Revisiting Earth's radial seismic structure using a Bayesian neural network approach

Utrecht Studies in Earth Sciences
No. 086

Members of the dissertation committee

Prof. dr. J.-P. Montagner
Université Paris-Diderot – Paris, France
Prof. dr. A. Morelli
Università di Bologna – Bologna, Italy
Prof. dr. J. Ritsema
University of Michigan – Ann Arbor, United States
Prof. dr. M. H. Ritzwoller
University of Colorado – Boulder, United States
Prof. dr. M. Sambridge
Australian National University – Canberra, Australia

The research described in this thesis was financially supported by the Netherlands Organisation for Scientific Research (NWO) and conducted at
The Seismology group
Office O.306 and beyond
Department of Earth Sciences
Faculty of Geosciences
Utrecht University
Budapestlaan 4
3584 CD Utrecht
The Netherlands

ISBN/EAN: 978-90-6266-398-9

Printed in the Netherlands by CPI Koninklijke Wöhrmann, Zutphen.

Cover illustration: artist impression of the Earth's interior structure (copyright Johan Swanepoel). The seismogram is from a magnitude 5.0 earthquake that occurred near Jan Mayen, an island in the Arctic Ocean, on June 15, 1995.

Revisiting Earth's radial seismic structure using a Bayesian neural network approach

De radiale seismische structuur van de Aarde geanalyseerd met behulp van een Bayesiaanse neurale netwerk techniek
(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor
aan de Universiteit Utrecht
op gezag van de rector magnificus, prof.dr. G.J. van der Zwaan,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen
op vrijdag 26 juni 2015 des ochtends te 10.30 uur

door

Ralph Willem Leonard de Wit

geboren op 25 augustus 1986
te Leiderdorp, Nederland

Promotor:
Prof. dr. J. A. Trampert

Copromotor:
Dr. A. P. Valentine

La théorie des probabilités n'est, au fond, que le bon sens réduit au calcul

The theory of probabilities is, at bottom, nothing but common sense reduced to calculus

— PIERRE-SIMON, MARQUIS DE LAPLACE, in *Théorie Analytique des Probabilités (1814)*

Contents

List of Figures

List of Tables

Acknowledgements

So this is it then. After four very enjoyable years at the Seismology group in Utrecht it is time to wrap up. And how better to do that than to thank the people who I've been involved with along the way? Foremost, I thank my promotor and supervisor Jeannot Trampert. Jeannot, when I accepted the position as PhD in this project, your description for the years to come was short but concise: "It's going to be fun". Of course, you were right, as was more often the case; I had a great time. I very much appreciate your guidance, sharp insights and the fact that your office door was always open, ready to answer any questions I had.

Further, I express my gratitude to my copromotor Andrew Valentine. Andrew, thanks for always being there, eager to help and willing to discuss just about anything, be it seismology, neural networks, the plights of academia, British alcoholic beverages or the latest rugby scores. I apologise for my inability to understand the beautiful game of cricket, despite your sincerest attempts to teach me. Perhaps you would like try once more over a pint at some point?

Thinking of my, or better, our office O.306, I would like to thank my office mate Paul Käufl. Paul, I can make this short and simple: you are an awesome guy. Despite your nationality. Oh no, wait, I didn't want to make another lame Dutch German-joke. But then I did. Last time, I promise! Honestly, I cannot put it any different than saying that I had an absolutely brilliant time, sharing an office, many a hotel room at conferences and even more thoughts with you for four years. I am convinced that I would not have gotten as far in my work, would it not have been for our continuous discussions and your willingness to listen to my (seemingly) endless chatter. I have vivid memories of our excursions at various QUEST meetings, 'the street that everything gravitates to' in Porto and a 'Pacific sunset with dolphins'-dinner in San Francisco to top it off. Although from now on we lack an office door to support pictures of us drinking beer, I hope we can have a few together in the future.

Hanneke, thanks for all the help you've given me over the years; I've always felt welcome in your office, ever since I worked on my BSc thesis under your guidance back in 2007. Theo, I am grateful for your IT support and patience whenever I messed things up. Having crashed the old and new pbsserv as an MSc student, I hope that I've not been too much of a nuisance during my PhD, aside from the occasional excess of a few tens of millions of files. Henk, thanks for all the coffee! Nothing beats the

welcoming chime of the coffee bell halfway through the morning. Moving further down the corridor, I very much enjoyed these years in the Utrecht Seismology group, so thank you Arie, Jacqueline, Laura, Suzanne, Denise, Maria, Nienke, Agnieszka, Fatemeh, Elmer, Nesli, Kabir, Arwen, Joop, Andreas, Benoit, Florian, Wouter, Tedi and Sonja.

Bram and Paul, thank you for being my paranymphs and standing by my side during my defence. I thank the organisers of the QUEST programme and all the researchers that were part of it; besides being in nice locations, these meetings were a great opportunity to learn and to meet inspiring scientists from different fields in seismology. During my time as a PhD at Utrecht University I've been involved with several PhD representative networks. The responsibilities and tasks at such organisations were a welcoming change from the daily activities behind the computer. As a bonus, I got to meet loads of amazing people. I thank Tim, Arjon, Jeroen, Negar, Sylvia, Martijn, Jules, Willem, Diana, Akshay, Maaike, Sophie, and all other people at PrOUt for their kindness and their continuing efforts to represent PhDs across our university. The people at UGG, Marlous, Lennart, Joyce, Willem, Wietse, Jelle, Joeri, Helen, Allard and the new generation; Utrecht Geo Graduates have never been so happy, thanks to you! We shared many a laugh and (too) many a beer at our 'evaluatieuitjes', which at some stage were organised more frequently than actual events.

I am glad I was able to continue playing futsal with the Roze Rakkers all those years, albeit in the Senior squad. I appreciate all those matches, goals, trainingskampen, derde helften and Bokma-Fristi, although my head did not always agree with that statement on the following Friday mornings. Thanks to the Mannen van Maandag, Alwin, Tim, Bram, Chris, Pepijn, Vincent, Mau, Ruben, LD and Bas, for making every first day of the week a good one. I am quite convinced that the steaks and beer with you guys on Monday ensures that I'm always ready to take on the other six! Most of all, I thank my parents for supporting me through the years. It might well have been your desire to travel, always taking Maite and me with you around the globe, that triggered my desire to study the Earth. I hope I can at some stage explain what it is I actually did the last four years. Finally, Marily, thank you for always being there for me and correcting my occasional nerdy scientist behaviour when necessary. I am glad we are together and very much look forward to the trip of our lifetime!

Ralph de Wit

Utrecht, May 9, 2015

1

Introduction

Ever since the days of the early explorers, mankind's desire to discover and understand its surroundings has been the driving force underlying many investigations. An expedition launched in 1519 by Ferdinand Magellan would turn out to be the first successful circumnavigation of the Earth[1]. These epic voyages only scratched the surface of our planet, however. Earth's inner structure remained enigmatic for centuries to come, and perhaps ever will be. Even now, we cannot sample the Earth's interior directly; attempts to do so have stranded in the crust. The deepest borehole in the world, the Kola Superdeep Borehole, was drilled between 1970 and 1994 by Russia in an effort to reach as deep into the Earth as possible. As was commonly the case in the Cold War-era, the goal was to beat the United States to it, although the publicity for this competition was overshadowed by the attention given to the contemporary space race. Despite the Russians winning this race towards the Earth's centre, the Kola borehole reaches a depth of a mere 12.2 kilometres below the surface, roughly a third of the Baltic continental crust and less than two per mille of the Earth's radius.

Without a means to look inside the Earth directly, one has to resort to indirect observations to learn about its internal properties. Solving the corresponding imaging or inverse problems is a fundamental task in the physical sciences and applications are vast. For instance, the human body can be analysed using non-invasive medical imaging techniques, such as X-ray Computed Tomography (CT) and Magnetic Resonance Imaging (MRI). Similarly, seismologists can infer the Earth's internal structure

[1] Magellan would never complete the expedition, however. After his death at the hands of Philippine tribesmen, Juan Sebastiá Elcano commanded the only ship (out of five) that would return safely to Spain on 6 September, 1522.

using surface recordings of seismic energy that has travelled through our planet. This thesis is concerned with solving such seismological inference problems using pattern recognition techniques. Searching for patterns is fundamental to extracting information from data. If automated using computer algorithms, the discovery of patterns in data can be classified as machine learning. In this thesis, I will use one particular machine learning approach, artificial neural networks, to solve seismological inverse problems in a Bayesian framework. This allows one to obtain a complete statistical description of earth model parameters, similar to more common sampling-based inversion techniques. The method is flexible and enables us to address specific hypotheses on Earth's structure. This first chapter will provide some more context on the historical development of seismological earth models, followed by the motivation and outline of the thesis.

1.1 Imaging the Earth's interior

The earliest thoughts on the Earth's internal structure relate to the recognition of some sort of underworld, lying beneath the surface of our own realm and often the place where the souls of the departed go. The idea that the Earth below us is a dark place associated with the afterlife, and commonly "hell", is remarkably widespread throughout both cultures and time (Kroonenberg, 2013). For instance, the Greek chtonic, *χθόνιος*, literally subterranean, the Chinese *diyu* and the *duat*, the realm of the dead and of the god Osiris in Egyptian mythology, all relate to similar concepts of an underworld. It was not until the seventeenth century, during the age of enlightenment, that more thought, and perhaps more scientific thought at that, was given to the Earth's internal structure. One of the main considerations concerned the question whether the Earth was solid or hollow, with Halley (1692), known best for the eponymous Halley's comet, actively supporting the latter hypothesis (Ellenberger, 1999). However, a more detailed description could not be given until the advent of seismology.

Since the start of the twentieth century, seismologists have been able to use the illumination of the Earth's interior by seismic waves to infer its structure (Figure 1.1). Oldham (1906) was the first to draw insights from seismic observations and hypothesised the existence of the Earth's core. Other notable early discoveries include the detection of the Moho, the seismic discontinuity between the crust and mantle (Mohorovičić, 1910, 1992), the depth of the core-mantle boundary (CMB, Gutenberg (1914)) and the break-down of the core into an inner and outer component (Lehmann, 1936). A historical overview of seismology can be found in Ben-Menahem (1995); Agnew (2002).

Since these groundbreaking observations, seismological models of the Earth's interior have gradually evolved into more and more complex images. The development of more sophisticated imaging techniques, in combination with the growth of available data and computational facilities, have paved the way for seismic models that depict the Earth's internal structure on length scales down to tens of kilometres. Current 3-D models typically contain thousands to hundreds of thousands of parameters,

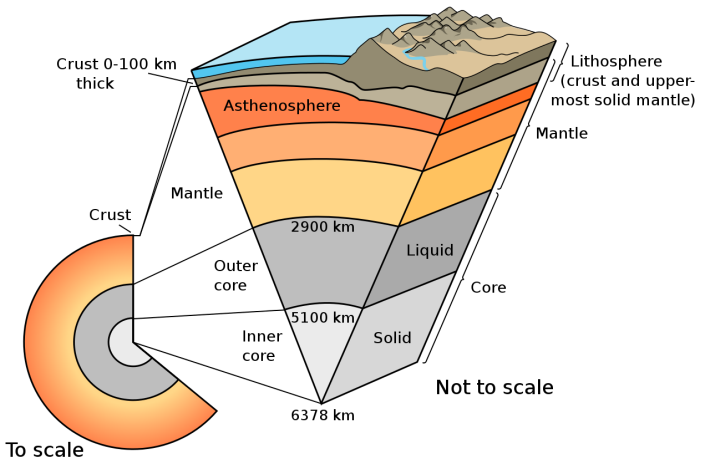


Figure 1.1: A schematic view of the Earth's internal structure. Courtesy of the U.S. Geological Survey.

e.g. Li et al. (2008); Fichtner et al. (2009); Ritsema et al. (2011); Zhu et al. (2012). For an overview of seismic tomography, see for instance Romanowicz (2003); Nolet (2008); Rawlinson et al. (2010). Nonetheless, the gross features of seismic observations can be explained by relatively simple spherically symmetric (1-D) models of wave velocities, density and attenuation, which describe the Earth's average (radial) structure (Figure 1.2). Such radial earth models are routinely used for the determination of seismic source locations and serve as a starting model for 3-D seismic tomography, e.g. Kennett (2006). 1-D seismological reference models are also successfully being used in conjunction with mineral physics data and geodynamical modelling to provide constraints on the Earth's thermochemical structure and its dynamics, e.g. Cammarano et al. (2005, 2011); Cobden et al. (2008, 2009).

Existing seismological reference models have been derived using seismic observables with different, yet complementary, sensitivities to the Earth's interior. The tables of Jeffreys and Bullen (1940) summarised the travel times for many different seismic phases in a 1-D earth model. The accumulation of measurements of the Earth's free oscillations made it possible to construct 1-D profiles of compressional (V_P) and shear (V_S) wave velocities and density (*1066A*, *1066B* (Gilbert, 1975)). Subsequently, parametric models were designed to simultaneously explain travel time, normal mode and regional surface wave dispersion data (*PEM*, Dziewoński et al. (1975)). A similar form of polynomial representation was used by Dziewoński and Anderson (1981) for the Preliminary Reference Earth Model (*PREM*), which was derived from body wave travel times, normal mode frequencies and attenuation measurements, augmented with constraints on the Earth's mass and moment of inertia, and has clearly outlived its 'preliminary' status. The models *iasp91* (Kennett and Engdahl, 1991) and

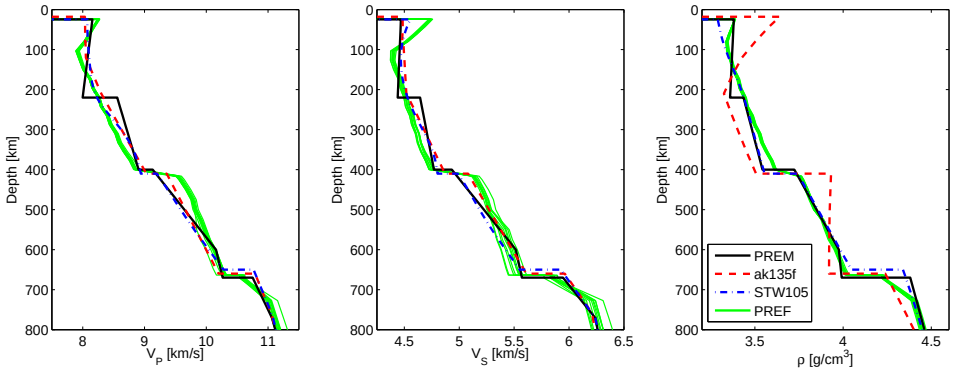


Figure 1.2: Radial (1-D) reference earth models in the upper mantle. Shown are the models *PREM* (Dziewoński and Anderson (1981), black, solid), *ak135f* (Kennett et al. (1995); Montagner and Kennett (1996), red, dashed), *STW105* (blue, dotted-dashed) and *PREF* (Cammarano et al. (2005), 10 out of an ensemble of 99 models, green) for the Voigt average isotropic velocities V_P (left-hand panel) and V_S (middle panel) and the density ρ (right-hand panel).

ak135 (Kennett et al., 1995) were constructed to explain the extensive catalogue of travel times documented by the International Seismological Centre (ISC). More recently, Cammarano et al. (2005) combined seismological and mineral physics data to construct 1-D physical reference models (*PREF*). Kustowski et al. (2008) derived the spherically symmetric model *STW105*, which serves as a basis for a 3-D tomographic mantle model of anisotropic shear wave velocity (*S362ANI*). These models were derived from body wave travel times, long-period waveforms and surface wave phase anomalies. In addition to the isotropic elastic structure, much attention has been given to the Earth's properties in terms of anisotropy, i.e. the direction-dependence of seismic wave propagation (Chang et al., 2014), and attenuation, i.e. the dissipation of seismic energy (Romanowicz and Mitchell, 2007).

1.2 Assessment of seismological models

But what is the quality of such models? The current ability to construct 3-D images of high *nominal* resolution, not uncommonly accompanied by an attractive and adaptive colour scheme, may draw attention away from this fundamental question. Seismic inverse problems are notoriously non-unique; different earth models can explain the data equally well, but may lead to incompatible interpretations of the nature of the Earth's interior and dynamics, e.g. Trampert and van der Hilst (2005); de Wit et al. (2012). Therefore, a quantification of uncertainties in any inferred earth model is essential to assess its quality and the robustness of any subsequent interpretation.

The aforementioned 1-D seismological reference models correlate with each other to a high degree (Figure 1.2), especially in the lower mantle and core, yet there is disagreement. As noted above, 3-D seismic tomographic models are often built upon

such radial earth models. The quality of a 3-D tomographic model is thus intrinsically linked to the details of the 1-D model and it becomes crucial to assess the quality of the latter. Further, seismological models are frequently used to determine earthquake locations. The spherically symmetric *ak135* model (Kennett et al., 1995), for instance, is used in the location algorithm of the International Seismological Centre (ISC). However, any imperfections in the earth model will map into the source location estimate, e.g. Valentine and Woodhouse (2010); Valentine and Trampert (2012a). Clearly, an accurate estimation of seismic source parameters requires a precise knowledge of the underlying earth model and its uncertainties.

However, quantifying model uncertainties presents a challenge in traditional seismological inverse problems; consequently, most existing techniques are pragmatic and based upon linear approximations. Indeed, in many geophysical inverse problems, a single 'optimal' solution is obtained via a linearised approach, e.g. Parker (1994); Tarantola (2005). In reality, the dependence of the data on the model often is non-linear. Various approaches are available to assess model quality. Resolution analyses, for instance using the linear framework by Backus and Gilbert (1968, 1970), can be employed to determine the robustness of the inferred earth models, e.g. Kennett (1998); Masters and Gubbins (2003). In seismic tomography, resolution and covariance matrices can provide some assessment of model quality, e.g. Aki et al. (1977); Boschi (2003); Vasco et al. (2003), but such measures are usually affected by subjective regularisation criteria, i.e. depend on assumptions which may not have a physical justification. The single 'optimal' solution to the inverse problem minimises a cost function that takes into account both the data misfit and the regularisation, which implies a trade-off between the two components. Due to the nature and distribution of data, such regularisation is necessary in most seismological inversions and therefore heavily used. However, the regularisation can have a significant effect on the final solution, e.g. Boschi and Dziewoński (1999); Resovsky and Ritzwoller (1999); Beghein and Trampert (2003); de Wit et al. (2012). Furthermore, it is very difficult or even impossible to assess what part of this final model is required by the data, and what is an artefact of the imposed regularisation. Other examples of uncertainty assessment for the linear case include exploring the model null space, or model non-uniqueness (de Wit et al., 2012), misfit mapping, e.g. in the context of source parameter determination (Valentine and Trampert, 2012a) and resolution tests using matrix probing (An, 2012; Trampert et al., 2013). Kennett et al. (1995) adopted a non-linear search procedure to determine the robustness of *ak135*. However, it would be better to take the non-linearity and model non-uniqueness into account in our inversion framework, rather than treating them *ex post facto*.

A more general approach, which allows us to solve a non-linear inverse problem and to quantify uncertainties, involves the description of our knowledge about earth parameters by probability distributions, e.g. Tarantola and Valette (1982); Tarantola (2005). Following Bayes' theorem (Bayes, 1763), our *posterior* knowledge of the model is given by our *prior* knowledge updated by the observed data, using a physical theory that relates the model to the data. A common approach is to directly sample the posterior model probability density, as is done in Markov Chain Monte Carlo

(MCMC) methods via a (guided) random walk, e.g. Mosegaard and Tarantola (1995); Sambridge and Mosegaard (2002); Tarantola (2005). The earliest studies to employ such techniques to explain seismic observations by an ensemble of 1-D earth models date back to those of Press (1968); Wiggins (1969). While such sampling methods are powerful for solving non-linear inverse problems, they quickly deteriorate as the dimension of the model space increases, a phenomenon which Bellman (1961) termed the *curse of dimensionality*. In practice, this often limits the use of sampling methods to inverse problems which involve relatively few (i.e. a few tens of) model parameters. It is noteworthy that Hamiltonian, or hybrid, Monte Carlo techniques (Duane et al., 1987; Neal, 2011) have recently been successfully applied to a wide range of problems that involve hundreds to thousands of parameters, e.g. Betancourt et al. (2014).

As an alternative to Monte Carlo techniques, I adopt a machine learning approach to draw Bayesian inferences based on samples of the prior model space. More specifically, I use artificial neural networks, which can be viewed as non-linear filters and are very common in pattern recognition problems. They can approximate an arbitrary non-linear mapping between two parameter spaces, inferring the relation from a set of training data, e.g. Bishop (1995); MacKay (2003). As such, neural networks can be very useful in situations where the forward relation is known, but the inverse mapping is unknown or difficult to establish by more conventional analytical or numerical methods. This situation applies to many geophysical inverse problems.

Neural networks have been applied in a wide variety of fields to solve regression and classification, feature extraction and function approximation and optimisation problems in science, finance and medicine. Applications include bankruptcy risk predictions, e.g. Odom and Sharda (2002), breast cancer detection, e.g. Baker et al. (1995), face recognition, e.g. Rowley (1998), landslide susceptibility estimation, e.g. Lee et al. (1998) and traffic flow forecasting, e.g. Jiang and Adeli (2005). van der Baan and Jutten (2000) and Poulton (2001, 2002) provide extensive reviews of geophysical applications of neural networks, which mostly involve feature identification, ranging from seismic phase picks (Dai and MacBeth, 1997) to seamounts (Valentine et al., 2013), data quality control, e.g. Valentine and Woodhouse (2010), and imaging or inversion. For instance, Röth and Tarantola (1994); Langer et al. (1996) inverted seismic waveforms for 1-D earth models. Other recent examples include Meier et al. (2007b); Shahraeeni et al. (2012); Käufl et al. (2014); Walker and Curtis (2014), in which artificial neural networks and a set of prior samples are used to solve various geophysical inverse problems.

