applied Mathematical Sciences. Springer New York, 1996. ISBN: 9780387945309. DOI: 10.1007/978-1-4419-8474-6.

[86] Köhn, D., Hellwig, O., De Nil, D., and Rabbel, W. “Waveform inversion in triclinic anisotropic media – a resolution study”. In: *Geophysical Journal International* 201.3 (2015), pp. 1642–1656. DOI: 10.1093/gji/ggv097.

[87] Köhn, D. “Time domain 2D elastic full waveform tomography”. Available at https://macau.uni-kiel.de/receive/dissertation_diss_00006786. PhD thesis. 2011.

[88] Komatitsch, D. and Martin, R. “An unsplit convolutional perfectly matched layer improved at grazing incidence for the seismic wave equation”. In: *Geophysics* 72.5 (2007), SM155–SM167. DOI: 10.1190/1.2757586.

[89] Komatitsch, D., Xie, Z., Bozdağ, E., Sales de Andrade, E., Peter, D., Liu, Q., and Tromp, J. “Anelastic sensitivity kernels with parsimonious storage for adjoint tomography and full waveform inversion”. In: *Geophysical Journal International* 206.3 (2016), pp. 1467–1478. DOI: 10.1093/gji/ggw224.

[90] Krebs, J. R., Anderson, J. E., Hinkley, D., Neelamani, R., Lee, S., Baumstein, A., and Lacasse, M.-D. "Fast full-wavefield seismic inversion using encoded sources". In: *Geophysics* 74.6 (2009), WCC177–WCC188. DOI: 10.1190/1.3230502.

[91] Kunert, M., Kurzmann, A., and Bohlen, T. "Application of 2D acoustic full waveform inversion to OBC-data in shallow water". In: *78th EAGE Conference and Exhibition*. 2016. DOI: 10.3997/2214-4609.201601549.

[92] Kunert, M. "Anwendung der 2D akustischen Wellenforminversion auf OBC-Daten". German. MSc. thesis. Karlsruher Institut für Technologie (KIT), 2015. DOI: 10.5445/IR/1000052718.

[93] Kurzmann, A. "Applications of 2D and 3D full waveform tomography in acoustic and viscoacoustic complex media". Available at https://publikationen.bibliothek.kit.edu/1000034421. PhD thesis. Karlsruher Institut für Technologie (KIT), 2012.

[94] Kuvshinov, B. N. and Mulder, W. A. "The exact solution of the time-harmonic wave equation for a linear velocity profile". In: *Geophysical Journal International* 167.2 (2006), pp. 659–662. DOI: 10.1111/j.1365-246X.2006.03194.x.

[95] Lailly, P. "The seismic inverse problem as a sequence of before stack migrations". In: *Conference on inverse scattering: theory and application*. https://books.google.de/books?id=PdEyM7kBkN4C. SIAM Philadelphia, PA. 1983, pp. 206–220.

[96] Lei, T., Bohlen, T., and Zhang, W. "Full-waveform inversion using truncated Newton optimization". In: *SEG Technical Program Expanded Abstracts*. 2018, pp. 1384–1388. DOI: 10.1190/segam2018-2998226.1.

[97] Lellmann, J., Lorenz, D. A., Schönlieb, C., and Valkonen, T. "Imaging with Kantorovich–Rubinstein Discrepancy". In: *SIAM Journal on Imaging Sciences* 7.4 (2014), pp. 2833–2859. DOI: 10.1137/140975528.

[98] Levander, A. R. "Fourth-order finite-difference P-SV seismograms". In: *Geophysics* 53.11 (1988), pp. 1425–1436. DOI: 10.1190/1.1442422.

[99] Li, X., Aravkin, A. Y., van Leeuwen, T., and Herrmann, F. J. "Fast randomized full-waveform inversion with compressive sensing". In: *Geophysics* 77.3 (2012), A13–A17. DOI: 10.1190/geo2011-0410.1.

[100] Lim, S. W. X. "Bayesian inverse problems and seismic inversion". Available at https://ora.ox.ac.uk/objects/uuid:ed60b058-3957-4414-bcb2-db6b5b3c0593. PhD thesis. University of Oxford, 2016.

[101] Lions, J. L. *Contrôle optimal de systemes gouvernés par des équations aux dérivées partielles*. English translation is published in 1971 by Springer: https://www.springer.com/gp/book/9783642650260. Dunod, 1968.