1.3 Motivation for the thesis

The main motivation for this thesis is the aforementioned necessity to simultaneously infer Earth structure and quantify the uncertainties in our estimates. The Bayesian framework will enable us to obtain complete statisical information on earth model parameters, represented by posterior probability density functions (pdfs). Furthermore, the final solution is not affected by subjective regularisation criteria that are used in

many seismological inversion schemes. As an alternative to commonly used Monte Carlo methods, I adopt machine learning methods, and more specificaly artificial neural networks. I envision the following advantages of such an approach. First, a trained neural network represents a continuous function that can interpolate between (input) samples. Therefore, a neural network can be applied to multiple unseen input data without the need to re-sample the model space, if the mapping is sufficiently smooth. This may offer significant computational savings. Second, the inferred mapping does not depend on the observed datum, which is used as the reference for data misfit evaluations in Monte Carlo methods. Thus, I can investigate whether or not the seismic data are sensitive to Earth structure, given the assumed data noise level, independent of the measured data. Finally, it is straightforward to invert for any combination of model parameters, again without the need to re-sample the model space, which could offer significant savings of computational budget. The method is flexible, as one is free to choose the output, or *target*, parameter for the neural network. I use one particular type of neural network, a Mixture Density Network (MDN, Bishop (1995)), to obtain posterior pdfs for the earth model parameter(s) of interest. An MDN takes the seismic data as input and outputs the parameters describing a probablity distribution. This allows us to ask specific questions, i.e. test hypotheses, about an arbitrary (combination of) model parameter(s), such as the depth of a seismic discontinuity or the average density in a region. Given the relevance of radial (1-D) earth models, I will apply this flexible method to revisit 1-D earth models and features therein.

In summary, the following concrete objectives of this thesis can be formulated:

1. Provide a complete quantitative framework to solve non-linear Bayesian seismological inverse problems. As an alternative to Monte Carlo methods, which sample the posterior model space, I use artificial neural networks to interpolate between samples from the prior model space.

2. Quantify the information on radial Earth structure that is contained in typical seismological observables, such as body waves, free oscillations and seismograms. This will facilitate a comparison of the relative information contained in various existing seismic data sets.

3. Apply the neural network method to these data sets to assess the robustness of features in existing reference earth models. The resulting posterior pdfs will provide a means to test hypotheses on Earth structure in a robust, quantitative manner. I investigate radial Earth structure in terms of elastic and anelastic structure, anisotropy and depths of major discontinuities.

4. In sampling-based methods, the maximum dimensionality of the model space is a limiting factor. When using neural networks, the dimensionality of the input data is another source of computational burden. I will investigate ways to alleviate such data dimensionality issues, which have become more and more relevant in the current age of big data.

1.4 Outline of the thesis

This thesis is outlined as follows. Chapter 2 describes the general methodology employed in this work. In Chapters 3 to 8, I apply the method to draw inferences on radial Earth structure from various seismic observables. Body-wave travel time data are inverted for the Earth's radial P-wave velocity structure in Chapter 3. Chapters 4 to 7 detail the inversion of normal mode data to test hypotheses on parameters of Earth's radial structure, in terms of velocity, density, attenuation and radial anisotropy. Chapter 8 is concerned with data dimensionality reduction using a novel neural network technique. This method is assessed in light of a seismic waveform inversion for uppermost mantle structure. Finally, the most important results of the thesis are reiterated, along with a perspective on the advantages and limitations of the method and future ideas in Chapter 9.

2

General methodology

This chapter presents the fundaments of inverse theory and machine learning upon which the results in this thesis are built. First, I present a general outline of inverse theory and highlight two approaches to solving inverse problems. Second, I provide a brief description of Bayesian probability theory. Third, the concept of machine learning is introduced, followed by a detailed exposition of artificial neural networks. The neural network method is illustrated by a simple toy problem. Finally, a measure for comparing probability distributions is introduced, which I will use extensively to quantify the information content of various seismic data sets. The aim of this chapter is to provide a backbone for the thesis, in which these techniques are used for probabilistic inversions of seismic data for 1-D Earth structure.

2.1 Inverse problem theory

A fundamental task in the physical sciences is to draw inferences on a physical system from data. The physical system may represent the human body for a medical doctor, the Milky Way for an astrophysicist or the Earth for a geophysicist. Most data are remotely sensed, i.e. are indirect measurements of some internal property of the system. In medical imaging for instance, measurements of the decay in intensity of X-rays provide information on the absorption properties of the human body. These data can be used to non-invasively construct an image of the body for diagnosis and analyse its functioning. In a typical geophysical application, one might be interested in determining the density within the Earth based on gravitational measurements made at its surface.

To make such inferences, Tarantola (2005) defines a very general three-step procedure:

1. The physical system has to be parametrised, i.e. one has to define a minimal set of model parameters whose values completely characterise the system.
2. *Forward* modelling, i.e. the discovery of physical laws that enable us to predict certain observations or measurements, given the values of the model parameters.
3. *Inverse* modelling, i.e. interrogating the observations to infer the values of the model parameters.

For instance, say we are interested in the density distribution in the Earth. As specified above, first we need to define a parametrisation of the system. While the Earth's density distribution is a continuous function of space, in practice the Earth is parametrised on a discrete set of points $\mathbf{m}(\mathbf{x})$, where $\mathbf{x}$ represents the spatial coordinates of a finite number of model parameters $\mathbf{m}$. The elements in this l-dimensional model vector may represent blocks, or voxels, in a 3-D spatial grid, or may represent the weights a_i in a linear combination of a finite number of basis functions ζ_i:

$$\mathbf{m}(\mathbf{x}) = \sum_{i=1}^{l} a_i \zeta_i(\mathbf{x}). \tag{2.1}$$

Commonly, the model parametrisation is defined prior to any forward or inverse modelling. As an alternative, recent work advocates including the specifics of the parametrisation, such as the number of parameters, as a free variable in the inverse problem, an approach that has been labelled transdimensional inversion, e.g. Sisson (2005); Sambridge et al. (2006, 2013).

The second step involves the application of the laws of physics to the model to make predictions for the measurements, as captured in the *forward* relation

$$\mathbf{d} = g(\mathbf{m}), \tag{2.2}$$

where the operation $g(\cdot)$ is an arbitrary, possible non-linear, function which relates the model $\mathbf{m}$ to the data $\mathbf{d}$. Again using a seismological example, the forward model may represent the seismic wave equation. By applying the wave equation to the earth model $\mathbf{m}$ and a seismic source (earthquake), one can make predictions for the data $\mathbf{d}$, the observed ground motion (particle velocity) recorded at the Earth's surface.

The third and final step involves solving the *inverse* or data inference problem. What can we say about the values of our model parameters, given the measurements and our understanding of the physical laws? Solving the inverse problem then boils down to extracting information from the data on the property of interest, a concept that is relevant to all empirical sciences. As Sivia (1996) puts it, solving an inverse problem is nothing more than a "dialogue with the data". Thus, the goal of the inference problem is to make quantitative statements about the features of the physical system under consideration that are consistent with both the measurements and whatever other data-independent information is available (Snieder and Trampert, 1999).

Note that I will use the terms 'inverse' and 'inference' interchangeably throughout this thesis to denote the same type of problem. Strictly speaking, this is incorrect, as the latter does not necessarily involve the use of an explicit inverse operator.

Ideally, there exists an exact physical theory that can be applied to noiseless data and reproduces the model. Following Sabatier (2000), the three requirements of a well-posed inverse problem, as proposed by Hadamard (1902), are defined as the "existence, uniqueness, and stability" of solutions. The latter condition relates to the continuous dependence of the solution on the data. In practice, however, an exact inverse relation is unavailable, the data are noisy and solving the inverse problem is non-trivial for various reasons. First, the parameter that one aims to determine is often a continuous function of spatial coordinates, such as the Earth's density distribution. Phrased differently, the distribution has infinitely many degrees of freedom. By contrast, in any realistic setup the number of measurements that can be made, and thus the number of data available to constrain the model, is finite. This means that we do not have enough data to uniquely determine the model. Equivalently, it means that there are infinitely many different models that will explain the data equally well. To overcome this relative lack of data, the model and forward relation are often discretised (Equation 2.1), resulting in a finite number of model parameters to invert for. However, this only solves part of the problem if the data do not sample the complete model homogeneously. This is the case in for instance seismic tomography, where the goal is to image the Earth's interior based on seismic data recorded at the Earth's surface. The heterogeneous distribution of both sources (earthquakes mainly occur along plate boundaries) and receivers (seismometers are located at the Earth's surface and mostly on continents) results in an uneven sampling of the Earth's interior. As a consequence, part of the earth model parameters are underdetermined, i.e. cannot be uniquely determined due to a lack of information, while for others there is more but possibly contradicting information. Such an inverse problem does not adhere to the above three conditions of well-posedness and is said to be *ill-posed*. The final solution may be very sensitive to small changes in the data.

Second, real data are always noisy:

$$\mathbf{d} = g(\mathbf{m}) + \epsilon, \tag{2.3}$$

in which the noisy data $\mathbf{d}$ consist of the combination of the exact response of the true physical system and the noise ϵ. The data may be corrupted by measurement errors, flaws in the measuring instrument or may be incomplete. These errors in the data can propagate, via the inverse operation, into the estimate of the model and can therefore corrupt the solution. Whenever small errors in the data cause large changes in the final solution, the problem is *ill-conditioned*. Many inverse problems are both ill-posed and ill-conditioned and consequently unstable (Snieder and Trampert, 1999). In the next sections, I consider several tactics that have been developed to solve inverse problems and elaborate on the importance of model quality assessment, which leads up to the introduction of Bayesian probability theory in Section 2.2. For a more detailed coverage of (geophysical) inverse theory, see for instance Parker (1994); Scales et al. (2001); Tarantola (2005).

2.1.1 Solving inverse problems

One can broadly define two classes of approaches to solving inverse problems. The first is a 'deterministic' approach that makes use of an explicit inverse operator, which usually consists of the forward operator and some regularisation procedure to address the ill-posedness and ill-conditioning of the inverse problem. The regularisation is often necessary to stabilise the inverse problem, but implicitly introduces prior information into the system that has no physical basis and can thus bias the solution (Trampert, 1998). For examples of this phenomenon, see for instance Boschi and Dziewoński (1999); Resovsky and Ritzwoller (1999); Beghein and Trampert (2003); de Wit et al. (2012). Furthermore, most non-linear inverse problems cannot be solved directly, since either the inverse operator does not exist or the computational resources are not sufficient to solve the problem. For weakly non-linear problems, the forward problem can be linearised around a reference point in model space; by performing this linearisation anew at each iteration, i.e. around an updated reference model, the problem can be solved in a quasi non-linear fashion, e.g. Woodhouse and Dziewonski (1984); Tarantola (2005). The solution can be described by a mean and covariance, but the quality of the solution depends on the validity of the linear approximation in the vicinity of the reference point. Similar approaches are adopted more and more in seismic tomography by using so-called adjoint techniques, which involve successively and iteratively solving the forward and its adjoint (time-reversed) wavefield, e.g. Tromp et al. (2005, 2008); Fichtner (2010). If this linear approximation is invalid or if the problem is highly non-linear, the Gaussian assumption may be equally invalid and the solution cannot be described by a simple estimate of mean and covariance (Tarantola, 2005).

In such a case, one has to resort to a second class of sampling-based techniques. The most straightforward option is an exhaustive search of the model space, i.e. evaluating the fit between the observed data and synthetic data for all possible models. However, this is computationally infeasible for most applications, as the number of evaluations increases exponentially with the number of model parameters. For instance, if a model consists of 20 parameters that can only take one of two values, one still has to evaluate $2^{20} \approx 10^6$ models. This exponential increase of the size of model space with the number of model dimensions is an important issue in inverse problems and was coined the *curse of dimensionality* by Bellman (1961).

A more efficient approach, which reduces the number of data misfit evaluations, is offered by Monte Carlo search methods, which employ random walks in the model space. The Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970) was the first to do so by successively accepting or rejecting candidate models from the prior model distribution based on the relative fit of the corresponding synthetic data to the observed data. The random walks are guided by the specific algorithm used, which seeks to preferentially sample the regions of model space that are compatible with the data. These random search methods are therefore very useful in tackling non-linear inverse problems, which may not be defined by a global minimum. For such problems, the algorithm needs to be capable of avoiding or escaping from

local misfit minima (or equivalently local likelihood maxima) during the search. Excellent reviews on the use of Monte Carlo methods in (geophysical) inverse problems are Mosegaard and Tarantola (1995); Mosegaard and Sambridge (2002); Sambridge and Mosegaard (2002). Related (global) optimisation methods that involve random sampling include genetic algorithms, which mimic evolutionary processes (Gallagher and Sambridge, 1994; Delsanto et al., 2006) and simulated annealing, which copies the process of annealing in metals (Kirkpatrick et al., 1983; Mosegaard and Vestergaard, 1991).

However, the main limitation of these sampling-based methods is similar to the exhaustive search described above: no method or algorithm escapes the aforementioned curse of dimensionality. The required number of sample evaluations can become large, i.e. $\mathcal{O}(10^6 - 10^7)$ or more, and put high demands on currently available computational facilities, e.g. Sambridge (1999a); Bodin and Sambridge (2009). Therefore, sampling-based methods are limited to the problems of moderate size, i.e. a few tens of dimensions, which is in stark contrast to the thousands of model parameters in for instance seismic tomographic models, e.g. Li et al. (2008); Ritsema et al. (2011).

It is noteworthy that in the case of a well-posed linear inverse problem, most or all techniques will arrive at the same correct solution. For a linear forward problem and a Gaussian prior pdf, the posterior pdf will also be a Gaussian distribution, which can be accurately described by a mean and covariance and can be found by both deterministic and probabilistic methods (Tarantola, 2005).

2.1.2 Assessing solution quality

In light of the above issues of model non-uniqueness and noisy data, it is crucial to determine the quality of any such solution. Scales and Snieder (2000) refer to the latter as the *appraisal* problem, while the determination of the solution is termed the *estimation* problem. Ideally, these two stages go hand in hand, the quality assessment being an integral part of the solution to the inverse problem. To make any inferences quantitative, one must answer three fundamental questions (Scales and Tenorio, 2001). First, how accurately are the data known, or equivalently what is signal and what is noise in the data? Second, how accurate is our forward model, i.e. the predicted response of the physical system for a given model? Third, do we have any other data-independent information on the physical system and model that we need to incorporate? In light of the non-uniqueness of most inverse problems, there may be infinitely many models that fit the data equally well (within their uncertainty), but some may be unreasonable given other independent constraints. For instance, seismic data may be explained equally well by either a smooth density distribution or a configuration in which the density in the mantle is much higher than in the core. By taking into account our knowledge on the Earth's mass and moment of inertia, we would be able to classify the latter model as unreasonable.

Thus, we need a systematic procedure for rejecting such unreasonable models. The inverse problem of inferring the Earth's properties from noisy data is essentially a data analysis problem. As such, the best tools to interrogate the data, and solve the inverse

problem, come from statistics. One approach is to use Bayesian statistics, in which the solution to the inverse problem is given by a probability distribution that reflects the conjunction of our existing knowledge on a model (or system) and the information on that model extracted from the data (Tarantola and Valette, 1982). In fact, the sampling-based techniques, introduced above, commonly employ the Bayesian framework; the solution is sought by directly sampling the posterior model space. If successful, the selected samples are representative of the posterior pdf, e.g. Mosegaard and Tarantola (2002). I will adopt a Bayesian approach in the remainder of this thesis and outline the main concepts in the next section.

2.2 The Bayesian paradigm

From a statistical point of view, inference concerns the problem of inferring the properties of an unknown probability distribution for the model parameters from data generated from that distribution. A unified and logical approach is given by the formulations of probability by Bayes (1763) and Laplace (Laplace, 1812), which were the products of an era in which people started to answer the question of how to reason when it is not possible to argue with certainty. The first to do so was perhaps Bernoulli (1713), who tried to apply the mechanics of deductive logic to the inductive logic required for the inference problems encountered in everyday life (Sivia, 1996). It was in fact Laplace (1812) who defined the current form of Bayes' theorem[1] and successfully applied this form of probability theory to problems in for instance celestial mechanics, jurisprudence and medical statistics. However, despite these early successes, their ideas were initially discredited and largely forgotten until they were rediscovered by Jeffreys (1939). More recently, probability theory has been expanded upon by for instance Jaynes (2003). Bayesian probability theory provides the means to solve the inference problems of interest in this thesis and I therefore give a brief description here. For a complete overview on basic probability theory and advanced topics, see for instance Gelman et al. (1995); Sivia (1996); Jaynes (2003); Tarantola (2005).

In the Bayesian formalism, all information is described by probability distributions that represent *degrees of belief* for each parameter. Bayes' theorem (Bayes, 1763) is given by

$$p(\mathbf{m}|\mathbf{d}) = \frac{p(\mathbf{d}|\mathbf{m})p(\mathbf{m})}{p(\mathbf{d})}, \tag{2.4}$$

where $\mathbf{m}$ and $\mathbf{d}$ are vectors of model and data parameters, respectively, and the notation $p(a|b)$ denotes the conditional probability of a, given b. All elements in Equation 2.4 are represented by probability density functions (pdfs). The prior $p(\mathbf{m})$ and the posterior $p(\mathbf{m}|\mathbf{d})$ pdfs contains all our knowledge on $\mathbf{m}$ before and after observing the data, respectively. $p(\mathbf{d}|\mathbf{m})$ is often replaced by the likelihood function $L(\mathbf{d}|\mathbf{m})$, a quantitative measure that reflects how well a particular model $\mathbf{m}$ explains the ob-

[1] As an aside, I point to a rather interesting exposition of the history of Bayesian inference and, in particular, the use of the adjective 'Bayesian' by Fienberg (2006).

served data $\mathbf{d}$. The likelihood function usually takes the form

$$L(\mathbf{m}) = \exp[-E(\mathbf{m}]), \tag{2.5}$$

where $E(\mathbf{m})$ is an error function that measures the distance between the observed data $\mathbf{d}$ and the predicted data $g(\mathbf{m})$ (Equation 2.2). In this form, mimising the error function is equivalent to maximising the likelihood. Under the Gaussian assumption, the error or data misfit function takes the form (Tarantola, 2005)

$$E(\mathbf{m}) = \frac{1}{2}[\mathbf{d} - g(\mathbf{m})]^t \mathbf{C}_\mathbf{d}^{-1}[\mathbf{d} - g(\mathbf{m})], \tag{2.6}$$

where the data covariance matrix $\mathbf{C_d}$ reflects the prior assumptions on the noise ϵ in the data (Equation 2.3). The discrepancy between observed and synthetic data is effectively weighted by the data covariance matrix and reflects the fact that we are only ever interested in explaining the signal in the observed data, and not its noisy parts (Scales and Snieder, 1998). In this thesis, I will use the common assumption that the data noise is uncorrelated, in which case $\mathbf{C_d} = \sigma_\mathbf{d}^2 \mathrm{I}$ with $\sigma_\mathbf{d}^2$ the data variance, which may be different for each measurement, and the identity matrix I.

The synthetic data are predicted by applying a known forward theory, such as the seismic wave equation, to a given model (Equation 2.2). The final term in Equation 2.4, $p(\mathbf{d})$, is commonly referred to as the *evidence* for the data $\mathbf{d}$, e.g. MacKay (2003), and is given by the integration of the product of the likelihood and the prior distribution over the model space, i.e. all possible models,

$$p(\mathbf{d}) = \int L(\mathbf{d}|\mathbf{m}) p(\mathbf{m}) \mathrm{d}m. \tag{2.7}$$

Some authors argue that the importance of the evidence is underestimated and that in fact the evidence ought to receive most attention, the posterior pdf being a convenient by-product, e.g. Skilling (2006). However, for the purpose of this thesis, it suffices to focus on the numerator of Equation 2.4 The evidence does not depend on the model $\mathbf{m}$, since $\mathbf{m}$ has been integrated out, and I treat the evidence as a normalisation constant k. Then, the solution to the general inverse problem can then be given by the conditional posterior probability distribution (Tarantola, 2005)

$$p(\mathbf{m}|\mathbf{d}) = kL(\mathbf{d}|\mathbf{m}) p(\mathbf{m}). \tag{2.8}$$

Thus, the posterior pdf is proportional to the product of the prior pdf and the likelihood function, which reflects how well a model explains the data. If the data do not offer new information, i.e. do not add knowledge that is not yet contained in the prior, we do not learn anything about the model and the posterior will equal the prior distribution.

The posterior pdf $p(\mathbf{m}|\mathbf{d})$ offers a full description of the posterior model space, in terms of uncertainties in the model and trade-offs between individual model parameters. In addition, one can study the *marginal* probability distribution for a subset of

the model parameters, i.e.

$$\begin{aligned} p(\mathbf{m}'|\mathbf{d}) &= kp(\mathbf{m}')L(\mathbf{m}'|\mathbf{d}) \\ &= \int p(\mathbf{m}|\mathbf{d})\mathrm{d}m_{c+1}\mathrm{d}m_{c+2}...\mathrm{d}m_l, \end{aligned} \tag{2.9}$$

where $\mathbf{m}'$ is a c-dimensional model vector (with c smaller than the number of parameters in the full model $\mathbf{m}$). The marginal posterior pdf represents the final state of knowledge of $\mathbf{m}'$, given the variations in the remaining $l - c$ model parameters. Usually, c is 1 or 2, in which case the marginal probability distribution in Equation 2.9 represents 1-D or 2-D marginal posterior pdfs, respectively. The former reflect our knowledge of a single model parameter, given the data and arbitrary values for all other model parameters. 2-D marginal pdfs are useful to investigate the correlation between any two parameters. It is noteworthy that the most probable model vector $\mathbf{m}'$, i.e. the peak of a marginal posterior pdf $p(\mathbf{m}'|\mathbf{d})$, does not have to coincide with the values of the corresponding model parameters in the most probable model $\mathbf{m}$, i.e. the peak of the full posterior pdf $p(\mathbf{m}|\mathbf{d})$.

In this thesis, I focus on 1-D marginal posterior pdfs, or "marginals". Marginal pdfs can be used to test hypotheses on individual model parameters, i.e. allow us to answer specific questions. Such hypothesis testing relates to the Popperian concept of falsifiability of hypotheses (Popper, 1963). Simply put, our prior pdf represents all possible hypotheses, in this case model parameter values, given our existing knowledge. After observing the data, our knowledge increases and higher probabilities are assigned to some values (hypotheses). By contrast, a low or zero probability is given to values incompatible with the data, which corresponds to falsifying these hypotheses. Obviously, 1-D marginals do not contain any information on higher-dimensional structures in the posterior model space, such as trade-offs between the parameters. This information is available in the full posterior model distribution, but is lost in the marginalisation process (Equation 2.9). However, as was discussed in Section 2.1.1, obtaining the full posterior pdf using sampling-based methods, whether they be Monte Carlo or machine learning techniques, is difficult or sometimes even impossible. I mostly concentrate my efforts on 1-D marginals in this thesis.