[102] Luo, S. and Sava, P. "A deconvolution-based objective function for wave-equation inversion". In: *SEG Technical Program Expanded Abstracts*. 2011, pp. 2788–2792. DOI: 10.1190/1.3627773.

[103] Malinowski, M., Operto, S., and Ribodetti, A. "High-resolution seismic attenuation imaging from wide-aperture onshore data by visco-acoustic frequency-domain full-waveform inversion". In: *Geophysical Journal International* 186.3 (2011), pp. 1179–1204. DOI: 10.1111/j.1365-246X.2011.05098.x.

[104] Marchuk, G. I. *Numerical methods for nuclear reactor calculations*. https://www.springer.com/us/book/9781489948922. Springer US, 1959. ISBN: 978-1-4899-4892-2.

[105] Marchuk, G. *Adjoint Equations and Analysis of Complex Systems*. Mathematics and Its Applications. Springer Netherlands, 1995. ISBN: 9780792330134. DOI: 10.1007/978-94-017-0621-6.

[106] Marchuk, G., Agoshkov, V., and Shutyaev, V. *Adjoint Equations and Perturbation Algorithms in Nonlinear Problems*. CRC Press, 1996. ISBN: 9781351468794. DOI: 10.1201/9781315136707.

[107] Martin, G. S., Wiley, R., and Marfurt, K. J. "Marmousi2: An elastic upgrade for Marmousi". In: *The Leading Edge* 25.2 (2006), pp. 156–166. DOI: 10.1190/1.2172306.

[108] Menke, W. *Geophysical Data Analysis: Discrete Inverse Theory*. Academic Press, 2018. ISBN: 978-0-12-813555-6. DOI: 10.1016/C2016-0-05203-8.

[109] Métivier, L., Brossier, R., Mérigot, Q., Oudet, E., and Virieux, J. "An optimal transport approach for seismic tomography: application to 3D full waveform inversion". In: *Inverse Problems* 32.11 (2016), p. 115008. DOI: 10.1088/0266-5611/32/11/115008.

[110] Métivier, L., Bretaudeau, F., Brossier, R., Operto, S., and Virieux, J. "Full waveform inversion and the truncated Newton method: quantitative imaging of complex subsurface structures". In: *Geophysical Prospecting* 62.6 (2014), pp. 1353–1375. DOI: 10.1111/1365-2478.12136.

[111] Métivier, L., Brossier, R., Mérigot, Q., Oudet, E., and Virieux, J. "Measuring the misfit between seismograms using an optimal transport distance: application to full waveform inversion". In: *Geophysical Journal International* 205.1 (2016), pp. 345–377. DOI: 10.1093/gji/ggw014.

[112] Métivier, L., Allain, A., Brossier, R., Mérigot, Q., Oudet, E., and Virieux, J. "Optimal transport for mitigating cycle skipping in full-waveform inversion: A graph-space transform approach". In: *Geophysics* 83.5 (2018), R515–R540. DOI: 10.1190/geo2017-0807.1.

[113] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. "Equation of state calculations by fast computing machines". In: *The Journal of Chemical Physics* 21.6 (1953), pp. 1087–1092. DOI: 10.1063/1.1699114.

[114] Miettinen, K. *Nonlinear Multiobjective Optimization*. International Series in Operations Research & Management Science. Springer US, 1999. ISBN: 9780792382782. DOI: 10.1007/978-1-4615-5563-6.

[115] Moghaddam, P. P., Keers, H., Herrmann, F. J., and Mulder, W. A. "A new optimization approach for source-encoding full-waveform inversion". In: *Geophysics* 78.3 (2013), R125–R132. DOI: 10.1190/geo2012-0090.1.

[116] Mora, P. "Elastic wave-field inversion of reflection and transmission data". In: *Geophysics* 53.6 (1988), pp. 750–759. DOI: 10.1190/1.1442510.

[117] Mora, P. "Nonlinear two-dimensional elastic inversion of multioffset seismic data". In: *Geophysics* 52.9 (1987), pp. 1211–1228. DOI: 10.1190/1.1442384.

[118] Mosegaard, K. and Tarantola, A. "Monte Carlo sampling of solutions to inverse problems". In: *Journal of Geophysical Research: Solid Earth* 100.B7 (1995), pp. 12431–12447. DOI: 10.1029/94JB03097.