2.2.1 Drawbacks of the Bayesian approach

Although intuitive and elegant, the Bayesian or probabilistic approach to solving inverse problems is hampered by several issues. First, a common critism of the Bayesian approach is the necessity to quantitatively describe all components in Equation 2.4 by pdfs, which usually requires some assumptions to be made. For instance, one has to decide on the shape of the prior model distribution $p(\mathbf{m})$. This introduces a certain degree of subjectivity in a method that essentially aims to provide a conjunction of all available independent information (Scales and Snieder, 1997; Scales and Tenorio, 2001). However, proponents of the Bayesian method argue that one always has to make assumptions to solve a problem, e.g. MacKay (2003). Furthermore, the subjectivity can be advantageous because it is explicit. All elements are defined prior to

actually solving the inference problem and the solution follows from a combination of these elements. By contrast, choices made to regularise ill-posed inverse problems are far more implicit and their effect on the final solution is often hard or impossible to assess afterwards.

In relation to the prerequisite that all prior information needs to be expressed by probability distributions, a second caveat is the complexity of doing so. Moulding existing knowledge in the quantitative form of a prior pdf may be harder than it may seem (Scales and Snieder, 1997). Given currently available information, defining a range of possible masses of the Earth is relatively straightforward, but how does one quantify for instance geological information on the types of sediments in a fault zone (Wood and Curtis, 2004)? To address these complexities, one often has to be make practical choices, i.e. choices that do not necessarily have a physical justification, when defining prior distributions.

2.3 Machine Learning

It is our ability to learn from observations that primarily shapes our knowledge of our surroundings. We learn to classify objects (an 'A' from a 'B'), predict events (it may rain soon if dark clouds form) and estimate unknown quantities (the weight of a PhD thesis based on its size). This ability of inductive learning is crucial to our functioning in daily life (Neal, 1994). But what underlying mechanism enables us to discern object 'A' from 'B', or produce a thought 'A' given thought 'B'? At the end of the nineteenth century, the psychologist James (1892) was the first to describe the ability of the brain to memorise in terms of some comprehensive and predictable fundamental structure. A crucial role in the learning process is taken up by the concept of associative memory, or, as James puts it, "This ultimate physiological law of habit among the neural elements is what runs the train." Studying the phenomenon of learning in itself is a challenge. Philosophers, psychologists and neurobiologists seek to describe the fundamental nature and underlying (biological) mechanisms for (inductive) learning. Alternative, one can study artificial intelligence, in which the learning process is translated to a computational environment. Engineers apply this knowledge in real-world devices, while statisticians develop methods to extract information from data.

In previous sections, I highlighted our interest in both forward (Equation 2.2) and inverse relations when solving an inverse or inference problem (Equation 2.4). In the case of a seismological inverse problem, the forward and inverse operators relate 'earth model space' and 'seismic data space'. Stated more generally, we are interested in learning or understanding the relation between two arbitrary spaces. Applying a mapping to some point in input space will yield a response in some output space. In the case of seismic tomography, the physical laws of wave propagation control the forward mapping. To solve either forward or inverse problems, we represent the physical laws by an analytical mathematical model. However, these models can become quite complex or computationally intensive. Moreover, sometimes such a physical theory is not even available. In these cases, one has to resort to statistical

modelling techniques, which include machine learning, to find the relation between the parameter spaces of interest.

In essence, machine learning algorithms learn or approximate relations from empirical data and subsequently make predictions or decisions given the data and the learned relation. They do not purely operate under explicitly programmed instructions, are able to adapt to new circumstances and to detect and extrapolate between patterns. Thus, machine learning has close ties to pattern recognition (Bishop, 1995, 2006) and is widely used in for instance robotics, speech and image recognition. Its niche lies in the fact that the available data always contain patterns, and in particular patterns that are not obvious to recognise analytically. We may not understand the process generating these patterns completely, but we are confident that we can construct an accurate and useful approximation to this process. If so, the approximation can be used to detect patterns or (ir)regularities, make predictions and learn about the process that generated the data (Breiman et al., 2001; Alpaydin, 2004). When the goal is to model the statistical characteristics of the patterns, i.e. to model probability distributions for the data, the approach is commonly referred to as statistical pattern recognition (Jain et al., 2000).

There are various ways to classify machine learning techniques. A first makes a distinction between the style of feedback a 'machine' receives during the learning process, for which in general three types are recognised (Russell and Norvig, 2009)

1. In *unsupervised* learning the algorithm learns patterns in the input even though no explicit feedback is supplied. The most common unsupervised learning task is clustering, i.e. detecting potentially significant clusters of input examples. For example, the algorithm might gradually learn to distinguish high from low quality seismograms, without ever being supplied with labelled examples of both types by a teacher. Note that it only recognises two distinct classes of seismograms; the algorithm does not determine what these two classes represent.

2. In *reinforcement* learning the program learns from a series of reinforcements, which are defined as either rewards or punishments. This type of learning is common in robotics, where a robot for instance needs to find out how to walk by trying several movements. Some movements fail, and the robot will fall down (punishment), while others will be successful, and the robot moves forward (reward). It is up to the robot, or the learning algorithm, to decide which of the actions prior to the reinforcement were most responsible for it.

3. In *supervised* learning the algorithm receives examples of input–output pairs and learns a function that maps from input to output. For instance, the inputs might be images of handwritten addresses and the outputs are the corresponding labels for letters and numbers.

In practice, these distinction are not always so clear-cut. In semi-supervised learning the algorithms is given a few labelled examples and must make what it can of a large collection of unlabelled examples. Furthermore, the labels may be inaccurate.

Thus, both data noise, be it random or systematic, and lack of labels create a continuum between supervised and unsupervised learning. In this thesis, I only consider supervised learning and will not discuss the other types further.

A second characterisation of algorithms considers the intended task for the 'machine', or the type of output of the system. The two main tasks are classification and regression, e.g. Hastie et al. (2009). In *classification*, inputs are divided into multiple classes, and the learner must develop a model that assigns unseen inputs to the correct class. This is usually done through supervised learning, by presenting the learner with labelled samples. Examples of applications include handwriting recognition, e.g. handwritten addresses on postcards, and spam filtering, i.e. learning a distinct classification between "spam" and "not spam". In classification, the output is discrete (the labels), in contrast to the second main task of *regression*, for which the output is continuous. As an example closer to home, we may wish to infer the Earth's continuous density distribution from geophysical data. Other machine learning tasks, such as data clustering and data density estimation, will not be considered in this thesis, except for a brief excursion into data dimensionality reduction in Chapter 8.

There exists a plethora of machine learning techniques, which can all be used to discover patterns in data, make predictions for future input variables and draw inferences from the data. The choice of modelling or learning technique is very much problem- and data-dependent. One factor to consider is the flexibility of the learning algorithm. For instance, there is little use in approximating an arbitrary non-linear mapping by a linear function. Commonly used machine learning techniques include artificial neural networks (ANNs), support vector machines (SVMs), k-Nearest Neighbours (KNN), naive Bayes and Gaussian process modelling. An overview of machine learning algorithms can be found in for instance MacKay (2003); Bishop (2006); Hastie et al. (2009). Given the strong dependence on the particular problem, it is not a goal of this thesis to perform an exhaustive comparison of machine learning methods on the same inference problem. Rather, I choose one particular type, namely artificial neural networks, that seems promising for the seismic inverse problems I wish to solve in this thesis.

2.3.1 The approach in a nutshell

In summary, I will use machine learning techniques to perform regression tasks on noisy data. This will be achieved in a supervised learning framework. Consider again a general non-linear unknown mapping, analogous to the physical forward theory in Equation 2.2,

$$\mathbf{t} = v(\mathbf{x}), \tag{2.10}$$

where the operation $v(\cdot)$ maps an I-dimensional real-valued input $\mathbf{x}$ to a K-dimensional real-valued output, or *target*, $\mathbf{t}$. At our disposal we have a synthetic data set $D = \{\mathbf{x}_n, \mathbf{t}_n\}$, where $n = 1, ..., N$ labels the statistically independent patterns in the data set. Every pattern consists of a pair of input and target vectors. Using the synthetic *training* data set D, we aim to approximate the relation in Equation 2.10 by some para-

metric function, i.e.

$$\tilde{v}(\mathbf{x};\mathbf{w}) \approx v(\mathbf{x}), \tag{2.11}$$

where $\mathbf{w}$ is a real-valued vector consisting of N_w free parameters, which control the approximate function $\tilde{v}(\mathbf{x};\mathbf{w})$. At this stage, it is worth pointing out one of the advantages of the flexible machine learning approach: the input and output in Equation 2.10 are in principle interchangeable. That is, we can infer $\mathbf{t}$ from $\mathbf{x}$ or $\mathbf{x}$ from $\mathbf{t}$. Thus, we are not restricted to approximating the forward relation in Equation 2.2 and may as well approximate the inverse relation $\mathbf{m} = g^{-1}(\mathbf{d})$. This is a key feature in the context of solving inverse problems.

The actual approximation or learning procedure corresponds to finding an optimal set of parameters $\mathbf{w}^*$. In this sense, 'optimal' is reserved for the set of parameters that minimises a distance measure between the predicted output $\mathbf{y}$ and the true (target) output $\mathbf{t}$ for the examples in the data set. Furthermore, the optimal function has to *generalise* well, i.e. should make similarly accurate predictions when applied to unseen input, that is, input patterns that were not contained in the training data set D. Once the optimal vector $\mathbf{w}^*$ has been found, the function $\tilde{v}(\mathbf{x};\mathbf{w}^*)$ is representative of the true unknown mapping in Equation 2.10.

2.4 Artificial Neural Networks

An artificial neural network (ANN) is essentially a mathematical model of an arbitrary mapping between two parameter spaces. They can be viewed as flexible non-linear regression devices or filters and are very common in pattern recognition problems (Bishop, 1995). By modifying the free parameters of the mathematical model during the training process, the mapping can be altered to represent the desired relation. Network training is driven by presenting the network with examples of corresponding input–output pairs. The fundamental idea is to represent the (potentially complicated) mapping as a combination of many simpler univariate *activation functions*. In that sense, neural networks are a form of connectionist model, a term introduced by Hebb (1949) to describe such a computational modelling approach to information processing. It is the non-linear nature of the activation functions that helps neural networks to approximate non-linear relations. Extensive studies of their properties show that their approximation capabilities are very general and therefore ANN are considered to be universal approximators (Cybenko, 1989; Hornik et al., 1989). Furthermore, the number of free parameters in ANNs scales favourably with the size of the input (Bishop, 1995). Applications in this thesis have more than 100 input parameters; for a polynomial of order n, this would result in 100^n free parameters. By contrast, in most neural networks the number of adjustable parameters grows only linearly or quadratically with the dimensionality of the input space. This makes it computationally feasible to apply neural networks to the regression problems of interest here. Finally, one can use neural networks to approximate both the forward and inverse relations; thus, it becomes possible to solve inverse problems directly.

First, I comment on the relation between biological and artifical neural networks, followed by a brief history of the development of the latter. Second, I explain the inner workings of a specific type of neural network, the Multi-Layer Perceptron (MLP), which will be used in the rest of this thesis. To solve Bayesian inverse problems, I extend our the method to so-called Mixture Density Networks (MDNs), which enable us to model conditional probability distributions by a Gaussian Mixture Model (GMM). Finally, I illustrate the functionality of an MDN through a toy problem.

2.4.1 Artificial versus biological neural networks

Artificial neural networks (ANN) derive their name from the biological equivalent on which they are based. They originate from attempts to construct mathematical models to describe information processing in the (human) brain and consequently their vocabulary draws heavily on that of cognitive psychology and neurophysiology. For instance, neural networks learn as a result of their training, instead of being programmed (Poulton, 2001). Although it is still debated to what degree ANNs are a good model for brain architecture, some fundamental concepts are shared by both organic and artificial neural networks. First, both are complex networks built up from many connected fundamental processing units, or neurons. Second, they can acquire knowledge and learn relations from their environment (data). Third, relations are learned and information is stored in the network by modifying the *synaptic* strenghts of the connections between the neurons. Biological neurons generate electrical signals, which can *activate* or excite connected neurons. A similar process occurs in ANNs, in which a change in the input signal to a neuron, represented by some activation function, can modify its output. Much of the development in artificial neural networks is still driven by the desire to better emulate biological neural networks (Müller et al., 1995). Alternatively, when neural networks are used as a computational tool to analyse data, the emphasis lies on ways to increase the speed and efficiency of existing algorithms. In this thesis, I will not concern myself with the biological plausibility of artificial neural networks, and instead use them as an effective tool for statistical pattern recognition, i.e. the statistical modelling of probability distributions based on samples.

2.4.2 A brief history

The following is a synopsis of the historical development of artificial neural networks. For a more extensive historical overview, see for instance Poulton (2001); Dawson (2013). The first efforts in neural computing can be traced back to McCulloch and Pitts (1943), who introduced a mathematical model to describe the behaviour of the brain. The associated McCulloch-Pitts neuron is a simple threshold unit, which was meant to mirror a single neuron in a biological nervous system

$$y = g(\mathbf{w}^T\mathbf{x} + b_0), \tag{2.12}$$

where the activation or threshold function $g(\cdot)$ was of the form of the Heaviside function

$$g(a) = \begin{cases} 0, & \text{if } a < 0 \\ 1, & \text{otherwise.} \end{cases} \tag{2.13}$$

The input $\mathbf{x}$ is given by the level of *activity* of other connected neurons, while the weights $\mathbf{w}$ represent the synaptic strengths of the connections between the neurons, i.e. the equivalent of the synapses in biological neural networks. The bias b_0 determines the threshold for a neuron to become active, or 'fire'.

The neuroscientist Hebb (1949) made the next vital contribution, in which he described the basic mechanism operating between neurons during learning. The concept, now known as "Hebbian learning", has been fundamental for the development of computational intelligence and is used as the basic structure for the weighted connections between the neurons in a network. As such, Hebbian learning shows parallels with the concept of associative memory and the law of habit proposed earlier by the psychologist James (1892). Hebb's theory triggered early computer experiments on memory systems (Rochester et al., 1956), which marked a major milestone as it illustrated the new possibility to test theories using computer simulations.

The next breakthrough came with the "Perceptron" (Rosenblatt, 1958), a two-layer network capable of learning how to classify certain patterns. The name derives from the creator's interest in the visual system and problems of perception. However, the Perceptron had some serious computational limitations. Foremost, it could only classify linearly separable classes and thus was incapable of solving the classic XOR (exclusive-or) problem. Widrow and Hoff Jr. (1960) developed the Adaptive Linear Network (ADALINE), which was similar to the perceptron but instead of the Heaviside function had a linear activation function that facilitated a more powerful learning algorithm. However, the ADALINE suffered from the same fundamental limitation that it could only solve linearly separable problems.

This was shown by Minsky and Papert (1969), who concluded that in practice these algorithms would fail to solve many interesting problems. Furthermore, computers at the time were not powerful enough to process the training of large neural networks. In combination with the criticisms of Minsky and Papert (1969), this resulted in a severe drop in both interest and funding for neural network research. Bishop (1995) argues that the criticisms of Minsky and Papert (1969) were more subtle than commonly portraited. In fact, a perceptron with multiple layers, in which only one of the layers has adaptable parameters, is capable of solving linearly inseparable problems, provided that the processing units (activation functions) in the fixed layer are chosen appropriately. He notes that the real issue with the perceptrons proposed by Rosenblatt (1958) was that these processing units were fixed *a priori* and could therefore not be adapted to the particular problem or data set of interest.

Later, researchers found that more complex non-linear relations could be learned if an intermediate layer of processing elements, so-called *hidden* units, was added. This hidden layer would process the input signal prior to sending it to the output layer. While devising learning rules for such neural networks, researchers encoun-

tered problems arising from the use of the Heaviside step function, which is non-linear but also discontinuous and therefore not differentiable. The solution lay in using a continuous approximation to the Heaviside function. A major breakthrough was made by Rumelhart et al. (1986), who published the non-linear training algorithm called *back-propagation*, although in reality the first to describe such back-propagation of errors was Werbos (1974). Back-propagation overcame many limitations of earlier network architectures, such as the Perceptron and ADALINE, and marked the revival of neural network research.

More recent developments include deep neural networks, which consist of multiple hidden layers and attempt to approximate more complex relationships (see for instance Hinton and Salakhutdinov (2006); Bengio (2009) and Chapter 8 of this thesis). Another class of algorithms receiving much interest are recurrent neural networks, which operate in time, similar to a recursive filter, and in which the neurons can send feedback to each other, similar to the processes occurring in the brain (Schmidhuber, 2014). As a result, these networks, such as the Long Short Term Memory (LSTM) network (Hochreiter and Schmidhuber, 1997), have some internal memory and can be a powerful tool when applied to sequential data.

2.4.3 The Multi-layer Perceptron (MLP)

One can use simple methods, such as polynomials, to approximate multi-variate non-linear mappings. However, these techniques rely on linear combinations of *fixed* basis functions and consequently the required number of such basis functions scales unfavourably with the size of the input (Bishop, 1996). In most practical applications, there will be significant correlations between the input variables so that the *intrinsic* dimensionality is significantly less than the number of inputs. The key to constructing a mapping that can exploit these correlations is to allow the basis functions themselves to be adapted to the data as part of the training process. In this case the number of such functions only needs to grow as the complexity of the problem itself grows, and not simply as the number of input variables grows (Bishop, 1995). Multi-layer neural networks with adjustable basis, or activation, functions are well-suited to take advantage of these correlations and are thus efficient in handling multi-dimensional problems.

Figure 2.1 shows a two-layer *feed-forward* Multi-layer Perceptron (MLP), which is probably the most commonly used neural network (Rumelhart et al., 1986; Bishop, 1995). I only consider MLPs in this thesis. This particular type of neural network consists of two layers of free parameters (*weights*), which are represented by lines in the figure. The weight $w_{ij}^{(1)}$ in the first layer connects the input unit x_i with the *hidden* neuron h_j, while the second layer weight $w_{jk}^{(2)}$ connects the hidden unit h_j to the output neuron z_k. In addition, the *biases* of the first ($b_j^{(1)}$) and second layer ($b_k^{(2)}$) provide a constant offset as input to the neurons in a subsequent layer. For the activation functions I use here, the bias controls the threshold at which the output of a neuron changes sign (Figure 2.2). Information flows only in the forward direction from the

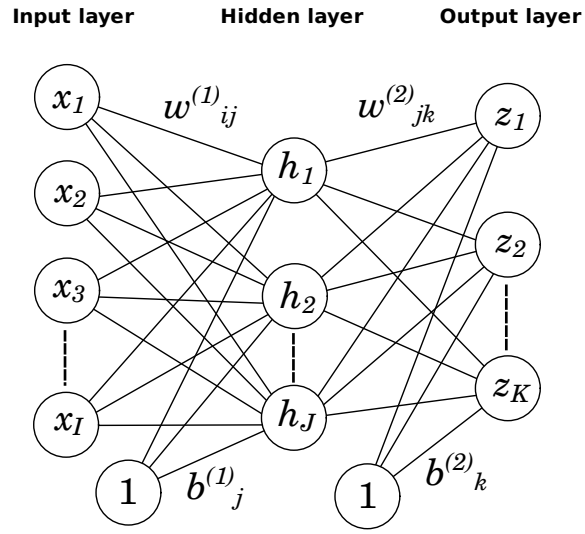


Figure 2.1: A two-layer feed-forward Multi-layer Perceptron (MLP). The lines represent the two layers of free parameters in the network, represented by the *weights* $w_{ij}^{(1)}$ and $w_{jk}^{(2)}$ and *biases* $b_j^{(1)}$ and $b_k^{(2)}$. The I input neurons x_i feed into the J hidden neurons h_j, which form the input to the K output units z_k. An additional input of value 1 feeds into the hidden and output layer, which is associated with the biases. Information flows only from the input to the output neurons (*feed-forward*).

input to the output neurons (*feed-forward*). The network output $\mathbf{z}(\mathbf{x};\mathbf{w})$ is an explicit function of both the input $\mathbf{x}$ and network parameters $\mathbf{w}$. As such, the MLP approximates the continuous input–output mapping or interest (Equation 2.11).

The K units in the MLP output layer are given by

$$z_k = g\left(\sum_j^J w_{jk}^{(2)} h_j + b_k^{(2)}\right), \tag{2.14}$$

where $g(\cdot)$ represents the activation function for the output neurons, $w_{jk}^{(2)}$ and $b_k^{(2)}$ are the second layer weights and biases, respectively, and h_j are the outputs of the J hidden neurons

$$h_j = f\left(\sum_i^I w_{ij}^{(1)} x_i + b_j^{(1)}\right). \tag{2.15}$$

Here, $f(\cdot)$ is the activation function for the hidden units, $w_{ij}^{(1)}$ and $b_j^{(1)}$ are the first layer weights and biases, respectively, and x_i represents the values of the I input units. Commonly used activation functions for the hidden layer are the logistic and hyperbolic tangent functions. I use the latter in this work, because symmetric sigmoids,

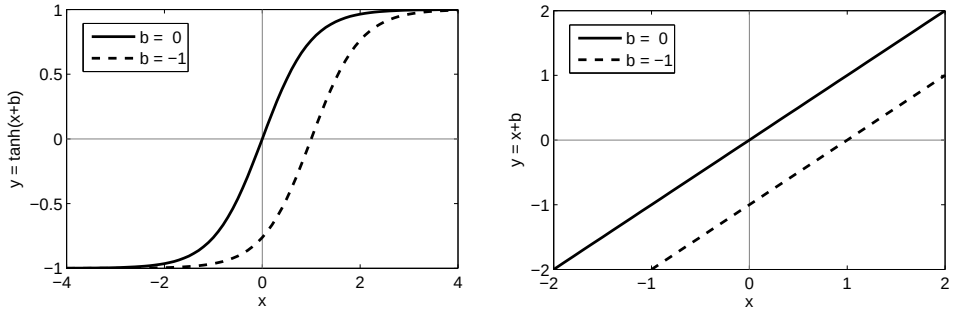


Figure 2.2: In this work, hyperbolic tangents (left-hand panel) are used as activation functions for the hidden neurons (Equation 2.15), while for the output neurons (Equation 2.14) linear activation functions are used (right-hand panel). The bias b provides a constant offset as input to the activation function, i.e. $y = f(x + b)$, and controls the threshold at which the output of a neuron y changes sign.

such as the hyperbolic tangent, often display better convergence properties during network training, e.g. LeCun et al. (1998). Thus, for the hidden neurons, I choose $f(a) = \tanh(a)$, while for the output units I use a linear activation function, $g(a) = a$ (Figure 2.2). These are common choices, and such an MLP can learn an arbitrary continuous mapping from a finite data set, provided the number of hidden units is sufficient (Cybenko, 1989; Hornik et al., 1989). Usually, a trial-and-error procedure is adopted to determine the appropriate number of hidden neurons.

Learning corresponds to the minimisation of a cost function with respect to the network weights. The cost function measures the difference between the network output and the desired output, the target vector. The necessary derivatives are given by the back-propagation algorithm, as introduced by Werbos (1974); Rumelhart et al. (1986). The network is trained on a synthetic data set $D = \{\mathbf{x}_n, \mathbf{t}_n\}$, where $n = 1, ..., N$ labels the statistically independent patterns in the data set. Every pattern consists of a pair of input and target vectors $\mathbf{x}$ and $\mathbf{t}$, respectively. Once successfully trained, the network can be applied to unseen input to produce an estimate of the unknown output.

2.4.4 The Mixture Density Network (MDN)

Bishop (1995) shows that an MLP, as shown in Figure 2.1, outputs the mean of the conditional probability distribution $p(\mathbf{t}|\mathbf{x})$ of the target $\mathbf{t}$, conditioned on the input $\mathbf{x}$. This will give meaningless results if the underlying function, which relates input and target, is multi-valued; therefore, it is desirable to obtain the full conditional distribution of the target, e.g. Bishop (1995); Meier et al. (2007b). I thus employ an MDN, an extension to the MLP which can model an arbitrary probability distribution, in the same fashion that an MLP can approximate an arbitrary function (McLachlan and Basford, 1988).

In this thesis, the network input $\mathbf{x}$ corresponds to seismic data $\mathbf{d}$ and the target $\mathbf{t}$ is

given by the earth model parameters of interest $\mathbf{m}'$, a subspace of the complete radial earth model $\mathbf{m}$ (Equation 2.9). The precise composition of $\mathbf{m}$, $\mathbf{m}'$ and $\mathbf{d}$ will be given in each chapter. An MDN gives a continuous approximation to the corresponding marginal posterior pdf $p(\mathbf{m}'|\mathbf{d})$ (Equation 2.9) as a linear sum of Gaussian kernels:

$$p(\mathbf{m}'|\mathbf{d};\mathbf{w}) \approx \sum_{j=1}^{M} \alpha_j(\mathbf{d};\mathbf{w})\phi_j(\mathbf{m}'|\mathbf{d};\mathbf{w}), \tag{2.16}$$

where α_j gives the relative importance of the jth kernel and the M Gaussian kernels ϕ_j are defined as

$$\phi_j(\mathbf{m}'|\mathbf{d};\mathbf{w}) = \frac{1}{(2\pi)^{c/2}[\sigma_j(\mathbf{d};\mathbf{w})]^c} \exp\left\{-\frac{\|\mathbf{m}' - \boldsymbol{\mu}_j(\mathbf{d};\mathbf{w})\|^2}{2[\sigma_j(\mathbf{d};\mathbf{w})]^2}\right\}, \tag{2.17}$$

where c is the dimensionality of the target vector $\mathbf{m}'$. Note the explicit dependence of the network output $p(\mathbf{m}'|\mathbf{d},\mathbf{w})$ on the network weights $\mathbf{w}$. The Gaussian Mixture Model is fully described by the mean vectors $\boldsymbol{\mu}_j(\mathbf{d};\mathbf{w})$ of size c, the variances $\sigma_j^2(\mathbf{d};\mathbf{w})$ and the mixing coefficients $\alpha_j(\mathbf{d};\mathbf{w})$. Each c-dimensional *spherical* Gaussian kernel ϕ_j is described by a single variance $\sigma_j^2(\mathbf{d};\mathbf{w})$, regardless of the dimensionality c of the target vector $\mathbf{m}'$. Alternatively, more complex Gaussian kernels, such as Gaussians with full covariance matrices (Williams, 1996), could be used. This is computationally more demanding, however, and I find that spherical kernels are flexible enough to model the probability densities of interest in this work.

Figure 2.3 illustrates an MDN, as introduced by Bishop (1995). The parameters describing the mixture model are related to the output $\mathbf{z}(\mathbf{d};\mathbf{w})$ of a conventional MLP, as shown in Figure 2.1. For M spherical Gaussian kernels, the MLP will have $(c+2)\cdot M$ output parameters $z_j^{[\alpha]}, z_j^{[\sigma]}$ and $z_j^{[\mu]}$. To make these raw parameters usable for describing a Gaussian Mixture Model, a few additional transformations have to be performed. The mixing coefficients α_j are required to sum to one, which can be achieved by applying the so-called softmax function to the network output:

$$\alpha_j = \frac{\exp\left(z_j^{[\alpha]}\right)}{\sum_{j=1}^{M}\exp\left(z_j^{[\alpha]}\right)}. \tag{2.18}$$

By definition, the standard deviations σ_j have to be positive, which can be ensured via the transformation

$$\sigma_j = \exp\left(z_j^{[\sigma]}\right). \tag{2.19}$$

The MLP output for the means $\boldsymbol{\mu}_j$ is directly usable, i.e.

$$\boldsymbol{\mu}_j = z_j^{[\mu]}. \tag{2.20}$$

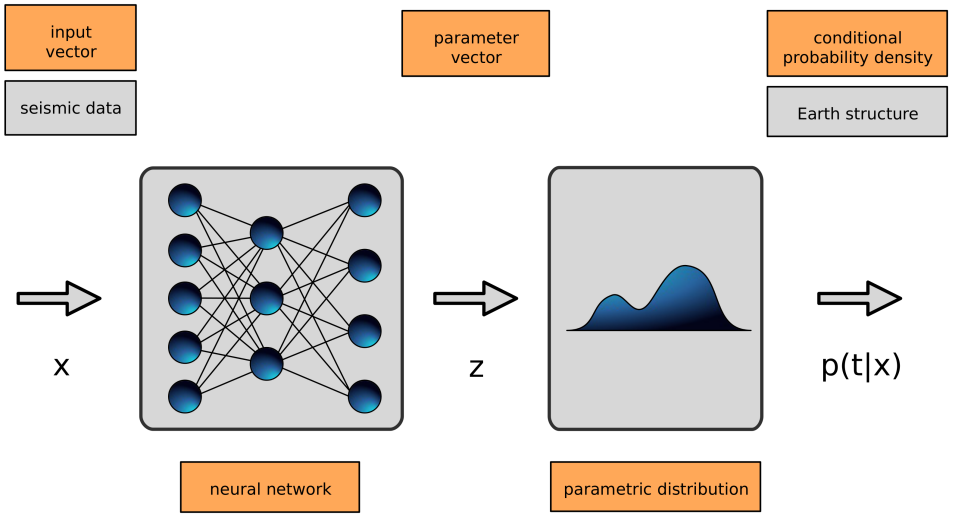


Figure 2.3: A schematic representation of a Mixture Density Network (MDN), as introduced by Bishop (1995). The output of an MDN approximates a parametric distribution $p(\mathbf{t}|\mathbf{x})$ for the target $\mathbf{t}$, conditioned on the input $\mathbf{x}$. The parameters describing this distribution are given by the output $\mathbf{z}$ of a neural network, such as the MLP shown in Figure 2.1. In this work, the input consists of seismic data $\mathbf{d}$, while the target represents the parameter(s) of interest $\mathbf{m}'$, which are a subset of a radial earth model $\mathbf{m}$ (Equations 2.8 and 2.9).

Once the parametric form of the probability distribution has been defined (Equations 2.16 and 2.17), the associated parameters can be found by training an MLP. Training corresponds to finding the weight values that maximise the likelihood for the desired pdf $p(\mathbf{m}'|\mathbf{d};\mathbf{w})$. Since maximising the likelihood is equivalent to minimising the negative logarithm of the likelihood, the error function for the MDN is defined as

$$E = -\sum_{n=1}^{N} \ln \left\{ \sum_{j=1}^{M} \alpha_j(\mathbf{d}_n;\mathbf{w}) \phi_j(\mathbf{m}'_n|\mathbf{d}_n;\mathbf{w}) \right\}, \tag{2.21}$$

where the outer summation runs over the N patterns in the synthetic data set $D = \{\mathbf{d}_n, \mathbf{m}'_n\}$. Analytical expressions for the derivatives of E with respect to the adjustable network parameters are straightforward to derive (Bishop, 1995) and allow an optimisation algorithm to be implemented.

I emphasise that an MDN models the posterior pdf in Equation 2.9 directly. Thus, the likelihood function $L(\mathbf{m}'|\mathbf{d})$ is not explicitly evaluated in the neural network, nor is the prior distribution $p(\mathbf{m}')$. The posterior distribution $p(\mathbf{m}'|\mathbf{d})$ is evaluated implicitly through network training.

2.4.5 Network training

MDN training corresponds to the minimisation of the cost function in Equation 2.21. Commonly, gradient-based optimisation algorithms are used for this task. I use the Scaled Conjugate Gradient (SCG) algorithm (Møller, 1993), which avoids the expensive line-search procedure of the basic conjugate gradient algorithm. Conjugate gradient methods acquire second order information about the error surface and are therefore more efficient than simpler gradient descent methods.

Gradient-based optimisation methods typically operate iteratively and thus require a user-defined starting point for the network weights. The starting point is crucial to ensure that the network training converges to an appropriate solution. For the hyperbolic tangent function, the summed input should be of order unity. If not, the activation functions become saturated, i.e. their first derivative tends to zero. Consequently, the error surface will become almost flat, so that training ceases to be useful. To aid the initialisation of the training process, it is common practice to pre-process the input and target vectors (Section 2.4.7).

The initial network weights are drawn from a Gaussian distribution of zero mean. The variance of this distribution is inversely proportional to the number of input units I for the first layer weights $w_{ij}^{(1)}$ and the number of hidden units J for the second layer weights $w_{jk}^{(2)}$ (Bishop, 1995). Further, the network parameters are initialised such that $p(\mathbf{m}'|\mathbf{d},\mathbf{w}) \approx p(\mathbf{m}')$, i.e. the initial posterior pdf resembles the prior pdf. This requires setting some initial values for the biases of the output layer in the MLP ($b_k^{(2)}$ in Figure 2.1). For the kernel means $\boldsymbol{\mu}_j(\mathbf{d};\mathbf{w})$, the initial values are found by applying the k-means algorithm (MacQueen, 1967), a simple iterative clustering technique, to the target data in the training set (a form of unsupervised learning, in fact). The initial variances $\sigma_j^2(\mathbf{d};\mathbf{w})$ are determined by the variance of the samples in each cluster, i.e. all samples that are closest to the centre (mean) of the j-th kernel. The initial mixing coefficients $\alpha_j(\mathbf{d};\mathbf{w})$ are set equal to the proportion of samples belonging to each cluster. Following Bishop (1995); Nabney (2002), such an initialisation leads to faster convergence and reduces the risk of the optimisation method getting stuck in a poor local minimum.

Regardless of the initialisation, every network run will be sensitive to the specific initial network weights $\mathbf{w}$. It is therefore common practice to train multiple networks with different random weight initialisations, all other settings being equal. The optimal weight vector $\mathbf{w}^*$, which minimises the cost function (Equation 2.21), is used to estimate the marginal posterior pdf $p(\mathbf{m}'|\mathbf{d},\mathbf{w}^*)$ through Equation 2.16.

2.4.6 Generalisation and regularisation

The goal of network training is to approximate the relationship between two parameter spaces. Such a mapping is in general believed to be found when the optimal network produces accurate results for previously unseen data, i.e. data that was not used for network training. In that case, the network is said to display a good *gener-*

alisation behaviour. This can be tested by using a data set that is independent of the training data.

The concept of generalisation is an important issue in practical applications of neural networks. A relatively simple network, i.e. a network with few free parameters, may produce a smooth mapping but will struggle to approximate non-linear functions. Consequently, its predictions will have a large bias, i.e. a large systematic error (Figure 2.4, left-hand panel). Conversely, a overly complex network may be too flexible and predict the training data exactly, but the approximated function will exhibit a large variance, referred to as *overfitting*. Such a network will generalise poorly and make inaccurate predictions for unseen input data (Figure 2.4, middle panel). There is a natural trade-off between the bias and variance and the problem is commonly referred to as the bias-variance dilemma (Geman et al., 1992). Either a large bias or a large variance will result in a large network prediction error (Figure 2.4, right-hand panel). The key is to achieve a compromise between these two effects during network training by controlling, or regularising, the effective complexity of the network (Bishop, 1995).

At this stage, it is worth pointing out that there is little use in fitting training data exactly when the goal is to draw inferences from real noisy data. Ideally, the network is flexible enough to map the signal in the input data to the target parameters, but ignores noise-level fluctuations. Intuitively, modelling the data noise seems to be a good starting point to enhance network generalisation and find the a proper balance between the bias- and variance-related error components. Therefore, I always add noise to the synthetic data to simulate the presence of measurement uncertainties in the real data. This discourages the network from fitting the details of the training data set, i.e. prevents overfitting. Instead, it enhances the generalisation behaviour by encouraging the network to map the underlying relation between input and output. Bishop (1995) shows that such noise addition is similar to using a regularisation constraint, thereby forcing the network to find a "smooth" mapping, i.e. a mapping that is insensitive to variations in the data on the order of the noise level. Meier et al. (2007b) demonstrated this concept in the context of a Bayesian inversion of surface wave data. The effect of the noise addition is similar to the inclusion of the assumed data noise in the likelihood function (Equations 2.5 and 2.6). In the above, I pointed out that subjective regularisation can bias the solutions to inverse problems and one might think that the regularisation due to noise addition, as I use here, causes similar problems. However, regularisation is natural if it is based on realistic prior knowledge. Data noise can be estimated *a priori*; by using it to regularise the training procedure, we can take into account that we are not interested in explaining the noisy data exactly.

In addition, I employ *early stopping*, which is a common procedure to improve generalisation. The network is trained using the conventional training set, but training is halted when the error (Equation 2.21) for an independent validation set starts to increase (Figure 2.4, right-hand panel). Since the validation set is used to determine the optimal set of network weights, a third (test) set is used to verify the accuracy of the network on unseen data.

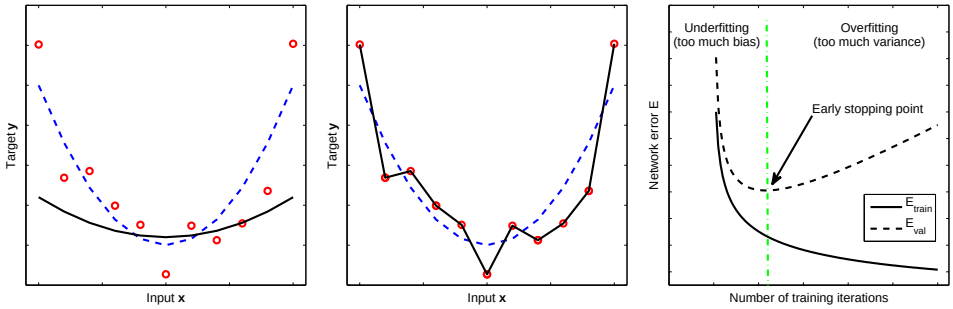


Figure 2.4: Illustration of the trade-off between bias and variance. An inflexible neural network (left-hand panel) approximates a smooth mapping (black, solid line) that gives a systematic error (bias) when predicting the noisy data (red circles) that were generated from the true function (blue, dashed line). Conversely, an overly complex network (middle panel) fits the noisy data exactly, but will generalise poorly to unseen input data and has a high variance. The total network prediction error for a given data set consists of the two components for bias and variance (right-hand panel). The key is to compromise between these two components during network training by controlling, or regularising, the effective complexity of the network. In this thesis this is achieved by adding noise to the synthetic data and *early stopping*, i.e. halting training when the error E_{val} for an independent validation set reaches a minimum (green, dashed-dotted line).

2.4.7 Data pre-processing

To facilitate the stability of the training process and enhance its convergence, it is common practice to pre-process the input and target vectors prior to network training (Bishop, 1995). I process the input $\mathbf{x}$ to have zero mean and unit variance for each input neuron x_i, which is commonly referred to as *standardising*:

$$\bar{x}_i = \frac{1}{N}\sum_{n=1}^{N} x_i^n \tag{2.22}$$

$$\mathrm{Var}(x_i) = \frac{1}{N-1}\sum_{n=1}^{N} (x_i^n - \bar{x}_i)^2, \tag{2.23}$$

where $i = 1, ..., I$ denotes the input units (Figure 2.1) and $n = 1, ..., N$ labels the patterns in the training data set. One can then apply a linear transformation to obtain a set of rescaled variables given by

$$\tilde{x}_i^n = \frac{x_i^n - \bar{x}_i}{[\mathrm{Var}(x_i)]^{1/2}}. \tag{2.24}$$

Effectively, this transformation leads to an equal weighting of the elements in the input vector. Note that as long as the same linear transformation is applied to every pattern in the data set, the information content of the data set is not altered. Network optimisation is driven by the differences between training patterns and not by absolute values; the rescaling due to the standardising does not change the relative differences between these training patterns. Furthermore, we note that the linear tranformation for each element could in theory be equal to the initial value for the network

weights in the input layer (Figure 2.1). Such network weights may take any value, and do not alter the information contained in the input data, which is independent of the weights.

In addition to pre-processing the input, I find that it is beneficial to pre-process the target $\mathbf{t}$ and thus perform a similar operation as in Equation 2.24. Obviously, this linear transformation turns the targets into dimensionless numbers. Therefore, once the network is trained, the linear transformation in Equation 2.24 has to be reversed and applied to the Gaussian kernel means $\boldsymbol{\mu}_j$ and standard deviations σ_j in the MDN output (Equation 2.17). By doing so, these parameters are given in the true physical dimensions of the earth model parameters. I do not correct the standard deviations σ_j for the translation in Equation 2.24, however, as the variance of a probability distribution is invariant under translations. The mixing coefficients α_j are not transformed, as they are dimensionless and sum to one due to the application of the softmax function (Equation 2.18).

Note that the validation and test sets are pre-processed, following Equation 2.24, by using the mean x_i and the variance $\mathrm{Var}(x_i)$ calculated for the training data set (Equations 2.22 and 2.23). Thus the same transformation is applied to the three different synthetic data sets. The same is true for the observed seismic data.

2.4.8 Ensembles of MDNs

Despite measures to maximise generalisation, such as the addition of noise, a neural network will be biased in its performance to the data sets used to train and validate it. Furthermore, a single network is sensitive to the initialisation of its weights. Ideally, we integrate out (marginalise over) the influence of the random initialisation of the network weights, i.e.

$$p(\mathbf{m}'|\mathbf{d}) = \int p(\mathbf{m}'|\mathbf{d}, \mathbf{w}) p(\mathbf{w}|D) \mathrm{d}w, \tag{2.25}$$

where $p(\mathbf{m}'|\mathbf{d}, \mathbf{w})$ is the posterior distribution for $\mathbf{m}'$ conditioned on both the data and the network weights and $p(\mathbf{w}|D)$ is the posterior weight distribution given the available training data $D = \{\mathbf{d}_n, \mathbf{m}'_n\}$. The marginalisation of so-called nuisance parameters plays a central role in the Bayesian framework, e.g. MacKay (2003). In practice, it is very difficult or even impossible to evaluate the integral in Equation 2.25, as this would require sampling from the posterior network weight distribution $p(\mathbf{w}|D)$, which is not known explicitly. By combining the output of multiple networks into one *ensemble*, I aim to reduce the influence of the random initialisation of the network weights. Ensembles of networks can achieve better generalisation, i.e. can make more accurate predictions for unseen data. Admittedly, the use of a simple ensemble as proposed here is no replacement of the integration over the full weight space (Bishop, 1995).

The output of an ensemble of C MDNs can be constructed from a weighted average

of the members (Käufl et al., 2014)

$$p(\mathbf{m}'|\mathbf{d}, \mathbf{w}^*_{i\in 1:C}) = \sum_{i=1}^{C} \frac{\omega_i}{\sum_j \omega_j} p(\mathbf{m}'|\mathbf{d}; \mathbf{w}^*_i), \tag{2.26}$$

where $p(\mathbf{m}'|\mathbf{d}; \mathbf{w}^*_i)$ is the output of a trained MDN (Equation 2.16) and the individual weights ω_i are determined by each network's performance on the same test set

$$\omega_i = \exp\left\{-\frac{E(D_{test}, \mathbf{w}^*_i)}{N}\right\}. \tag{2.27}$$

N is the number of samples in the test set $D_{test} = \{\mathbf{d}_n, \mathbf{m}'_n\}$ and $E(D_{test}, \mathbf{w}^*_i)$ is the error for the ith member (Bishop, 1995)

$$E(D_{test}, \mathbf{w}^*_i) = -\sum_{n=1}^{N} \ln[p(\mathbf{m}'_n|\mathbf{d}_n; \mathbf{w}^*)]. \tag{2.28}$$

The corresponding ensemble error E_{ens} is given by

$$E_{ens}(D_{test}, \mathbf{w}^*_{i\in 1:C})) = -\sum_{n=1}^{N} \ln[p(\mathbf{m}'_n|\mathbf{d}_n, \mathbf{w}^*_{i\in 1:C})], \tag{2.29}$$

with $p(\mathbf{m}'_n|\mathbf{d}_n, \mathbf{w}^*_{i\in 1:C})$ as defined in Equation 2.26. Effectively, the output of an ensemble of C MDNs is a Gaussian Mixture Model with $C \cdot M$ kernels of relative importance

$$\beta_{M\cdot i+j} = \frac{\omega_i}{\sum_j \omega_j} \cdot (\alpha_j)_i, \tag{2.30}$$

where $(\alpha_j)_i$ is the relative importance of the jth kernel in the ith ensemble member, akin to α_j in Equation 2.16.

Bishop (1995) shows that the upper bound on the ensemble error is given by the average error of the individual networks, i.e. $E_{ens} \leq E_{avg}$ with

$$E_{avg} = \frac{\sum_{i=1}^{C} E(D_{test}, \mathbf{w}^*_i)}{C} \tag{2.31}$$

and $E(D_{test}, \mathbf{w}^*_i)$ as defined in Equation 2.28. Figure 2.5 illustrates the advantage of an ensemble of MDNs. The performance of the ensemble on a test set is favourable compared to the performance of the individual members: the ensemble error is similar or slightly lower than the error for the best performing individual network in the committee. In addition, the ensemble encompasses a larger volume of the weight space than any individual member. In this thesis, I use MDN ensembles in Chapters 4 to 8. In Chapter 3, I adopt the simpler approach of training several networks and selecting the one which performs best on a given test set.