[119] Mosegaard, K. and Tarantola, A. "Probabilistic Approach to Inverse Problems". In: *International Handbook of Earthquake and Engineering Seismology, Part A*. Ed. by W. H. Lee, H. Kanamori, P. C. Jennings, and C. Kisslinger. Vol. 81. International Geophysics. Academic Press, 2002, pp. 237 –265. DOI: 10.1016/S0074-6142(02)80219-4.

[120] Mulder, W. A. and Hak, B. "An ambiguity in attenuation scattering imaging". In: *Geophysical Journal International* 178.3 (2009), pp. 1614–1624. DOI: 10.1111/j.1365-246X.2009.04253.x.

[121] Neal, R. M. "MCMC using Hamiltonian dynamics". In: *Handbook of Markov Chain Monte Carlo* 2.11 (2011). Available at https://arxiv.org/pdf/1206.1901.pdf., p. 2.

[122] Nocedal, J. and Wright, S. "Numerical optimization". In: *Springer Science* 35.67-68 (1999), p. 7. DOI: 10.1007/978-0-387-40065-5.

[123] Noether, E. "Invariante Variationsprobleme". ger. In: *Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse* 1918 (1918). English translation is available at https://arxiv.org/pdf/physics/0503066.pdf., pp. 235–257.

[124] Operto, S., Gholami, Y., Prieux, V., Ribodetti, A., Brossier, R., Metivier, L., and Virieux, J. "A guided tour of multiparameter full-waveform inversion with multicomponent data: From theory to practice". In: *The Leading Edge* 32.9 (2013), pp. 1040–1054. DOI: 10.1190/tle32091040.1.

[125] Operto, S., Miniussi, A., Brossier, R., Combe, L., Métivier, L., Monteiller, V., Ribodetti, A., and Virieux, J. "Efficient 3-D frequency-domain mono-parameter full-waveform inversion of ocean-bottom cable data: application to Valhall in the visco-acoustic vertical transverse isotropic approximation". In: *Geophysical Journal International* 202.2 (2015), pp. 1362–1391. DOI: 10.1093/gji/ggv226.

[126] Osypov, K., Yang, Y., Fournier, A., Ivanova, N., Bachrach, R., Yarman, C. E., You, Y., Nichols, D., and Woodward, M. "Model-uncertainty quantification in seismic tomography: Method and applications". In: *Geophysical Prospecting* 61.6 (2013), pp. 1114–1134. DOI: 10.1111/1365-2478.12058.

[127] Parker, R. *Geophysical Inverse Theory*. Princeton Series in Geophysics. https://press.princeton.edu/titles/5527.html. Princeton University Press, 1994. ISBN: 9780691036342.

[128] Pekeris, C. L. "Theory of Propagation of Sound in a Half-Space of Variable Sound Velocity under Conditions of Formation of a Shadow Zone". In: *The Journal of the Acoustical Society of America* 18.2 (1946), pp. 295–315. DOI: 10.1121/1.1916366.

[129] Pica, A., Diet, J. P., and Tarantola, A. "Nonlinear inversion of seismic reflection data in a laterally invariant medium". In: *Geophysics* 55.3 (1990), pp. 284–292. DOI: 10.1190/1.1442836.

[130] Plessix, R.-É. "A review of the adjoint-state method for computing the gradient of a functional with geophysical applications". In: *Geophysical Journal International* 167.2 (2006), pp. 495–503. DOI: 10.1111/j.1365-246X.2006.02978.x.

[131] Plessix, R.-É. "Three-dimensional frequency-domain full-waveform inversion with an iterative solver". In: *Geophysics* 74.6 (2009), WCC149–WCC157. DOI: 10.1190/1.3211198.

[132] Plessix, R.-É. and Cao, Q. "A parametrization study for surface seismic full waveform inversion in an acoustic vertical transversely isotropic medium". In: *Geophysical Journal International* 185.1 (2011), pp. 539–556. DOI: 10.1111/j.1365-246X.2011.04957.x.

[133] Plessix, R.-É., Baeten, G., de Maag, J. W., ten Kroode, F., and Rujie, Z. "Full waveform inversion and distance separated simultaneous sweeping: a study with a land seismic data set". In: *Geophysical Prospecting* 60.4 (2012), pp. 733–747. DOI: 10.1111/j.1365-2478.2011.01036.x.