As a final comment, I highlight another advantage of network ensembles which I experienced. When training neural networks, it is common practice to use more training samples than there are network weights (LeCun et al., 1998). One 'rule of thumb'

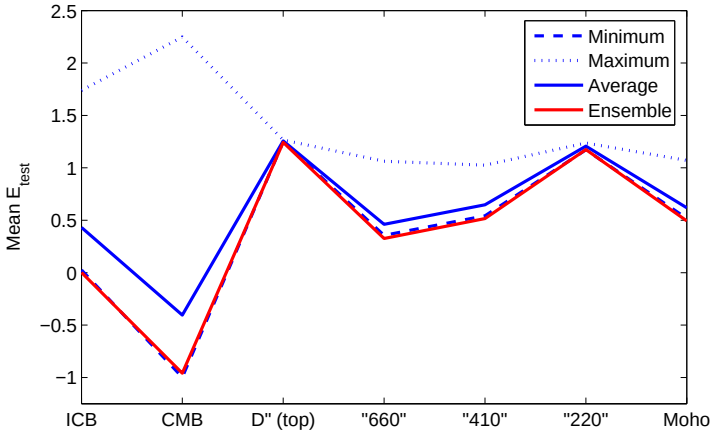


Figure 2.5: Illustration of the advantage of an ensemble of MDNs, comparing the ensemble error (solid red, Equation 2.29) with the average (solid blue, Equation 2.31), minimum (dashed blue) and maximum (dotted blue) error for the ensemble members. All network members are applied to the same test set. The ensemble performs similar or better than the best member and, in addition, samples a larger portion of 'network weight space'. The 1-D targets consist of the seven discontinuity depths, which are shown along the horizontal axis.

suggests that a network will achieve a 90% prediction accuracy if there were ten times as many training patterns as free parameters, e.g. Duda et al. (2001). However, such rules are more relevant for classification problems, in which the labelling can only be right or wrong, than for the regression problems which I solve here, as the MDN outputs a conditional probability distribution for continuous model parameters. Furthermore, in my experience the actual ratio of the number of training samples versus the number of weights is less relevant when using network ensembles. If relatively few training patterns are available, I expect a stronger dependence on the network initialisation and thus a larger variation in performance and output between individual networks. By combining these networks in an ensemble, which corresponds to a weighted average of their output (Equation 2.26), such variations are averaged out and the final ensemble output is more robust than for any individual network. Thus, I do not strictly follow the above rule of thumb, except for Chapter 3, which is the only chapter in which I do not use MDN ensembles. Nonetheless, in this thesis I always ensure that I have more training patterns than there are network parameters, which is good practice in general.

2.4.9 A toy problem

I illustrate the general methodology through a simple toy problem, in which a 1-D inference problems is solved using a Mixture Density Network (Figure 2.6). The toy problem is similar to that used by Bishop (1995) and also serves as a benchmark for

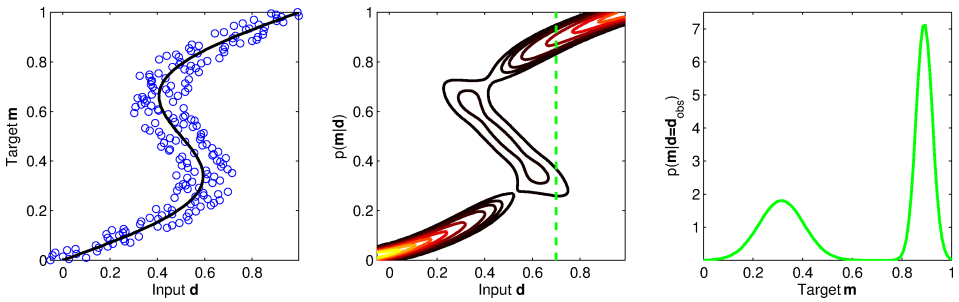


Figure 2.6: Illustrating the functionality of a Mixture Density Network (MDN) using a simple toy problem, inspired by Bishop (1995). Both the input $\mathbf{d}$ and the target $\mathbf{m}$ are one-dimensional. Synthetic samples (blue circles) are generated from an exact function $\mathbf{d} = g(\mathbf{m})$ (black solid line). The goal is to approximate the conditional distribution the $p(\mathbf{m}|\mathbf{d})$ corresponding to the inverse mapping $\mathbf{m} = g^{-1}(\mathbf{d})$ (left-hand panel). To make network training insensitive to observational noise, random Gaussian noise is added to the synthetic samples. Using the noisy synthetic examples, an MDN is trained on the inverse mapping and outputs the conditional pdf $p(\mathbf{m}|\mathbf{d})$ for the full range of input values $\mathbf{d}$ (middle panel). Once the MDN is trained, it can be applied to an observed datum $\mathbf{d}_{obs}$ (green-dashed line), which yields the 1-D marginal posterior pdf (right-hand panel). An MDN outputs a Gaussian Mixture Model, which has the flexibility to represent arbitrary multi-modal and asymmetric distributions.

the implementation of the method. The approach used here is similar to the workflow adopted in the remainder of this thesis.

Here, both the model $\mathbf{m}$ and the data $\mathbf{d}$ are one-dimensional. Synthetic samples are generated from an exact function $\mathbf{d} = g(\mathbf{m})$, similar to the forward relation in Equation 2.2. In this thesis, I am interested in solving inverse problems. As pointed out in Section 2.3.1, the input and output (target) of the neural network are interchangeable. The goal of this toy problem is to infer the posterior pdf $p(\mathbf{m}|\mathbf{d})$, i.e. train the MDN on an input $\mathbf{d}$ and a target $\mathbf{m}$ (Figure 2.6, left-hand panel). To make network training insensitive to observational noise and enhance the generalisation capabilities of the MDN, random Gaussian noise is added to the synthetic samples (Section 2.4.6). Using the noisy synthetic examples, an MDN is trained on the inverse mapping and outputs the conditional pdf $p(\mathbf{m}|\mathbf{d})$ for the full range of input values $\mathbf{d}$ (middle panel). The MDN consisted of five hidden units, for which hyperbolic tangents were used as activation functions, and three Gaussian kernels for the output (mixture model). Only one MDN was trained for this example; no use was made of MDN ensembles yet.

Once training is successful, the idea is to 'invert' the observed data by presenting it as input to the trained MDN. In this context, success is determined by evaluating the MDN performance on test samples not used during training. The various measures of prediction accuracy will be introduced in the relevant sections of later chapters. Here, the MDN is applied to the hypothetical observed datum $\mathbf{d}_{obs} = 0.7$, which yields the 1-D marginal posterior pdf (right-hand panel). Note that in this 1-D problem the marginal pdf corresponds to the full posterior pdf (Equation 2.9). The MDN outputs a Gaussian Mixture Model, which can approximate arbitrary multi-modal and asymmetric distributions, given a sufficient number of Gaussian kernels.

2.5 Kullback-Leibler divergence

In this thesis, I deal with many prior and posterior pdfs for earth model parameters. Following Bayes' theorem (Equations 2.4 and 2.8), the difference between a posterior and prior model distribution is due to the information on the model that is contained in the data **d**. It is often useful to compare these distributions in a quantitative manner, i.e. quantify the information gain or content of the data. Furthermore, such a measure would enable us to quantitatively compare the information on Earth structure contained in different seismic data sets.

A straightforward option is the Kullback-Leibler divergence, or relative entropy (Kullback and Leibler, 1951; MacKay, 2003), which measures the difference between two probability distributions A and B

$$D_{KL}(A||B) = H(A,B) - H(A), \tag{2.32}$$

where $H(A,B)$ is the cross-entropy of A and B and $H(A)$ is the entropy of A. The measure can be interpreted as the amount of information lost when using B to approximate A. The entropy $H(A)$ represents the average uncertainty, surprise or unpredictability of the value of a random variable x that is distributed according to $A(x)$. More specifically, the entropy measures the average Shannon information content (Shannon, 1948) for an ensemble X, which consists of a set of values for x drawn from the probability distribution $A(x)$. If the random variable is continuous, the (Shannon) entropy is commonly referred to as the differential or continuous entropy. The cross-entropy $H(A,B)$ measures the average information needed to identify a sample x in the ensemble X using a distribution $B(x)$, if in truth the ensemble was generated from the distribution $A(x)$.

For a continuous random variable x, D_{KL} can be expressed by the integral

$$D_{KL}(A||B) = \int_{-\infty}^{+\infty} \log_2 \left(\frac{A(x)}{B(x)} \right) A(x) dx, \tag{2.33}$$

which measures D_{KL} in units of bits for a logarithm taken to base 2. The Kullback-Leibler divergence can be shown to satisfy Gibbs' inequality (MacKay, 2003)

$$D_{KL}(A||B) \geq 0, \tag{2.34}$$

i.e. is always non-negative, with equality only if $A = B$. Thus, the entropy of a distribution A is always equal to or less than its cross-entropy with another distribution B. This makes sense intuitively: when identifying a sample in the ensemble X, the average uncertainty is lower (less information is needed) if the true distribution $A(x)$ is used instead of a different proposed distribution $B(x)$. Note that I only consider continuous random variables in this thesis.

For a single model parameter m, I quantify the information gain upon observing the data by calculating the Kullback-Leibler divergence D_{KL} between the 1-D marginal posterior and prior probability distributions. A similar measure was used by for instance Meier et al. (2007a); Käufl et al. (2014). In the above integral (Equation 2.33),

$A(x)$ is the prior pdf $p(m)$ and $B(x)$ represents the 1-D marginal posterior pdf $p(m|\mathbf{d})$, with the observed data $\mathbf{d}$. If the posterior pdf equals the prior pdf, $D_{KL} = 0$ and our knowledge on the parameter m remains unchanged after observing the data $\mathbf{d}$. For reference, consider a 1-D Gaussian distribution with mean μ and standard deviation σ; the difference with a second distribution with the same mean and standard deviation $\frac{1}{2}\sigma$, as measured by the information gain, is 1.16 bits.

2.6 Concluding remarks

In this chapter the key components of the methodology were discussed. To place this work in perspective, I outlined the main issues and approaches in (geophysical) inverse problems. The machine learning approach, which uses neural networks, and in particular Mixture Density Networks, will be used to solve seismological inverse problems. The method relies on Bayesian probability theory, which underlies the inferences on the Earth's internal properties from seismic data in the remainder of this thesis.

The machine learning approach differs from more common sampling-based (Monte Carlo) methods in a few distinct ways. First, I use artificial neural networks as a non-linear regression device using samples from the prior model space. By contrast, Monte Carlo methods aim to sample from the posterior model space directly. If computationally feasible, the latter approach would be preferable. However, sampling directly from the posterior model space can prove difficult, and search algorithms not seldomly arrive in local minima of the error function, or fail to converge to a global minimum. If the mapping of interest is sufficiently smooth, the interpolating capacities of neural networks will prove useful. A trained neural network represents a continuous function that can interpolate between (input) samples and can thus be applied to new unseen data without the need to re-sample the model space. Second, the inferred mapping does not depend on the observed datum, which is used as the reference for data misfit evaluations in Monte Carlo methods. Thus, I can investigate whether or not the seismic data are sensitive to Earth structure, given the assumed data noise level, without requiring the real data. Third, the approach is flexible, in the sense that I can use the same set of prior samples from the model space to test hypotheses on earth model parameters of interest. The trained neural networks approximate the 'inverse' relation between seismic data and earth model parameters.

Since the approximated function is by definition not exact, one can expect (minor) mispredictions with respect to exact inverse operators, which may result in additional errors in our inferences on earth model parameters. However, such true inverse mappings are rarely available in seismological inverse problems, which are often ill-posed and biased by imposed regularisation criteria. Further, our final solution does not constitute a single value for a model parameter. The marginal posterior pdfs represent a range of possible model values, given the data, our prior model distribution, parametrisation and errors in the approximated inverse mapping. To account for the effect of the latter, I have introduced ensembles, or committees, of MDNs. The varia-

tions between the network outputs reflect part of the uncertainty in the approximated function and relate to the variation when training randomly initialised neural networks. From a Bayesian perspective, ideally one integrates (marginalises) over the whole network parameter (weight) space. This is computationally infeasible, however, and the use of the ensembles should be seen as an effective first-order effort to do so.

In summary, the main advantages of artificial neural networks relate to their ability to deal with non-linearities in the mapping of interest, robustness with respect to data noise and the ability to recognise patterns or extract features, that is, patterns that are not perceptible by direct human analysis or more conventional linear statistical methods. As such, neural networks can be very useful in situations where the forward relation is known, but the inverse mapping is unknown or difficult to establish by more conventional analytical or numerical methods. It is worth pointing out that some of these abilities are also met by other machine learning techniques, such as SVMs; by no means do I claim that neural networks are the only possible approach to solve the problems addressed in this thesis. Furthermore, the number of free parameters in ANNs scales favourably with the size of the input, compared to statistical modelling techniques that use fixed basis functions. This makes it computationally feasible to apply neural networks to the non-linear regression problems of interest in this thesis.

3

Bayesian inference of Earth's radial seismic structure from body-wave travel times using neural networks

Abstract

How do body-wave travel times constrain the Earth's radial (1-D) seismic structure? Existing 1-D seismological models underpin 3-D seismic tomography and earthquake location algorithms. It is therefore crucial to assess the quality of such 1-D models, yet quantifying uncertainties in seismological models is challenging and thus often ignored. Ideally, quality assessment is an integral part of the inverse method. Our aim in this study is two-fold: (i) we show how to solve a general Bayesian non-linear inverse problem and quantify model uncertainties, and (ii) we investigate the constraint on spherically symmetric P-wave velocity (V_P) structure provided by body-wave travel times from the EHB bulletin (phases *Pn*, *P*, *PP* and *PKP*). Our approach is based on artificial neural networks, which are very common in pattern recognition problems and can be used to approximate an arbitrary function. We use a Mixture Density Network to obtain 1-D marginal posterior probability density functions

The content of this chapter was published in: de Wit, R. W. L., A. P. Valentine, and J. Trampert, 2013. Bayesian inference of Earth's radial seismic structure from body-wave traveltimes using neural networks. *Geophysical Journal International 195*, 408–422.

(pdfs), which provide a quantitative description of our knowledge on the individual earth parameters. No linearisation or model damping is required, which allows us to infer a model which is constrained purely by the data.

We present 1-D marginal posterior pdfs for the 22 V_P parameters and seven discontinuity depths in our model. P-wave velocities in the inner core, outer core and lower mantle are resolved well, with standard deviations of $\sim$0.2% to 1% with respect to the mean of the posterior pdfs. The maximum likelihoods of V_P are in general similar to the corresponding *ak135* values, which lie within one or two standard deviations from the posterior means, thus providing an independent validation of *ak135* in this part of the radial model. Conversely, the data contain little or no information on P-wave velocity in the D" layer, the upper mantle and the homogeneous crustal layers. Further, the data do not constrain the depth of the discontinuities in our model. Using additional phases available in the ISC bulletin, such as *PcP*, *PKKP* and the converted phases *SP* and *ScP*, may enhance the resolvability of these parameters. Finally, we show how the method can be extended to obtain a posterior pdf for a multi-dimensional model space. This enables us to investigate correlations between model parameters.

3.1 Introduction

Since the start of the twentieth century, the illumination of the Earth's interior by seismic waves has enabled seismologists to infer its seismic velocity and density structure. Current 3-D tomographic models show structural variations in great detail (see e.g. Nolet (2008); Rawlinson et al. (2010) for an overview). Such tomographic inversions are often built upon radial (1-D) earth models. The quality of a 3-D tomographic model is thus intrinsically linked to the robustness of a 1-D model; therefore, it is crucial to assess the quality of the latter. Further, seismological models are frequently used to determine earthquake locations. The spherically symmetric *ak135* model (Kennett et al., 1995), for instance, is used in the location algorithm of the International Seismological Centre (ISC). However, any imperfections in the earth model will map into the source location estimate, e.g. Valentine and Trampert (2012a). Clearly, an accurate estimation of seismic source parameters requires a precise knowledge of the underlying earth model and its uncertainties. However, determining the quality of earth models is non-trivial.

In many geophysical inverse problems, a single 'optimal' solution is obtained via a linearised approach, e.g. Parker (1994); Tarantola (2005). In reality, the dependence of the data on the model often is non-linear. Further, not all model parameters are equally resolved by the data, and seismological inversions usually suffer from a strong model non-uniqueness, e.g. de Wit et al. (2012). Therefore, it is important to understand the uncertainties and resolution associated with the one 'optimal' model. Neglecting these is likely to lead to flaws in any interpretation of the final model. Various approaches are available to assess model quality. For instance, Kennett et al. (1995) used a non-linear search procedure to assess the robustness of the spherically

symmetric *ak135* model in the lower mantle and core, but the velocity bounds for the search procedure itself were based on the final model and relatively narrow, that is, within 0.5% from *ak135*. de Wit et al. (2012) showed how to explore the model null-space to investigate the non-uniqueness of a 3-D tomographic model. Alternatively, a resolution analysis is often employed to investigate the robustness of the inferred Earth structure, using for instance the linear framework provided by Backus and Gilbert (1968, 1970). For example, the resolution of 1-D density structure, as determined from normal mode data, was investigated using both linear (Masters and Gubbins, 2003) and non-linear techniques (Kennett, 1998). However, it would be better to take the non-linearity and model non-uniqueness into account in our inversion framework, rather than treating them *ex post facto*.

A more general approach, which allows us to solve a non-linear inverse problem and to quantify uncertainties, involves the description of our knowledge about earth parameters by probability distributions, e.g. Tarantola and Valette (1982); Tarantola (2005). Following Bayes' theorem (Bayes, 1763), our *posterior* knowledge of the model is our *prior* knowledge updated by the observed data, using a physical theory that relates the model to the data. The aim of Bayesian inference is to obtain the posterior probability distribution $p(\mathbf{m}|\mathbf{d})$ of the model $\mathbf{m}$, conditioned on the observed data $\mathbf{d}$ (Equation 2.4). A common approach is to directly sample the posterior model probability density using Monte Carlo techniques, e.g. Mosegaard and Tarantola (1995); Sambridge (1999a,b); Resovsky and Trampert (2003); Käufl et al. (2013). Beghein et al. (2006) constructed probability density functions (pdfs) to assess whether radial anisotropy in existing 1-D mantle models is robust. While such sampling methods are powerful for solving non-linear inverse problems, they quickly deteriorate as the dimension of the model space increases, a phenomenon which Bellman (1961) termed the *curse of dimensionality*. In practice, this often limits the use of sampling methods to inverse problems which involve relatively few model parameters, i.e. a few tens. As an alternative to Monte Carlo techniques, we use artificial neural networks to solve the Bayesian inverse problem. Neural networks can be viewed as non-linear filters and are very common in pattern recognition problems. They can approximate an arbitrary non-linear relation between two parameter spaces, inferring the mapping from a set of training data. See the detailed explanation given in Section 2.4 and common references on artificial neural networks, e.g. Bishop (1995); MacKay (2003).

Here we perform a Bayesian inversion of P-wave travel time curves for the Earth's spherically symmetric P-wave velocity (V_P) structure. We use travel times from the EHB bulletin (Engdahl et al., 1998) for the *Pn*, *P*, *PP* and *PKP* phases. The inverse problem is non-linear, as the ray paths of the seismic phases depend on the underlying velocity structure of the Earth. Our focus is two-fold. First, we demonstrate how to solve a Bayesian non-linear inverse problem and assess model uncertainties using neural networks. Second, we quantify the constraint on radial V_P structure which is provided by the travel time data for these major seismic phases. To solve our non-linear inverse problem, we use a particular class of neural networks, known as a Mixture Density Network (MDN, Section 2.4.4, Bishop (1995)). An MDN outputs a parametric distribution, which approximates the continuous posterior model proba-

bility density. This distribution reflects our updated state of knowledge on the earth model parameters.

First, we describe the model parametrisation and the travel time data used for this study. We refer the reader to Chapter 2 for an outline of the Bayesian inversion framework (Section 2.2) and an explanation of the neural network technique used here (Section 2.4). Second, we invert the travel time data using neural networks and show the 1-D marginal probability density functions (pdfs) for P-wave velocity parameters and discontinuity depths. We emphasise that our focus lies on the constraints on individual model parameters; we do not present a new 1-D earth model.

3.2 Model parametrisation

We adopt a piecewise continuous representation for the P-wave velocity model, as was used for the *prem* (Dziewoński and Anderson, 1981) and *iasp91* (Kennett and Engdahl, 1991) models. The piecewise continuous functions can be used to evaluate the model at any depth exactly. We parametrise the depths of seven discontinuities in the V_P profile: the inner core boundary (ICB) and core-mantle boundary (CMB), the top of the D″ layer, the discontinuities around 660, 410 and 210 km depth and the Moho. We define the lower mantle (LM) as the region between the top of the D″ layer and the 660 km discontinuity, while the transition zone (TZ) spans the region between the 660 and 410 km discontinuities.

Between the discontinuities, we parametrise the P-wave velocity structure at L depths (knots). Subsequently, we construct the V_P profile $f(z)$, i.e. the 1-D velocity structure between the discontinuities, by interpolating between the L knots using a set of L natural cubic spline functions

$$f(z) = \sum_{i=1}^{L} a_i \psi_i(z). \tag{3.1}$$

Each spline function $\psi_i(z)$ is a function of the depth z and equals 1 at one knot, while being 0 at the remaining $L-1$ knots (Figure 3.1). The coefficients a_i represent the V_P values at each knot. This yields a piecewise continuous representation of the model.

The transition zone and the region between the 410 and 210 km discontinuities are parametrised using Equation 3.1 with $L = 2$. This results in linear velocity gradients with depth in these layers. We separate the region between the Moho and 210 km discontinuity in two linear ($L = 2$) segments, i.e. 210–120 km and 120-Moho km, as the linear velocity gradient in these two segments differs significantly in existing 1-D models such as *ak135*. Thus, the velocity profile is continuous at 120 km depth, but its first derivative is not. The depth of the transition at 120 km is not varied. The lower mantle and outer core are parametrised by 4 knots, and the inner core by 3, which results in non-linear gradients with depth (Figure 3.1). The crust is parametrised by two homogeneous layers. No sediment or water layers are present. We thus have 29 parameters in our model $\mathbf{m}$: 22 V_P parameters (the coefficients a_i in Equation 3.1) and 7 discontinuity depths.

Tables 3.1 and 3.2 define the prior model distribution $\rho(\mathbf{m})$ (Equation 2.8). Discontinuity depths are independently drawn from uniform priors, as are the V_P values directly below the discontinuities (Table 3.1). The prior distributions are centred on the corresponding values of the *ak135* model. We choose conservative priors by allowing for a large variation in the independent model parameters, i.e. $\pm 3\%$ with respect to *ak135* for V_P in the core and lower mantle and $\pm 5\%$ in the upper mantle. We emphasise that by choosing such broad prior distributions, we ensure that the results of our probabilistic inversion are not driven by the actual values in the *ak135* model.

The V_P values at the other $L-1$ knots in each region are calculated using the new value at the first knot (m_d^1 in Table 3.1) and the gradient of the *ak135* model. Subsequently, these values are perturbed, with the amount of perturbation drawn from a uniform prior (Table 3.2). This introduces a correlation between the V_P parameters in each region. In general, radial P-wave velocity increases with depth, i.e. the velocity gradient is mostly positive. By using the gradient in *ak135*, we aim to exclude physically implausible models and restrict the size of our model space.

We generate 99 862 synthetic models, which are drawn from the prior model distribution $\rho(\mathbf{m})$. Ten synthetic models in the training set are shown in Figure 3.2, along with the *ak135* model. Note that locally negative velocity gradients can still exist in the models.

3.3 Travel time data

3.3.1 EHB data

The EHB bulletin (Engdahl et al., 1998) contains millions of routinely determined travel time measurements, which have been corrected for source mislocation. We select travel time data for the *Pn*, *P*, *PP*, *PKPab*, *PKPbc* and *PKPdf* phases for the years 2001–2008 (Figure 3.3). For simplicity, we exclude *PnPn*, the upgoing phases *pP*, *pwP* and *sP* and the crustal phases *Pb* and *Pg*.

Several corrections are provided with the EHB bulletin. We correct the raw EHB data for the Earth's ellipticity and station elevation. We do not apply the regionally smoothed station corrections, which perform regional averaging ($5 \times 5^\circ$ patches) to smooth out effects of lateral heterogeneities in the upper mantle. In our setup, the imprint of 3-D structure on the travel time measurements is treated as noise; therefore, such a correction is unnecessary here.

We select the travel times for which the EHB estimated source depth lies between 14 and 16 km. Note that this range of depths is chosen to approximate a fixed source around 15 km depth and not to represent the uncertainty in EHB source depth estimates; we will discuss our data noise model below in Section 3.3.3. Each event in the EHB bulletin is given a three-letter label that characterises the quality of hypocentral determination. We exclude the EHB measurements for which the source depth is fixed to a standard depth by Engdahl, denoted by FEQ in the EHB data, and solutions for which the uncertainty in source depth is expected to be $>$15 km (LEQ, XEQ). The

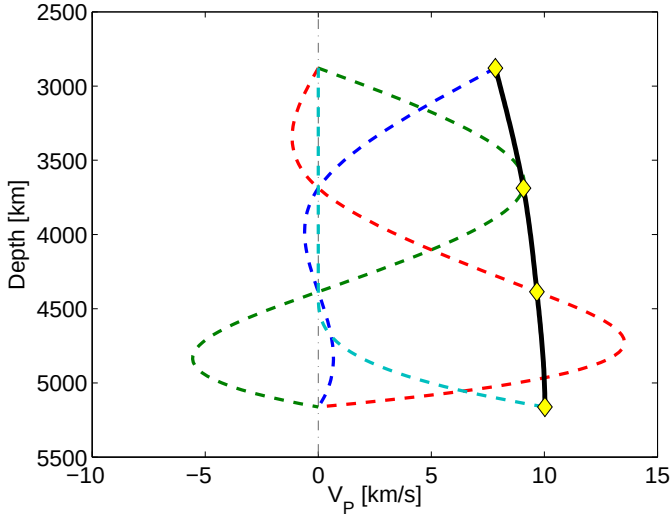


Figure 3.1: An example of the natural cubic splines used to construct the 1-D earth models, following Equation 3.1. The coefficients a_i represent V_P values in the outer core (yellow diamonds), drawn from the prior model distribution $\rho(\mathbf{m})$. The black line shows the resulting V_P profile $f(z)$ in the outer core, which is constructed by summing the products $a_i\psi_i(z)$ for the four splines (dashed).

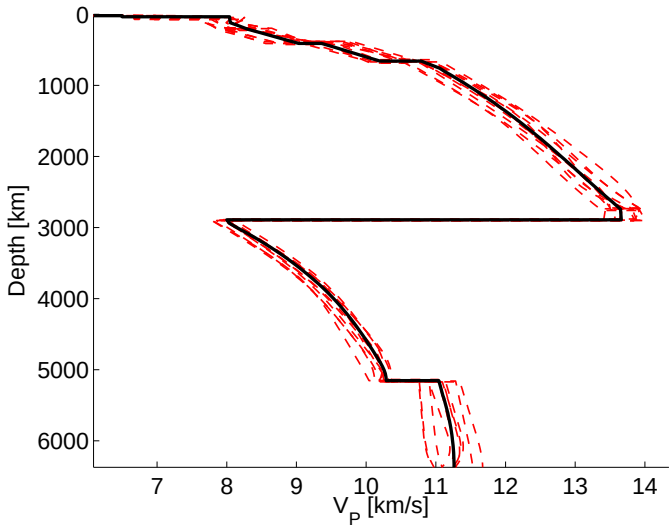


Figure 3.2: Ten model realisations (red), drawn from the prior distribution $\rho(\mathbf{m})$, and *ak135* (black) for V_P.

Table 3.1: Prior information on *independent* model parameters. Prior distributions are uniform over the specified ranges. The ranges for the P-wave velocity parameters are given as percentile perturbations from *ak135* (Kennett et al., 1995), except for V_P in the two crustal layers. V_P parameters are indicated by m_d^1 and represent the knots located directly below a discontinuity d. Note that the tops of the transition zone (TZ) and the lower mantle (LM) are formed by the 410 and 660 km discontinuities, d_{410} and d_{660}, respectively. The interpolation style for the V_P profile in every region, following Equation 3.1, is given in the last column.

Discontinuity	Parameter	Range [km]	
Inner-outer core (ICB)	d_{ICB}	$5133.5 - 5173.5$	
Core-mantle (CMB)	d_{CMB}	$2871.5 - 2911.5$	
D″ layer (top)	$d_{\mathrm{D''}}$	$2720 - 2760$	
660 discontinuity	d_{660}	$630 - 690$	
410 discontinuity	d_{410}	$380 - 440$	
210 discontinuity	d_{210}	$190 - 230$	
Moho	d_{Moho}	$25 - 75$	
Region	**Parameter**	**Range [±%]**	**Interpolation style**
Inner core (IC)	m^1_{IC}	3	3 cubic splines
Outer core (OC)	m^1_{OC}	3	4 cubic splines
D″ layer	$m^1_{\mathrm{D''}}$	3	linear
Lower mantle (LM)	m^1_{LM}	3	4 cubic splines
Transition zone (TZ)	m^1_{TZ}	5	linear
410–210	m^1_{210}	5	linear
210–Moho	m^1_{Moho}	5	
210–120			linear
120–Moho			linear
Region	**Parameter**	**Range [km/s]**	
Lower crust (LC)	m_{LC}	$6.4 - 7.4$	
Upper crust (UC)	m_{UC}	$5.6 - 6.3$	

Table 3.2: Prior information on *dependent* model parameters. Prior distributions are uniform over the specified ranges, which are given as percentile perturbations from the updated model value (see text). The indices m_d^i represent the correlated model parameters in every region d, with higher indices i corresponding to deeper V_P knots in the parametrisation. The corresponding independent parameters m_d^1 are listed in Table 3.1.

Region	Parameter	Range [±%]
Inner core (IC)	$m_{\mathrm{IC}}^{2,3}$	1
Outer core (OC)	$m_{\mathrm{OC}}^{2,3,4}$	1
D″	$m_{\mathrm{D''}}^{2}$	1
Lower mantle (LM)	$m_{\mathrm{LM}}^{2,3,4}$	1
Transition zone (TZ)	m_{TZ}^{2}	2
410–210	m_{210}^{2}	2
210–Moho	$m_{\mathrm{Moho}}^{2,3}$	2

remaining data correspond to 1100 events that were registered at 5268 different stations (Figure 3.4). Both sources and receivers are globally distributed, that is, within the typical limitations of seismological data coverage.

3.3.2 Synthetic data

Neural network training and validation requires a data set containing many examples of input-output pairs. For this purpose, we calculate synthetic first-arrival travel time curves for 99 862 synthetic models using the TauP package (Crotwell et al., 1999). The synthetic models are parametrised as described in the previous section. All 29 model parameters are allowed to vary in each model realisation. The source depth is fixed to 15 km for all synthetic data. Thus, source depth is a fixed parameter in our setup and any uncertainties in EHB depth estimates are regarded as a source of data noise, as will be discussed below in Section 3.3.3. The three PKP branches are calculated separately. Travel times are computed for a phase-specific range of epicentral distances at one degree intervals (Table 3.3). For these phases, the distance ranges are comparable to those used by Kennett et al. (1995), who used travel time data collected by the ISC. We include *Pn*, which refracts along the Moho, to provide a better constraint on the structure of the uppermost mantle.

If for a given distance no arrival is computed by TauP, the travel time is set to zero. By doing so, the number of elements in every travel time branch is constant. This is a requirement of the network architecture, which only permits input vectors of constant dimension. The zeros represent gaps in the travel time curves, which commonly result from low-velocity zones, and associated negative gradients, in the earth model. This information is therefore available to the neural network. Note that for the epicentral

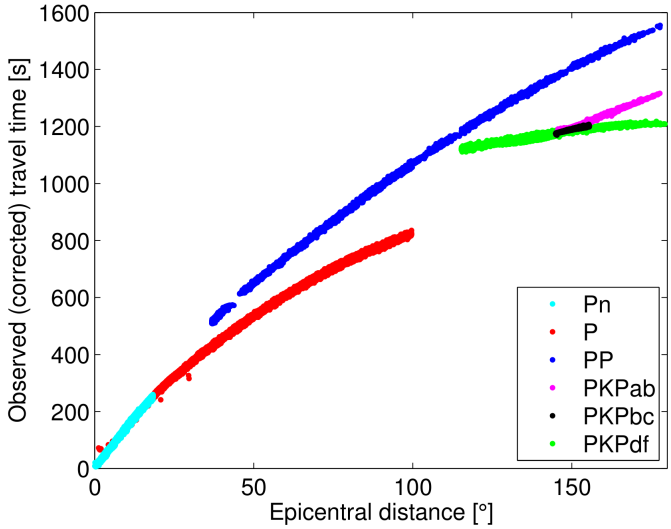


Figure 3.3: Travel time measurements in the EHB bulletin for 2001–2008. Event depths are restricted to lie between 14 and 16 km.

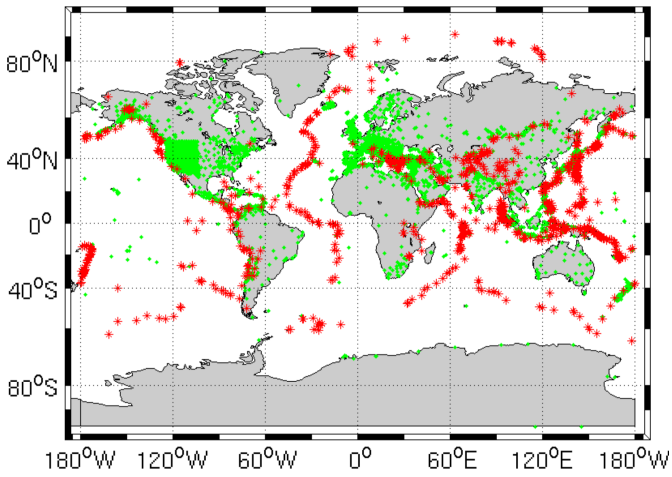


Figure 3.4: Locations of 1100 sources (red asterisks) and 5268 stations (green dots) in the EHB data for 2001–2008. Event depths are restricted to lie between 14 and 16 km.

Table 3.3: Epicentral distance range for the seismic phases. The ranges used by Kennett et al. (1995) are added as a reference.

Distance range [°]	Pn	P	PP	PKPab	PKPbc	PKPdf
This study	3:18	25:88	50:173	145:174	145:155	122:179
Kennett et al. (1995)	—	25:99	53:180	156:178	151:153	118:180

Table 3.4: Phase-specific measurement error ϵ_{ISC} for the travel time data, as documented in ISC (2008), which serves as a minimum for the standard deviation ϵ_{ip} (Figure 3.5).

Phase	Pn	P	PP	PKPab	PKPbc	PKPdf
$\epsilon_{ISC}[s]$	0.8	0.8	1.3	1.3	1.3	1.3

distance ranges used (Table 3.3), no gaps occur in the globally distributed EHB data (Figure 3.3). This could indicate that there are no global low-velocity zones in the parts of the Earth sampled by the data, or that some of the EHB travel times do not represent a direct geometric arrival (or a combination of both).

3.3.3 Data uncertainties

Uncertainties exist in both the epicentral distance, through the source location estimate, and the travel time measurements. We add noise to the synthetic data to simulate these two types of uncertainty.

For every synthetic travel time measurement, we draw a perturbation to the epicentral distance from a uniform distribution $\mathcal{U}(-\epsilon_{\text{dist}}, +\epsilon_{\text{dist}})$ with $\epsilon_{\text{dist}} = 0.1^\circ$ (~ 10 km). The value of ϵ_{dist} is similar to the average test-event mislocation reported by Engdahl et al. (1998). The corresponding travel time is updated by applying the local gradient in the travel time curve to this perturbation.

Second, the synthetic data have to be corrupted to reflect noise in the travel time data. The scatter in the observed travel time data (Figure 3.3) originates from lateral heterogeneities in the Earth, measurement errors, phase misidentifications and uncertainties in the estimated source depth. For a given phase p and epicentral distance i, we estimate the noise as the spread in the EHB travel times. The half-width of this spread ϵ_{ip} (Figure 3.5) is used to define a Gaussian noise distribution $\mathcal{N}(0, \epsilon_{ip}^2)$, i.e. with zero mean and standard deviation ϵ_{ip}. A random sample from this noise distribution is added to every synthetic datum. The scatter in the data may be small if few data are available, which would result in an unrealistically low noise estimate. Therefore, we impose a minimum phase-specific noise level (Table 3.4), which is based on measurement error estimates documented in a recent ISC report (ISC, 2008).

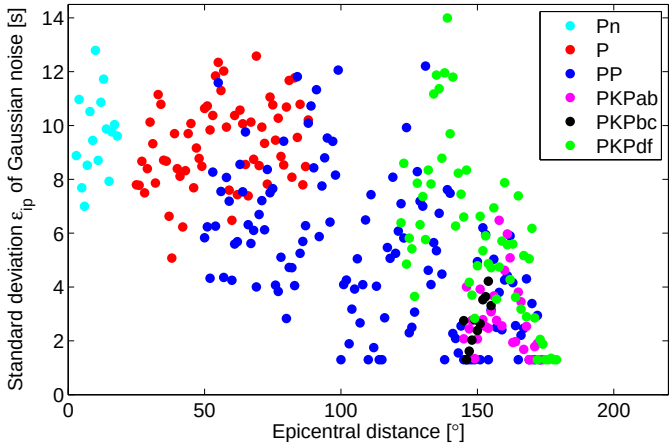


Figure 3.5: The half-width of the spread in the observed travel times ϵ_{ip} is used as the standard deviation of the Gaussian noise model, which differs for every phase p and epicentral distance i. Different colours denote different phases.

3.3.4 Data processing

The input to the neural network is a concatenation of the travel time curves for the different phases $\mathbf{d}$. Since the curves are rather smooth, a large (linear) correlation exists between the travel time at different epicentral distances. Therefore, we sample the travel time curves at intervals of 2°. This reduces the size of the input vector and thus the number of network parameters, thereby making network training faster. In light of the strong correlations, we assume that this downsampling does not result in a significant loss of information on our Earth model. The resulting input vector consists of 152 travel times, which is a concatenation of the data for the Pn (8 travel time measurements), P (32), PP (62), PKPab (15), PKPbc (6) and PKPdf (29) phases.

For every synthetic model, the input pattern is constructed by sampling the synthetic data at 2° intervals and subsequently adding a random noise sample (Figure 3.6). The input vectors for the observed data are constructed in a slightly different fashion. For each distance sample d_{ip}, with phase p and epicentral distance i, we extract all observations from the EHB data for which the epicentral distance lies within the range $d_{ip} \pm \epsilon_{\text{dist}}$, where $\epsilon_{\text{dist}} = 0.1°$ as before. Consequently, multiple EHB observations are available at each distance point. We draw one random sample from these multiple possibilities. This yields the EHB travel time curves that serve as input to the trained networks (Figure 3.7). The differences between these curves are regarded as noise (see Section 3.3.3). The variations in the real data vectors are significantly smaller than the variations in synthetic training patterns (cf. Figure 3.6). Both observed and synthetic data contain noise, but the variations in the synthetic data are larger due to the differences in the underlying synthetic earth models.

Note that upon applying a trained network to one input pattern for the EHB data, only 152 measurements are 'inverted'. For a given distance and phase, however, all available observations should contain the same information on the radial earth model, given the measurement noise defined above. When repeated with a different EHB input pattern, constructed as described here, the inversion should yield similar results.

3.4 Results

We present inversion results for all 22 P-wave velocity and seven discontinuity depth parameters. For each model parameter m_i, we investigate the constraint that is provided by the travel time data and quantify the associated uncertainty. The network input $\mathbf{x}$ corresponds to the concatenated body-wave travel time curves $\mathbf{d}$ and the target $\mathbf{t}$ is given by the individual parameters in the radial P-wave velocity earth model $\mathbf{m}$ (Figure 2.3). Thus, we train MDNs on 1-D target vectors $\mathbf{m}' = m_i$ (Equation 2.9). Since all model parameters are allowed to vary in the synthetic models, the output of each MDN forms a 1-D marginal posterior pdf $p(m_i|\mathbf{d})$. This is equivalent to marginalising the full joint posterior pdf $p(\mathbf{m}|\mathbf{d})$ over all model parameters other than m_i via the integration in Equation 2.9. $p(m_i|\mathbf{d})$ reflects our knowledge of m_i given the variations in the other 28 model parameters.

3.4.1 Network configuration

For all results presented in this study, we train MDNs with 40 hidden units and a Gaussian mixture consisting of 15 Gaussian kernels ($M = 15$). We verified that the precise number of hidden units is not of paramount importance to final network performance. The same applies to the number of Gaussian kernels. During training, the mixing coefficient α_j can be set close to zero for redundant kernels, or kernels can be combined by having a similar mean and variance (Bishop, 1995).

For a one-dimensional target ($c = 1$), the MDN has $(c + 2) \cdot M = 45$ output parameters: the means μ_j, the variances σ_j^2 and the mixing coefficients α_j (Equations 2.16 and 2.17). In combination with the 152-D input pattern (Figures 3.6, 3.7), the corresponding MLP has 7725 free parameters (the weights $w_{ij}^{(1)}$ and $w_{jk}^{(2)}$ and biases $b_j^{(1)}$ and $b_k^{(2)}$ in Figure 2.1). We use 80% of the 99 862 patterns in the synthetic data set for training, 15% for the validation set, which is used to evaluate the early stopping criterion, and the remaining 5% for the test set.

Theoretically, there is no limit to the size of a neural network. However, a larger network consists of more free parameters and thus takes longer to train. More importantly, more network parameters require a larger training set to succesfully train the network. Therefore, computational facilities restrict the network size. In general, the number of free parameters should not exceed the number of training patterns, e.g. Bishop (1995); Duda et al. (2001). In this study, we ensure that the training set is approximately a factor of 10 larger than the number of network parameters. The main computational requirement thus lies in the generation of synthetic training patterns,

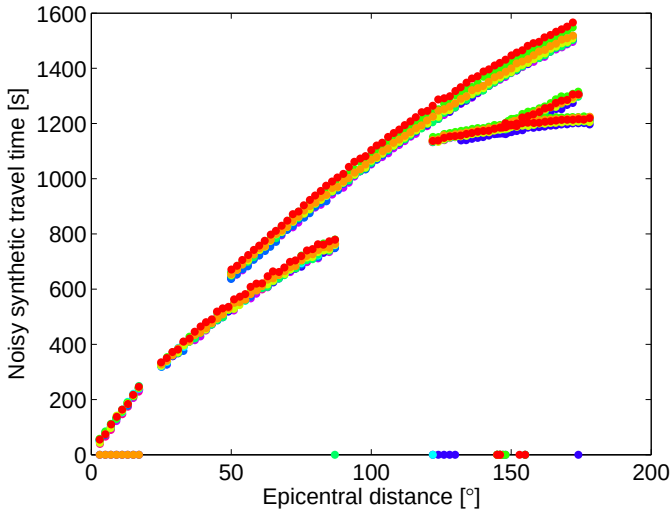


Figure 3.6: Examples of noisy synthetic travel time data that form the input to the network. The synthetic data are sampled at distance intervals of 2° for the epicentral distance ranges specified in Table 3.3. Note the zeros in the travel time curves, which indicate that TauP did not compute an arrival at the corresponding distance.

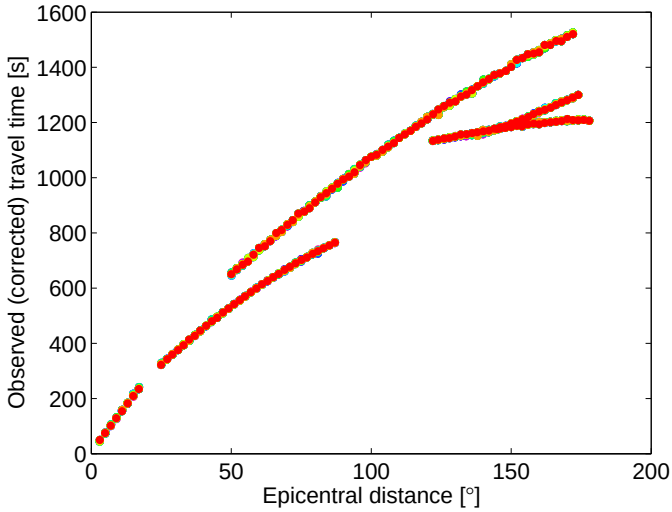


Figure 3.7: Ten input patterns, which are constructed from the EHB data (Figure 3.3). Note that the variation in these real data vectors is significantly smaller than the variations between synthetic training patterns (Figure 3.6).