[134] Plessix, R.-É., Milcik, P., Rynja, H., Stopin, A., Matson, K., and Abri, S. "Multiparameter full-waveform inversion: Marine and land examples". In: *The Leading Edge* 32.9 (2013), pp. 1030–1038. DOI: 10.1190/tle32091030.1.

[135] Pratt, R. G. "Inverse Theory Applied to Multi-Source Cross-Hole Tomography. Part 2: Elastic wave-equation method". In: *Geophysical Prospecting* 38.3 (1990), pp. 311–329. DOI: 10.1111/j.1365-2478.1990.tb01847.x.

[136] Pratt, R. G. "Medical ultrasound tomography: lessons from exploration geophysics". In: *Proceedings of the International Workshop on Medical Ultrasound Tomography: 1.-3. Nov. 2017, Speyer, Germany*. KIT Scientific Publishing. 2018, p. 65. DOI: 10.5445/KSP/1000071328.

[137] Pratt, R. G. "Seismic waveform inversion in the frequency domain, Part 1: Theory and verification in a physical scale model". In: *Geophysics* 64.3 (1999), pp. 888–901. DOI: 10.1190/1.1444597.

[138] Pratt, R. G. and Shipp, R. M. "Seismic waveform inversion in the frequency domain, Part 2: Fault delineation in sediments using crosshole data". In: *Geophysics* 64.3 (1999), pp. 902–914. DOI: 10.1190/1.1444598.

[139] Pratt, R. G. and Worthington, M. H. "Inverse Theory Applied to Multi-Source Cross-Hole Tomography. Part 1: Acoustic Wave-Equation Method". In: *Geophysical Prospecting* 38.3 (1990), pp. 287–310. DOI: 10.1111/j.1365-2478.1990.tb01846.x.

[140] Pratt, R. G., Shin, C., and Hick, G. "Gauss–Newton and full Newton methods in frequency–space seismic waveform inversion". In: *Geophysical Journal International* 133.2 (1998), pp. 341–362. DOI: 10.1046/j.1365-246X.1998.00498.x.

[141] Pratt, R. G., Huang, L., Duric, N., and Littrup, P. "Sound-speed and attenuation imaging of breast tissue using waveform tomography of transmission ultrasound data". In: *Proceedings SPIE 6510, Medical Imaging: Physics of Medical Imaging*. 2007. DOI: 10.1117/12.708789.

[142] Pratt, R. G., Song, Z.-M., Williamson, P., and Warner, M. "Two-dimensional velocity models from wide-angle seismic data by wavefield inversion". In: *Geophysical Journal International* 124.2 (1996), pp. 323–340. DOI: 10.1111/j.1365-246X.1996.tb07023.x.

[143] Press, F. "Earth models obtained by Monte Carlo Inversion". In: *Journal of Geophysical Research* 73.16 (1968), pp. 5223–5234. DOI: 10.1029/JB073i016p05223.

[144] Ravaut, C., Operto, S., Improta, L., Virieux, J., Herrero, A., and Dell'Aversana, P. "Multiscale imaging of complex structures from multifold wide-aperture seismic data by frequency-domain full-waveform tomography: application to a thrust belt". In: *Geophysical Journal International* 159.3 (2004), p. 1032. DOI: 10.1111/j.1365-246X.2004.02442.x.

[145] Rawlinson, N., Fichtner, A., Sambridge, M., and Young, M. K. "Chapter One – Seismic Tomography and the Assessment of Uncertainty". In: ed. by R. Dmowska. Vol. 55. Advances in Geophysics. Elsevier, 2014, pp. 1 –76. DOI: 10.1016/bs.agph.2014.08.001.

[146] Sambridge, M. and Mosegaard, K. "Monte Carlo methods in geophysical inverse problems". In: *Reviews of Geophysics* 40.3 (2002), pp. 3–1–3–29. DOI: 10.1029/2000RG000089.

[147] Scales, J. A. and Snieder, R. "The anatomy of inverse problems". In: *Geophysics* 65.6 (2000), pp. 1708–1710. DOI: 10.1190/geo2000-0001.1.

[148] Sears, T., Singh, S., and Barton, P. "Elastic full waveform inversion of multi-component OBC seismic data". In: *Geophysical Prospecting* 56.6 (2008), pp. 843–862. DOI: 10.1111/j.1365-2478.2008.00692.x.

[149] Shalev-Shwartz, S. and Ben-David, S. *Understanding Machine Learning: From Theory to Algorithms*. New York: Cambridge University Press, 2014. ISBN: 1107057132, 9781107057135. DOI: 10.1017/CBO9781107298019.

[150] Shin, C., Jang, S., and Min, D.-J. "Improved amplitude preservation for prestack depth migration by inverse scattering theory". In: *Geophysical Prospecting* 49.5 (2001), pp. 592–606. DOI: 10.1046/j.1365-2478.2001.00279.x.

[151] Shore, J. and Johnson, R. "Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy". In: *IEEE Transactions on Information Theory* 26.1 (1980), pp. 26–37. DOI: 10.1109/TIT.1980.1056144.

[152] Sirgue, L. and Pratt, R. G. "Efficient waveform inversion and imaging: A strategy for selecting temporal frequencies". In: *Geophysics* 69.1 (2004), pp. 231–248. DOI: 10.1190/1.1649391.

[153] Smithyman, B., Pratt, R. G., Hayles, J., and Wittebolle, R. "Detecting near-surface objects with seismic waveform tomography". In: *Geophysics* 74.6 (2009), WCC119–WCC127. DOI: 10.1190/1.3223313.

[154] Stuart, A. M. "Inverse problems: A Bayesian perspective". In: *Acta Numerica* 19 (2010), 451–559. DOI: 10.1017/S0962492910000061.

[155] Sun, W. and Fu, L.-Y. "Two effective approaches to reduce data storage in reverse time migration". In: *Computers & Geosciences* 56 (2013), pp. 69 –75. DOI: 10.1016/j.cageo.2013.03.013.

[156] Symes, W. W. "Algorithmic aspects of extended waveform inversion". In: *77th EAGE Conference and Exhibition - Workshops*. 2015. DOI: 10.3997/2214-4609.201413492.

[157] Symes, W. W. "Extended Waveform Inversion". In: *79th EAGE Conference and Exhibition - Workshops*. 2017. DOI: 10.3997/2214-4609.201701711.

[158] Symes, W. W. "Migration velocity analysis and waveform inversion". In: *Geophysical Prospecting* 56.6 (2008), pp. 765–790. DOI: 10.1111/j.1365-2478.2008.00698.x.

[159] Symes, W. W. "Reverse time migration with optimal checkpointing". In: *Geophysics* 72.5 (2007), SM213–SM221. DOI: 10.1190/1.2742686.

[160] Symes, W. W. "The seismic reflection inverse problem". In: *Inverse Problems* 25.12 (2009), p. 123008. DOI: 10.1088/0266-5611/25/12/123008.

[161] Takougang, E. M. T. and Calvert, A. J. "Application of waveform tomography to marine seismic reflection data from the Queen Charlotte Basin of western Canada". In: *Geophysics* 76.2 (2011), B55–B70. DOI: 10.1190/1.3553478.

[162] Tarantola, A. "A strategy for nonlinear elastic inversion of seismic reflection data". In: *Geophysics* 51.10 (1986), pp. 1893–1903. DOI: 10.1190/1.1442046.

[163] Tarantola, A. *Elements for physics: quantities, qualities, and intrinsic theories*. Springer Science & Business Media, 2006. DOI: 10.1007/978-3-540-31107-2.

[164] Tarantola, A. *Inverse problem theory and methods for model parameter estimation*. Vol. 89. SIAM, 2005. DOI: 10.1137/1.9780898717921.

[165] Tarantola, A. *Inverse problem theory: Methods for data fitting and model parameter estimation*. 1st ed. Vol. 613. Available at http://www.ipgp.fr/~tarantola/Files/Professional/Papers_PDF/InverseProblemTheory.pdf. 1987.

[166] Tarantola, A. "Inversion of seismic reflection data in the acoustic approximation". In: *Geophysics* 49.8 (1984), pp. 1259–1266. DOI: 10.1190/1.1441754.

[167] Tarantola, A. *Mapping of Probabilities: Theory for the Interpretation of Uncertain Physical Measurements (unfinished book)*. Available at http://www.ipgp.fr/~tarantola/Files/Professional/Books/MappingOfProbabilities.pdf. 2007.

[168] Tarantola, A. "Popper, Bayes and the inverse problem". In: *Nature physics* 2.8 (2006), pp. 492–494. DOI: 10.1038/nphys375.