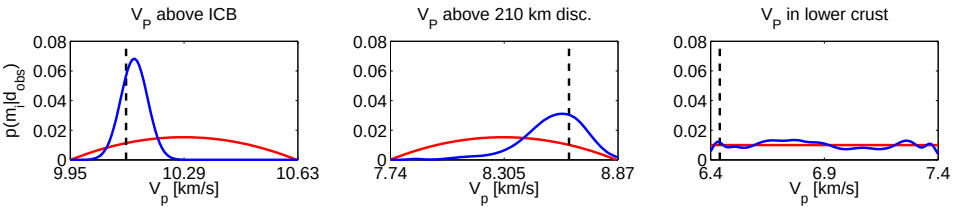


Figure 3.8: 1-D marginal posterior pdf (blue line), prior pdf (red) and true model value (black, dashed) for one pattern in the test data set for V_P (left-hand frame) directly above the ICB, (middle frame) directly above the 210 km discontinuity and (right-hand frame) in the lower crust.

i.e. repetitively solving the forward problem. For the $\sim$ 100,000 patterns, this took $\sim$ 100 hours on a standard desktop computer. Once the training set is available, network training is relatively fast: the training time for a single network is on the order of tens of minutes in this study.

Due to the random initialisation of the network weights, the optimisation algorithm can become stuck in local minima of the error surface. To verify that network training converges properly, we train 30 independent networks. For each of these networks, training commences at a different point in weight space due to the random initialisation. To reduce the chance of overfitting, all synthetic patterns are divided randomly over the training, validation and test sets before the training of each independent network commences. We find that results for the 30 independent networks are similar and choose the network that produces the lowest pattern-averaged error for the test set.

3.4.2 Network evaluation

For each V_P parameter, we apply the optimal MDN to the $\sim$5000 patterns in the synthetic test set, which are not used for network training. Network performance can be evaluated by comparing the resulting $\sim$5000 1-D marginal pdfs with the true synthetic model value. As an example, we show 1-D marginal posterior pdfs for one test pattern for three model parameters: V_P directly above the ICB, directly above the 210 km discontinuity and in the lower crust (Figure 3.8). For the P-wave velocity near the ICB (left-hand frame) and 210 km discontinuity (middle frame), most probability mass in the posterior distribution lies close to the target value (black line). We conclude that network performance is accurate for this particular input pattern. It is clear from the marginal pdf of V_P in the lower crust (right-hand frame) that the travel time data do not constrain this part of the model. Consequently, the MDN output represents the uniform prior distribution for this parameter. The difference between the MDN output and the true uniform prior pdf is due to the fixed shape of the finite number of Gaussian kernels.

It is difficult to quantitatively evaluate network performance from such marginal distributions, however. A more pragmatic validation method is to analyse the correlation between the target value and the mean of the dominant Gaussian kernel in the

MDN output (Bishop, 1995). We take the kernel associated with the largest mixing coefficient α_j (Equation 2.16) to be dominant. Although this simple measure ignores the information provided in the full posterior pdf, such an analysis provides a practical way of evaluating the trained networks.

For all 22 V_P parameters, Figure 3.9 shows the mean μ_j of the dominant Gaussian kernel versus the target value for all patterns in the test data set. Every row in the figure represents a region between two discontinuities, with the depth of the V_P parameter (knot) decreasing from left to right. The corresponding correlation coefficient R is given above every frame ($R = 0$ indicates no correlation, whereas $R = 1$ represents a perfect correlation). Network performance on these unseen input patterns is good ($R \geq 0.87$) for the P-wave velocity in the inner and outer core and lower mantle (first, second and fourth row, respectively). For V_P in the D'' layer (third row), the upper mantle (fifth and sixth row) and crust (bottom row), correlations are in general low or absent ($R \approx 0$).

Besides network evaluation, the performance on synthetic input is a good indicator of the constraint that the data provide on the model. For the P-wave velocity directly above the ICB, for instance, the 1-D marginal posterior pdf is unimodal and narrow relative to the width of the prior pdf (left frame, Figure 3.8). This parameter is constrained well by the data, as indicated by $R = 0.94$ in Figure 3.9 (second row, first column). Conversely, for V_P in the lower crust the mean of the 'dominant' kernel does not relate to the true value ($R = 0.05$, Figure 3.9, seventh row, first column). The travel time data do not constrain this part of the model, as is apparent from the corresponding marginal pdf in Figure 3.8 (right).

One can expect similar results, for both resolved and unresolved model parameters, when applying the trained networks to the observed travel time data. As the data provide very little or no constraint on the seven discontinuity depth parameters, we do not show the corresponding performance on the test set here and restrict ourselves to the application to the EHB data for these parameters.

3.4.3 Application to EHB data

Figure 3.10 shows 1-D marginal pdfs for P-wave velocities for the ten EHB input patterns in Figure 3.7. Recall that these ten input patterns are random realisations from the available EHB data set, as described in Section 3.3.4. The differences between these input patterns are regarded as noise. The network should be insensitive to such variations, since we have used a similar noise level during network training. Consequently, network output should be approximately the same for these different input vectors.

The data constrain V_P in the outer core (OC) and lower mantle (LM) best (second and fourth row, respectively). PKPdf, the only seismic phase in our data set that is sensitive to the inner core (IC), constrains V_P in this region (first row). The most notable feature is the proximity of the posterior maxima to the *ak135* values, which are indicated by the green dashed lines. However, in addition to a most likely model value, we obtain uncertainties in the P-wave velocity estimate. Recall that our posterior pdfs

are based on our conservative prior pdfs and are therefore taken to be independent of the actual values in the *ak135* model.

For each V_P parameter in the inner core, outer core and lower mantle, we extract statistics from the 10 posterior distributions in Figure 3.10. Since the ten distributions are similar for these parameters, we calculate the mathematical expectation $\langle V_P \rangle$ and standard deviation σ_{V_P} of the ten pdfs combined (Table 3.5). When expressed as a percentage of the mean $\langle V_P \rangle$, the standard deviation of the posterior pdfs σ_{V_P} is smaller than 1% for every model parameter. The corresponding *ak135* values are given as a reference and lie within one standard deviation from $\langle V_P \rangle$ for the parameters in Table 3.5, except for $m^{2,3}_{\mathrm{LM}}$ in the lower mantle, for which *ak135* lies within two standard deviations. It is difficult to compare the uncertainties found here, represented by the standard deviations σ_{V_P}, as uncertainty estimates are scarce in the literature. Kennett et al. (1995) used a non-linear search procedure to evaluate a range of models around *ak135* for various data misfit measures. They used velocity bounds of ± 0.02 km/s for the lower mantle and ± 0.04 km/s for the lowermost mantle and core. The σ_{V_P} values we find here are on the order of these velocity bounds or a factor of 2-3 larger (Table 3.5). We emphasise that our uncertainty estimates are not representative of the uncertainties in *ak135*, given the differences in the data selection and the model parametrisation. However, the constraint on individual model parameters, as investigated here, may be indicative of the uncertainties in *ak135* and similar models.

The 1-D marginals for V_P between the Moho and the 210 km discontinuity (Figure 3.10, sixth row) indicate that the travel time data contain some information on this region. We find that this is mainly due to the addition of the Pn phase, which refracts along the Moho. The data indicate a (very) weak preference for velocities slightly higher than in *ak135*. We should point out, however, that the posterior distributions include all but the very low P-wave velocities. Thus, for these model parameters the data cannot falsify any of the assumptions on V_P contained in our model prior.

The data contain no or at most a very limited signal on the parameters of the D″ layer (third row, Figure 3.10), the transition zone (TZ, fifth row), the region between the 410 and 210 km discontinuities (410–210, fifth row) and the two homogeneous crustal layers (seventh row). The poor constraint on upper mantle structure is not surprising, given the teleseismic epicentral distances for which *P* (above 25°) and *PP* (above 50°) travel times are used (Table 3.3).

The travel time data we invert here contain practically no information on the depth of the discontinuities (Figure 3.11). For the upper mantle, this may be explained by the near-vertical incidence under which the teleseismic rays travel through the discontinuities in this region. The inclusion of phases that reflect of discontinuities, e.g. *PcP*, which reflects of the CMB, could improve the constraint on discontinuity depths.

3.5 Discussion

We trained neural networks to invert P-wave travel time data for the radial P-wave velocity structure of the Earth. We obtained a continuous probabilistic description of

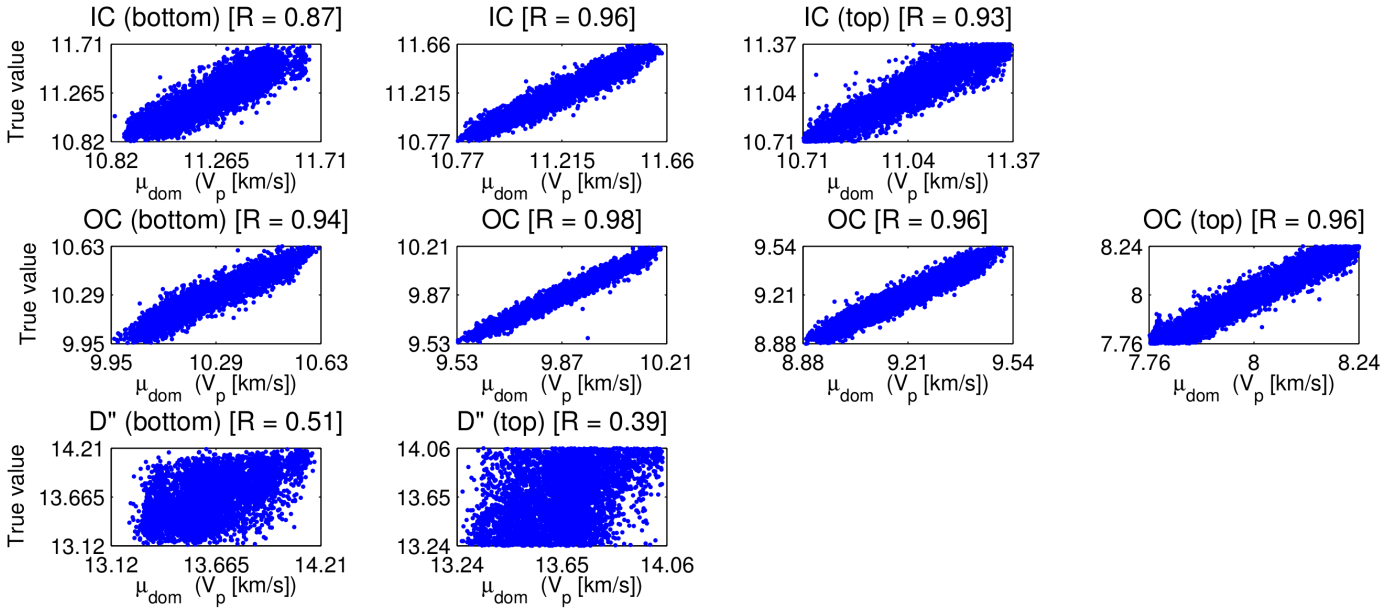


Figure 3.9: Mean μ_j (Equation 2.17) of the dominant Gaussian kernel (maximum α_j in Equation 2.16), labelled μ_{dom} in the figure, versus the true synthetic value for all patterns in the test data set for the 22 V_P parameters (Tables 3.1 and 3.2). The regions between discontinuities are displayed on different rows. For every region, depth decreases from left to right in the figure. The corresponding correlation coefficient R is given above each frame.

Figure 3.9 (continued)

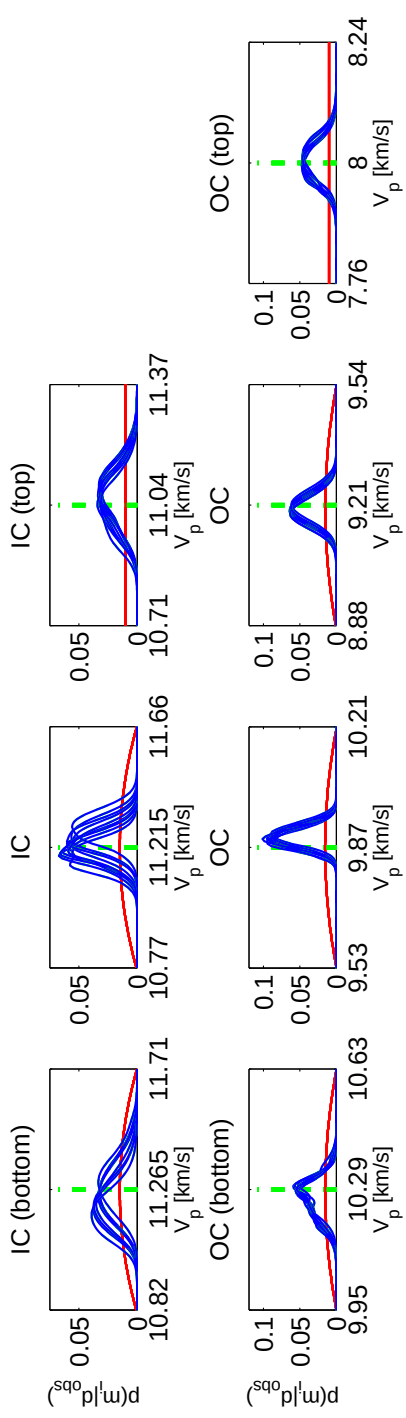


Figure 3.10: 1-D marginal posterior pdf (blue line), prior pdf (red) and the corresponding *ak135* value (green, dashed) for the 22 V_P parameters (Tables 3.1 and 3.2). The same trained networks as used for Figure 3.8 are applied to 10 different EHB input patterns (Figure 3.7). The regions between discontinuities are displayed on different rows. For every region, depth decreases from left to right in the figure. Note that the range of the vertical axis, i.e. normalised probability, differs between rows.

Figure 3.10 (continued)

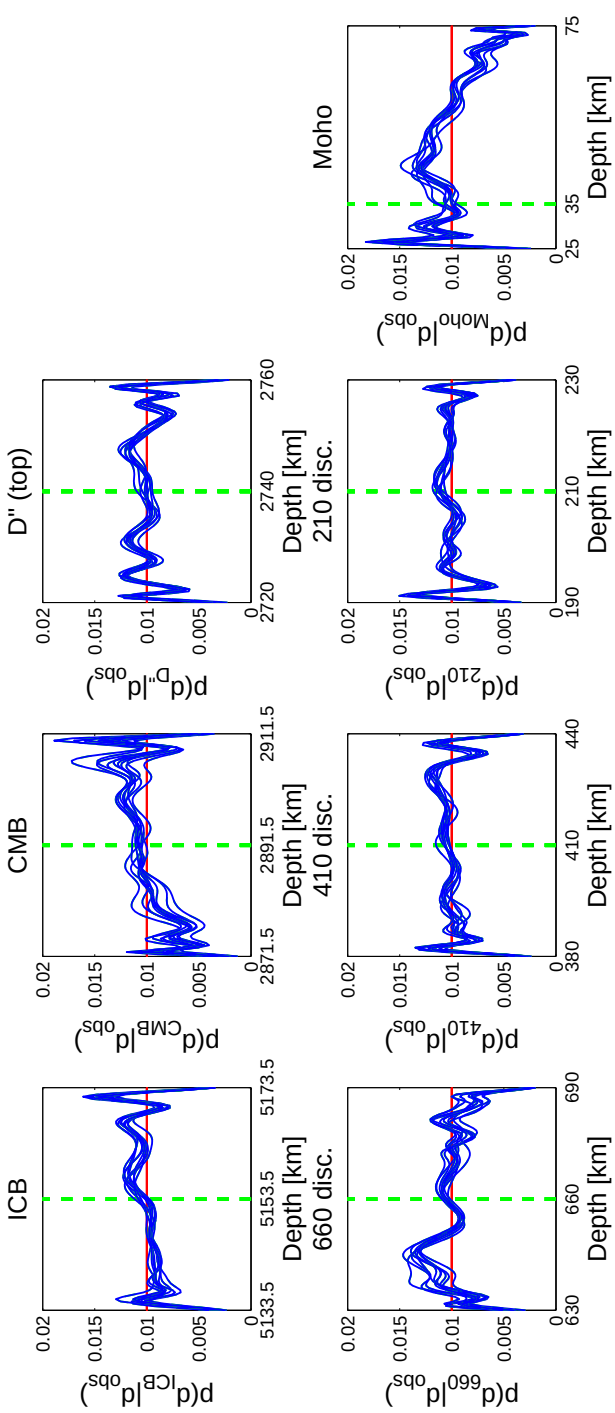


Figure 3.11: 1-D marginal posterior pdf (blue line), prior pdf (red) and the corresponding *ak135* value (green, dashed) for the seven discontinuity depths (Table 3.1). Trained networks are applied to 10 different EHB input patterns (Figure 3.7).

Table 3.5: Mean $\langle V_P \rangle$ and standard deviation σ_{V_P} of the ten 1-D marginal posterior pdfs for the P-wave velocity parameters (Figure 3.10) in the inner core (IC), outer core (OC) and lower mantle (LM). All values are in [km/s], except for the fourth column, which shows the standard deviation σ_{V_P} as a percentage of the mean $\langle V_P \rangle$. The corresponding value in *ak135* is given for comparison (V_P^{ak135}). Recall that the depth of the knots m^i decreases with decreasing index number i (see Tables 3.1 and 3.2).

Parameter	$\langle V_P \rangle$	σ_{V_P}	σ_{V_P}[%]	V_P^{ak135}
m^3_{IC}	11.248	0.107	0.952	11.265
m^2_{IC}	11.229	0.066	0.589	11.215
m^1_{IC}	11.063	0.081	0.736	11.040
m^4_{OC}	10.266	0.056	0.545	10.290
m^3_{OC}	9.899	0.030	0.304	9.870
m^2_{OC}	9.201	0.043	0.466	9.210
m^1_{OC}	8.004	0.041	0.506	8.000
m^4_{LM}	13.638	0.086	0.629	13.650
m^3_{LM}	12.819	0.031	0.239	12.860
m^2_{LM}	12.034	0.039	0.320	11.995
m^1_{LM}	10.793	0.057	0.528	10.790

the individual model parameters from the conjunction of our prior knowledge with the information contained in the data. The 1-D marginal posterior pdfs enable us to assess the uncertainty in the model parameters, which reflects the non-uniqueness of the non-linear inverse problem.

A visual comparison of the prior and posterior pdfs enables us to assess how well a model parameter is resolved by the data. Alternatively, one can quantify the constraint on a model parameter by comparing the information content of the prior and posterior pdfs (Tarantola and Valette, 1982), as was done by for instance Meier et al. (2007a). Quantifying the information content, or gain, is useful when quantitatively comparing the resolving power of various data types or when it is not possible to show the posterior pdfs for all model parameters. We do not include such a measure in this study, as we explicitly show the prior and posterior distributions for all 29 model parameters (Figures 3.10 and 3.11).

We use neural networks as an alternative to Monte Carlo methods. A succesful comparison of the two types of technique was presented by for instance Meier et al. (2007b); Shahraeeni and Curtis (2011). The $\sim$100,000 samples in our data set can be used to sample the posterior pdf with the straightforward Independent Metropolis-Hastings algorithm. However, we find that the data set is insufficient to produce robust results. To obtain the posterior pdf, we would need many more samples or a more sophisticated approach, such as the Neighbourhood Algorithm (Sambridge, 1999a). This illustrates the efficiency of the neural network to interpolate between the limited number of available samples.

In general, the deeper parts of the model (inner and outer core, lower mantle) are constrained well by the data and the corresponding posterior pdfs contain V_P values similar to those in *ak135* (Kennett et al., 1995). By contrast, the same data provide little information on the upper mantle V_P structure and the depth of discontinuities in the radial V_P profile. Kennett et al. (1995) derived the *ak135* model from the travel time data provided by the ISC. They used the same phases as we use here, except for *Pn*, and in addition used *PcP*, *PKKP*, *P'P' (PKPPKP)* and the converted phases *ScP* and *SP*. The inclusion of such complementary phases could have enhanced our knowledge on parts of the V_P model. These phases, however, are not included in the EHB bulletin for 2001–2008 and we decided to restrict our inversion to the phases listed in Table 3.3.

Ideally, the output for all ten input patterns in Figures 3.10 and 3.11 is similar (see Section 3.3.4). This is the case for the well-resolved parameters in the outer core and lower mantle. Differences between the posterior pdfs are larger for the inner core parameters, as they are for the P-wave velocity in the uppermost mantle (210–Moho). The trained networks are thus not completely insensitive to random variations in the input on the order of the noise level. In our view, however, these differences are minor and we argue that similar inferences would be drawn from the ten 1-D marginal posterior pdfs for these parameters.

For all parameters, the networks have been trained on the same type of travel time data as input. Obviously, any information redundancy between the individual units in the input pattern is inefficient from the perspective of network training. This is the case for the inner core parameters, for which we know only PKPdf carries information. It would thus be most efficient if the networks trained for the inner core parameters would only take PKPdf as input. We verified that prediction accuracy is similar, regardless of whether the networks are trained on the full input pattern or only on PKPdf travel times. Thus, network training is able to focus on the systematic relation between the PKPdf data and the inner core P-wave velocities.

Our final results comprise the 1-D marginal posterior pdfs $p(m_i|\mathbf{d})$ of the model parameters m_i. These represent all available knowledge on the individual model parameters, given the travel time data, associated measurement errors, our choices regarding the model parametrisation and the variations in the other model parameters. Such 1-D distributions do not contain information on any correlations between model parameters. Further, we emphasise that the maximum likelihood values of the individual model parameters, when taken together, do not necessarily represent the maximum likelihood model. Therefore, it is often desirable to analyse the joint posterior pdf, i.e. the posterior probability distribution of the full model $p(\mathbf{m}|\mathbf{d})$ (Equation 2.8). This distribution could be obtained by training a network on the complete 29-D model $\mathbf{m}$. Such a network, however, would contain a lot of free parameters and thus require a large training set. Further, network training may converge slowly or not at all for such a high-dimensional target space.

Alternatively, a joint distribution for an n-dimensional model can be constructed

from the product of conditional and marginal pdfs, e.g. Tarantola (2005)

$$\begin{aligned} p(m_1, .., m_n|\mathbf{d}) = & \; p(m_n|m_1, .., m_{n-1}, \mathbf{d}) \\ & \times \; p(m_1, .., m_{n-1}|\mathbf{d}). \end{aligned} \tag{3.2}$$

For each pdf in the r.h.s. product, a separate network would be trained. Note that it is straightforward to train networks on conditional pdfs such as $p(m_n|m_1, .., m_{n-1}, \mathbf{d})$. This only requires the input pattern to be extended with the model parameters on which the distribution is conditioned. Once successfully approximated, the joint pdf has to be sampled to analyse its properties. This requires the evaluation of the individual pdfs in the decomposition for many random model realisations and performing the necessary multiplication following Equation 3.2. This can be done efficiently, as for every trained network computing the probability for a specific input datum and model value is very fast, i.e. consumes a fraction of a second on a standard desktop computer. By doing so, one can approximate the joint posterior model distribution via Equation 3.2 and construct a representative ensemble of models. This will aid the interpretation of the information on Earth structure that is contained in seismological data.