[169] Tarantola, A. "Theoretical background for the inversion of seismic waveforms including elasticity and attenuation". In: *Pure and Applied Geophysics* 128.1-2 (1988), pp. 365–399. DOI: 10.1007/BF01772605.

[170] Tarantola, A. and Valette, B. "Generalized nonlinear inverse problems solved using the least squares criterion". In: *Reviews of Geophysics* 20.2 (1982), pp. 219–232. DOI: 10.1029/RG020i002p00219.

[171] Tarantola, A. and Valette, B. "Inverse problems = Quest for information". In: *Journal of Geophysics* (1982). Available at http://www.ipgp.jussieu.fr/~tarantola/Files/Professional/Papers_PDF/IP_QI_original.pdf., pp. 159–170.

[172] ten Kroode, F., Bergler, S., Corsten, C., de Maag, J. W., Strijbos, F., and Tijhof, H. "Broadband seismic data – The importance of low frequencies". In: *Geophysics* 78.2 (2013), WA3–WA14. DOI: 10.1190/geo2012-0294.1.

[173] Thurin, J, Brossier, R, and Métivier, L. "Ensemble-based uncertainty estimation in full waveform inversion". In: *79th EAGE Conference and Exhibition*. 2017. DOI: 10.3997/2214-4609.201701007.

[174] Thurin, J., Brossier, R., and Métivier, L. "An ensemble-transform Kalman filter: Full-waveform inversion scheme for uncertainty estimation". In: *SEG Technical Program Expanded Abstracts*. 2017, pp. 1307–1313. DOI: 10.1190/segam2017-17733053.1.

[175] Tikhonov, A. N. "On the solution of ill-posed problems and the method of regularization". In: *Doklady Akademii Nauk*. Vol. 151. 3. Available at http://www.mathnet.ru/links/9afd901ed0d325bbe16fed7cb09150f5/dan28329.pdf. Soviet Academy of Sciences. 1963, pp. 501–504.

[176] Tribus, M. and McIrvine, E. C. "Energy and information". In: *Scientific American* 225.3 (1971), pp. 179–190. DOI: 10.1038/scientificamerican0971-179.

[177] Tromp, J., Tape, C., and Liu, Q. "Seismic tomography, adjoint methods, time reversal and banana-doughnut kernels". In: *Geophysical Journal International* 160.1 (2005), pp. 195–216. DOI: 10.1111/j.1365-246X.2004.02453.x.

[178] Ulrych, T. and Sacchi, M. *Information-Based Inversion and Processing with Applications*. Handbook of Geophysical Exploration: Seismic Exploration. Elsevier Science, 2005. ISBN: 9780080461342. DOI: 10.1016/s0950-1401(05)x8001-6.

[179] van Leeuwen, T. and Herrmann, F. J. "Fast waveform inversion without source-encoding". In: *Geophysical Prospecting* 61 (2013), pp. 10–19. DOI: 10.1111/j.1365-2478.2012.01096.x.

[180] van Leeuwen, T. and Herrmann, F. J. "Mitigating local minima in full-waveform inversion by expanding the search space". In: *Geophysical Journal International* 195.1 (2013), pp. 661–667. DOI: 10.1093/gji/ggt258.

[181] van Leeuwen, T. and Mulder, W. A. "A correlation-based misfit criterion for wave-equation traveltime tomography". In: *Geophysical Journal International* 182.3 (2010), pp. 1383–1394. DOI: 10.1111/j.1365-246X.2010.04681.x.

[182] van Leeuwen, T. and Mulder, W. A. "A variable projection method for waveform inversion". In: *71st EAGE Conference and Exhibition incorporating SPE EUROPEC*. 2009. DOI: 10.3997/2214-4609.201400381.

[183] van Leeuwen, T., Aravkin, A. Y., and Herrmann, F. J. "Seismic Waveform Inversion by Stochastic Optimization". In: *International Journal of Geophysics* (2011), pp. 1 –18. DOI: 10.1155/2011/689041.

[184] Vapnik, V. *Estimation of dependences based on empirical data*. Springer Science & Business Media, 2006. DOI: 10.1007/0-387-34239-7.

[185] Vapnik, V. "Principles of risk minimization for learning theory". In: *Advances in Neural Information Processing Systems*. Available at https://pdfs.semanticscholar.org/9642/a175637a400b425f0ac0cb6a2b067cc8fe6b.pdf. 1992, pp. 831–838.

[186] Vapnik, V. *The nature of statistical learning theory*. Springer Science & Business Media, 2013. DOI: 10.1007/978-1-4757-3264-1.

[187] Virieux, J. "P-SV wave propagation in heterogeneous media: Velocity-stress finite-difference method". In: *Geophysics* 51.4 (1986), pp. 889–901. DOI: 10.1190/1.1442147.

[188] Virieux, J. "SH-wave propagation in heterogeneous media: Velocity-stress finite-difference method". In: *Geophysics* 49.11 (1984), pp. 1933–1942. DOI: 10.1190/1.1441605.

[189] Virieux, J. and Operto, S. "An overview of full-waveform inversion in exploration geophysics". In: *Geophysics* 74.6 (2009), WCC1–WCC26. DOI: 10.1190/1.3238367.

[190] Virieux, J., Asnaashari, A., Brossier, R., Métivier, L., Ribodetti, A., and Zhou, W. "An introduction to full waveform inversion". In: *Encyclopedia of Exploration Geophysics*. Society of Exploration Geophysicists, 2017. DOI: 10.1190/1.9781560803027.entry6.

[191] Warner, M. and Guasch, L. "Adaptive waveform inversion: Theory". In: *Geophysics* 81.6 (2016), R429–R445. DOI: 10.1190/geo2015-0387.1.

[192] Warner, M., Ratcliffe, A., Nangoo, T., Morgan, J., Umpleby, A., Shah, N., Vinje, V., Štekl, I., Guasch, L., Win, C., Conroy, G., and Bertrand, A. "Anisotropic 3D full-waveform inversion". In: *Geophysics* 78.2 (2013), R59–R80. DOI: 10.1190/geo2012-0338.1.

[193] Warner, M., Stekl, I., and Umpleby, A. "3D wavefield tomography: synthetic and field data examples". In: *SEG Technical Program Expanded Abstracts*. 2008, pp. 3330–3334. DOI: 10.1190/1.3064036.

[194] Wolpert, D. H. and Macready, W. G. "No free lunch theorems for optimization". In: *IEEE Transactions on Evolutionary Computation* 1.1 (1997), pp. 67–82. DOI: 10.1109/4235.585893.

[195] Wolpert, D. H. and Macready, W. G. *No free lunch theorems for search*. Tech. rep. Available at http://delta.cs.cinvestav.mx/~ccoello/compevol/wolpert96.pdf. Santa Fe Institute, 1995.

[196] Wu, R.-S., Luo, J., and Wu, B. "Seismic envelope inversion and modulation signal model". In: *Geophysics* 79.3 (2014), WA13–WA24. DOI: 10.1190/geo2013-0294.1.

[197] Yang, P., Brossier, R., Métivier, L., Virieux, J., and Zhou, W. "A Time-Domain Preconditioned Truncated Newton Approach to Visco-acoustic Multiparameter Full Waveform Inversion". In: *SIAM Journal on Scientific Computing* 40.4 (2018), B1101–B1130. DOI: 10.1137/17M1126126.

[198] Yang, P., Brossier, R., Métivier, L., and Virieux, J. "A review on the systematic formulation of 3-D multiparameter full waveform inversion in viscoelastic medium". In: *Geophysical Journal International* 207.1 (2016), pp. 129–149. DOI: 10.1093/gji/ggw262.

[199] Yang, P., Brossier, R., Métivier, L., and Virieux, J. "Wavefield reconstruction in attenuating media: A checkpointing-assisted reverse-forward simulation method". In: *Geophysics* 81.6 (2016), R349–R362. DOI: 10.1190/geo2016-0082.1.

[200] Yang, Y. and Engquist, B. "Analysis of optimal transport and related misfit functions in full-waveform inversion". In: *Geophysics* 83.1 (2018), A7–A12. DOI: 10.1190/geo2017-0264.1.

[201] Yang, Y., Engquist, B., Sun, J., and Hamfeldt, B. F. "Application of optimal transport and the quadratic Wasserstein metric to full-waveform inversion". In: *Geophysics* 83.1 (2018), R43–R62. DOI: 10.1190/geo2016-0663.1.

[202] Zhu, H. and Fomel, S. "Building good starting models for full-waveform inversion using adaptive matching filtering misfit". In: *Geophysics* 81.5 (2016), U61–U72. DOI: 10.1190/geo2015-0596.1.