As a simple example, we construct 2-D marginal posterior pdfs using Equation 3.2, i.e. for a 2-D $\mathbf{m}'$ (Equation 2.9), for one of the patterns in the test set. Figure 3.12 shows the three distributions in Equation 3.2 for two different combinations of parameters: V_P at the top of the inner core (m^1_{IC}) and the ICB depth d_{ICB} (first row) and V_P in the lower mantle (m^1_{LM} and m^2_{LM}, second row). The inner core V_P is resolved well, whereas the ICB depth is unresolved, as was apparent from the 1-D marginals in Figures 3.10 and 3.11. No correlation between these parameters is observed in the 2-D marginal pdf (Figure 3.12). The two shallowest V_P parameters in the lower mantle are resolved well (cf. Figure 3.10, fourth row). Despite the good constraint provided by the data, a (weak) positive correlation between the two parameters is visible in the corresponding 2-D marginal posterior pdf (Figure 3.12).

3.6 Concluding remarks

We used artifical neural networks to solve a non-linear Bayesian inverse problem. Neural networks are flexible and can be used to approximate an arbitrary function. No linearisation or model damping is required, which allows for an optimal use of the information on the model that is contained in the data. We used a Mixture Density Network to acquire a continuous probabilistic description of each model parameter. Each 1-D marginal posterior pdf represents our knowledge of the parameter and provides the necessary quantification of uncertainties, which plays a crucial role in any interpretation of seismological models.

We investigated the information on the Earth's radial P-wave velocity structure that is available in the EHB travel time data for the *Pn*, *P*, *PP*, *PKPab*, *PKPbc* and *PKPdf* phases. Our results comprise 1-D marginal posterior probabiliy distributions for the 22 V_P parameters and seven discontinuity depths in our model. These 1-D

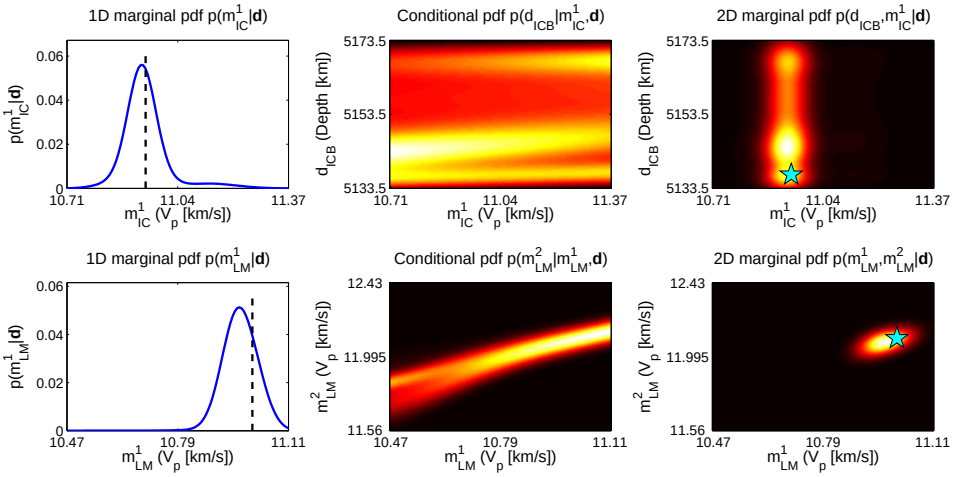


Figure 3.12: Construction of 2-D marginal posterior pdfs via Equation 3.2 for (first row) V_P at the top of the inner core (m^1_{IC}) and the ICB depth (d_{ICB}) and (second row) V_P in the lower mantle (m^1_{LM} and m^2_{LM}). See Tables 3.1 and 3.2 for the model parametrisation. The three panels in each row show (left-hand panel) the 1-D marginal pdf, (middle panel) the conditional pdf and (right-hand panel) the 2-D marginal pdf for one of the patterns in the test set. Lighter colours denote higher probabilities. The corresponding target values are denoted by the black line (left-hand panel) and the cyan star (right-hand panel).

marginal pdfs enable us to assess the uncertainty in the individual model parameters. We have shown how the method can be extended to obtain a posterior pdf for a multi-dimensional model space. This enables us to investigate potential correlations between model parameters.

The P-wave velocities in the inner core, outer core and lower mantle are resolved well, i.e. standard deviations of $\sim$0.2% to 1% with respect to the means of the 1-D marginal posterior pdfs. The maximum likelihoods of V_P are in general similar to the corresponding *ak135* values, which lie within one or two standard deviations from the means of the posterior pdfs (Table 3.5). This provides an independent validation of this part of the *ak135* model, which is often used in 3-D seismic tomography and earthquake location algorithms. Conversely, the data contain little or no information on P-wave velocity in the D" layer, the upper mantle and the homogeneous crustal layers. For the upper mantle, this is not surprising, given that the travel time data used here are of a teleseismic nature, i.e. $>25^\circ$ epicentral distance. Using additional phases available in the ISC bulletin, such as *PcP*, *PKKP* and the converted phases *SP* and *ScP*, may enhance the resolvability of our model parameters. However, the major phases we used here give a good indication of how much information on the radial V_P structure is contained in typical body-wave travel time data. We included *Pn*, which led to a weak constraint on the V_P structure in the uppermost mantle. The data do not constrain the depth of the discontinuities in our model. Again, this is common

knowledge, as teleseismic rays tend to travel perpendicular to discontinuities and thus provide a poor sampling of these structures. Reflected phases, such as *PcP*, which reflects of the CMB, are known to contain much more information on discontinuities.

Seismograms contain more information on the seismic source and the Earth's structure. We aim to apply Mixture Density Networks to Bayesian seismic waveform inversion in the future. However, for such an application the dimensionality of the data is much larger than for the travel time inversion performed in this study, which presents additional challenges that must be overcome.

Acknowledgements

We thank Malcolm Sambridge and an anonymous referee for constructive reviews. We appreciate helpful discussions with Paul Käufl and Hanneke Paulssen. Ralph de Wit and Andrew Valentine are funded by the Netherlands Organisation for Scientific Research (NWO) under the grant ALW Top-subsidy 854.10.002. Computational resources for this work were provided by the Netherlands Research Center for Integrated Solid Earth Science (ISES 3.2.5 High End Scientific Computation Resources).

4

Bayesian inversion of free oscillations for Earth's radial (an)elastic structure

Abstract

We perform a Bayesian inversion of degree-zero spheroidal mode splitting function measurements for radial (1-D) Earth structure, in terms of the Voigt averages of P-wave (V_P) and S-wave (V_S) velocities, density, bulk and shear attenuation, using neural networks. The method is flexible and allows us to assess the robustness of features in existing reference models, such as *PREM*. The Bayesian framework provides a means for quantifying uncertainties in the model parameters and for measuring the information content of the data. The analysis of the information content suggests that the free oscillations constrain most parameters better than body wave travel time data.

Our most important findings can be summarised as follows. The data prefer an inner-outer core boundary (ICB) that lies in the depth range 5154.7–5165.7 km, i.e. deeper than in existing reference models; the effect on the travel time of inner-core–sensitive seismic phases is comparable to the estimated noise in such measurements. The density contrast at the ICB (0.73 g cm^{-3}) is larger than in *PREM* (0.60 g cm^{-3})

The content of this chapter was published as: de Wit, R. W. L., P. J. Käufl, A. P. Valentine, and J. Trampert, 2014. Bayesian inversion of free oscillations for Earth's radial (an)elastic structure. *Physics of the Earth and Planetary Interiors 237*, 1–17.

and *ak135f* (0.56 g cm^{-3}), but our range including uncertainties (0.52–0.94 g cm^{-3}) encompasses all previous estimates in the literature. The average V_P and V_S in the D" region are smaller than in *PREM*, whereas the mean density is probably larger. The data cannot uniquely determine whether this density excess is restricted to the D" region or distributed throughout the lower(most) mantle. The data cannot determine with certainty the presence or absence of a discontinuity at 220 km depth for V_P, V_S and density. If present, the jump in both velocities is likely smaller than in *PREM*. Shear attenuation parameters in the mantle deviate from *PREM* in a similar fashion to results from more recent studies. We find a non-homogeneous shear attenuation in the inner core, reinforcing the hypothesis that a distinct 'innermost inner core' may exist. The bulk attenuation in the mantle and the outer core is stronger than in *PREM*. We investigate the influence of radial anisotropy on the inversions and analyse possible trade-offs between (anisotropic) parameters. The largest trade-offs are observed in regions that are believed to be anisotropic, such as the D" region. This illustrates the need to constrain anisotropy in the (deep) mantle.

4.1 Introduction

Most of our current knowledge of the Earth's internal structure has been inferred from seismological observations made at its surface. The gross features of these measurements can be explained by relatively simple spherically symmetric (1-D) models of wave velocities, density and attenuation, which describe the Earth's average (radial) structure. Such radial earth models are routinely used for the determination of seismic source locations and serve as a starting model for 3-D seismic tomography: see for example Kennett (2006) and references therein. 1-D seismological reference models are also successfully being used in conjunction with mineral physics data and geodynamical modelling to provide constraints on the Earth's thermochemical structure and its dynamics, e.g. Cammarano et al. (2005, 2011); Cobden et al. (2008, 2009).

Existing seismological reference models have been derived using seismic observables with different, yet complementary, sensitivities to the Earth's interior. The tables of Jeffreys and Bullen (1940) summarised the travel times for many different seismic phases in a 1-D earth model. The accumulation of measurements of the Earth's free oscillations made it possible to construct 1-D profiles of compressional (V_P) and shear (V_S) wave velocities and density (*1066A*, *1066B* (Gilbert, 1975)). Subsequently, parametric models were designed to simultaneously explain travel time, normal mode and regional surface wave dispersion data (*PEM*, Dziewoński et al. (1975)). A similar form of polynomial representation was used by Dziewoński and Anderson (1981) for the Preliminary Reference Earth Model (*PREM*), which was derived from body wave travel times, normal mode frequencies and attenuation measurements, augmented with constraints on the Earth's mass and moment of inertia, and has clearly outlived its 'preliminary' status. The models *iasp91* (Kennett and Engdahl, 1991) and *ak135* (Kennett et al., 1995) were constructed to explain the extensive catalogue of travel times documented by the International Seismological Centre (ISC). More re-

cently, Cammarano et al. (2005) combined seismological and mineral physics data to construct 1-D physical reference models (*PREF*). Kustowski et al. (2008) derived the spherically symmetric model *STW105*, which serves as a basis for a 3-D tomographic mantle model of anisotropic shear wave velocity (*S362ANI*). These models were derived from body wave travel times, long-period waveforms and surface wave phase anomalies.

Besides the elastic structure, the anelastic properties of the Earth have received quite some attention. This interest relates, particularly, to the temperature dependence of attenuation processes in the Earth, through which elastic (seismic) energy is transformed into heat. The Earth's absorption properties can be inferred from the attenuation of free oscillations and surface waves. For instance, *PREM* includes a model for bulk and shear attenuation, represented by their inverses Q_κ and Q_μ, respectively. Montagner and Kennett (1996) aimed to reconcile free oscillation and travel time observations by supplementing *ak135* with density and Q profiles (*ak135f*). Other estimates of the Earth's Q structure include *PAR3C* (Okal and Jo, 1990), *QM1* (Widmer et al., 1991), *QL6* (Durek and Ekström, 1996), Resovsky et al. (2005) and Cammarano and Romanowicz (2008).

Another crucial feature of seismological models concerns anisotropy in elastic parameters. Mantle flow is one mechanism that might introduce anisotropy; therefore, the identification of seismic anisotropy can provide important constraints on the Earth's dynamics. Whereas *PREM* is only transversely isotropic in the uppermost mantle, more recent studies suggest that the deeper parts of the Earth also exhibit radial anisotropy. Anisotropy has been inferred in the inner core, e.g. Morelli et al. (1986); Woodhouse et al. (1986); Beghein and Trampert (2003); Deuss et al. (2010), and the lowermost mantle, e.g. Montagner and Kennett (1996); Panning and Romanowicz (2004), while consensus appears to have been reached that the rest of the lower mantle is devoid of anisotropic structure (see Chang et al. (2014) for a review). Furthermore, the presence of anisotropy is important for estimates of the Earth's seismic structure due to trade-offs between (anisotropic) parameters in earth models. For instance, Beghein et al. (2006) investigated the robustness of radial anisotropy in existing 1-D mantle reference models and found that the strength of anisotropy in V_P trades off with density structure, which was later confirmed by Kustowski et al. (2008).

The aforementioned 1-D seismological reference models correlate with each other to a high degree (Figure 4.1), especially in the Earth's deep interior, yet there is disagreement. The fundamental shortcoming of most existing models is the lack of a quantitative assessment of their accuracy, which renders it impossible to determine the significance of the differences between these models. Seismic inverse problems are notoriously non-unique; different earth models can explain the data equally well, but may lead to incompatible interpretations of the nature of the Earth's interior and dynamics, e.g. Trampert and van der Hilst (2005); de Wit et al. (2012). Therefore, a quantification of uncertainties in any inferred earth model is essential to assess its quality and the robustness of any subsequent interpretation.

However, quantifying model uncertainties presents a challenge in traditional seismological inverse problems; consequently, most existing techniques are pragmatic

and based upon linear approximations. Resolution analyses, for instance using the framework by Backus and Gilbert (1968, 1970), can be employed to determine the robustness of the inferred earth models, e.g. Kennett (1998); Masters and Gubbins (2003). In seismic tomography, resolution and covariance matrices can provide some assessment of model quality, e.g. Aki et al. (1977); Boschi (2003); Vasco et al. (2003), but such measures are usually affected by subjective regularisation criteria. Other examples for the linear case include exploring the model null space, or model non-uniqueness (de Wit et al., 2012), misfit mapping, e.g. in the context of source parameter determination (Valentine and Trampert, 2012a) and resolution tests using matrix probing (An, 2012; Trampert et al., 2013). Kennett et al. (1995) adopted a non-linear search procedure to determine the robustness of *ak135*.

An assessment of model uncertainty is natural in a Bayesian framework, in which all inferences are probabilistic. Any inference made about a model is the result of the conjunction of our current (*prior*) knowledge and the ability of the model to explain the observations, e.g. Tarantola and Valette (1982). The *posterior* knowledge on the model, i.e. the knowledge after observing the data, represents the updated degree of belief in the model, expressed by a probability density function (pdf). Ample examples of Bayesian inference exist in the seismological literature; this involves sampling the model space, as is done in Markov Chain Monte Carlo (MCMC) methods via a (guided) random walk, e.g. Mosegaard and Tarantola (1995); Sambridge and Mosegaard (2002); Tarantola (2005).

As an alternative, we propose to employ machine learning techniques to make inferences based on samples of the prior model space. Recent examples in geophysics include Meier et al. (2007b); Shahraeeni and Curtis (2011); de Wit et al. (2013) and Käufl et al. (2014), where artificial neural networks and a set of prior samples are used to solve various geophysical inverse problems. Neural networks are very common in pattern recognition problems and can be used to infer an arbitrary non-linear mapping between two parameter spaces, e.g. Bishop (1995); MacKay (2003).

To solve the inverse problem, which in our framework involves obtaining posterior pdfs on earth model parameters, we use a Mixture Density Network (MDN, Bishop (1995)). An MDN takes the seismic data as input, and outputs the parameters describing the posterior marginal pdf for the earth model parameter(s) of interest; for a full description, see e.g. de Wit et al. (2013); Käufl et al. (2014). In the Bayesian paradigm, a 1-D marginal distribution represents our knowledge of (and uncertainties in) a single model parameter, given the variations in all other model parameters. The method is flexible, as we are free to choose the output, or *target*, parameter for the MDN. This allows us to ask specific questions, i.e. test hypotheses, about an arbitrary (combination of) model parameter(s), such as the depth of a seismic discontinuity or the average density in a region. In addition, we can construct 2-D pdfs to investigate the trade-offs between parameters, given the constraint offered by the available data.

We investigate the information on radial Earth structure that is contained in various seismic observations and assess the robustness of features in existing reference models, such as *PREM*. Our aim is threefold. First, we illustrate the flexibility of the method to investigate specific parameters in our earth model. In the process, we can

assess the uncertainties in the corresponding estimates and the information content of the data. Second, we exploit the constraint on earth model parameters provided by recently-measured normal mode splitting functions (Deuss et al., 2013; Koelemeijer et al., 2013; Koelemeijer, 2014). We perform a non-linear Bayesian inversion of the available data for radial Earth structure in terms of V_P, V_S, density, bulk and shear attenuation. We focus on specific parameters in the radial distributions rather than presenting a new radial earth model. Third, we investigate potential trade-offs between parameters in the context of radial anisotropy.

This paper is structured as follows. First, we describe the earth model parametrisation and the normal mode data. Second, we train MDNs to construct 1-D marginal posterior pdfs for the parameters in our radial earth model. Third, we address trade-offs between model parameters related to anisotropy. Finally, we discuss the efficacy of a joint inversion of normal mode and travel time data by analysing the information content of the various data sets.

4.2 Model parametrisation

We base our model parametrisation on that used for *PREM* and parametrise the radial (1-D) structure of the Earth in terms of V_P, V_S, density (ρ), the anisotropic parameter η and bulk and shear attenuation ($1/Q_\kappa$ and $1/Q_\mu$, respectively). The model is parametrised on a discrete set of 185 grid points (as one of the options in the Mineos package (Masters et al., 2011)) and the depths of discontinuities are allowed to vary. These points, or knots, are used by Mineos for a cubic spline interpolation to obtain a continuous representation with depth (between discontinuities). No correlations between physical parameters are imposed, i.e. velocity, density, η and attenuation profiles are constructed independently from each other. Within each profile, except for attenuation, we introduce correlations between adjacent points away from discontinuities to exclude physically implausible, i.e. non-smooth or oscillatory, models and restrict the size of the model space. In addition, we impose constraints on the mass and moment of inertia of the earth models using estimates from Chambat and Valette (2001). The details of the parametrisation are given in Appendix A.

We consider two different classes of parametrisation for the anisotropic structure. In a first setup, we allow for radial anisotropy in the uppermost mantle between the Moho and the 220 km discontinuity ("220"), as in *PREM*, which is parametrised by the vertically (V_{PV}, V_{SV}) and horizontally (V_{PH}, V_{SH}) polarised wave velocities and η (Dziewoński and Anderson, 1981). In a second setup, we adopt a similar anisotropic parametrisation in the whole mantle and in the inner core.

We generate 100 000 synthetic models, which are randomly drawn from the prior model distribution. Figure 4.1 shows the parameter range spanned by the prior model space and a number of existing 1-D reference models for the upper mantle. Prior ranges for the various parameters in our model are given in Tables A.2, A.3 and A.4.

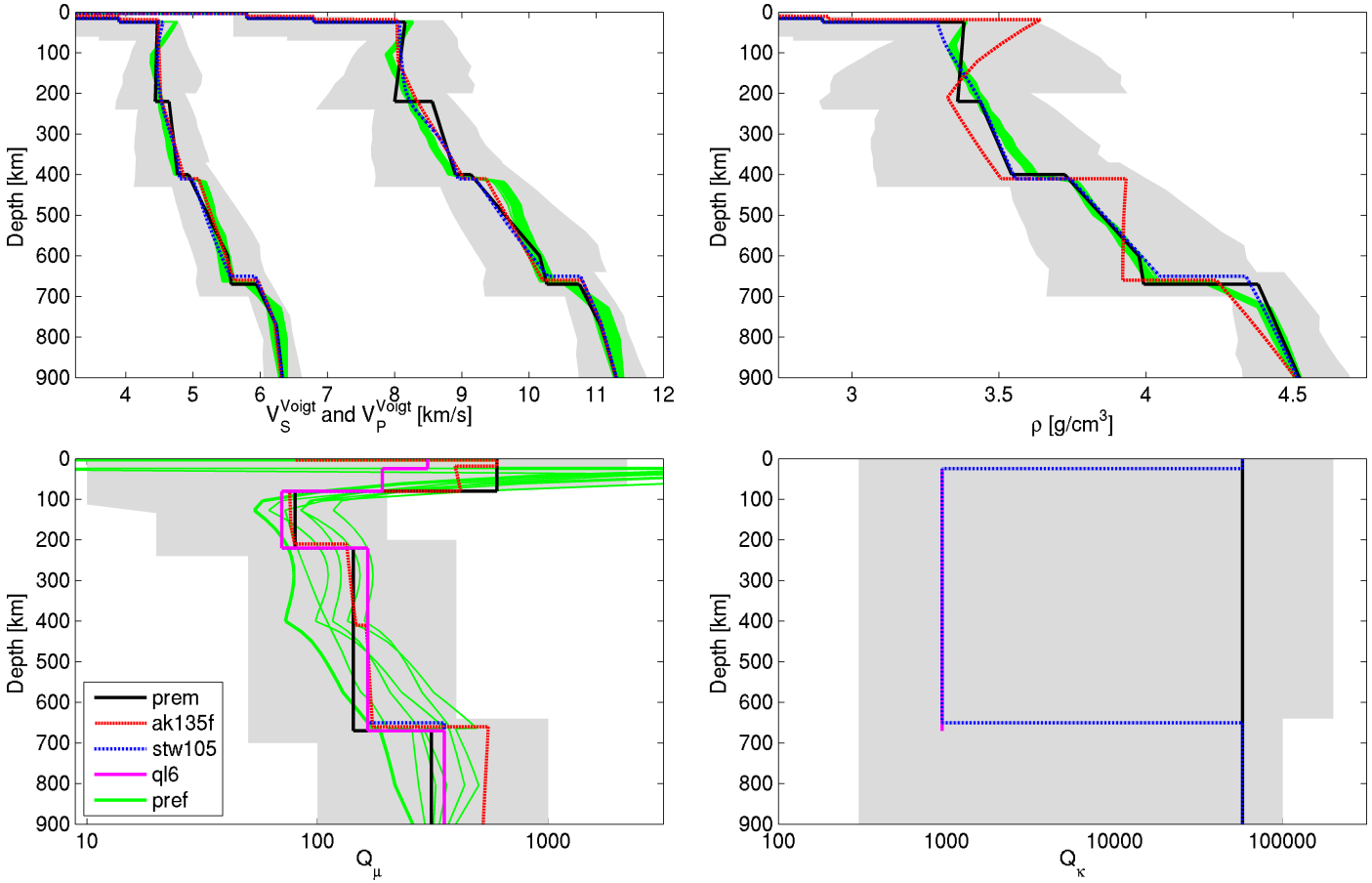


Figure 4.1: Radial earth models in the upper mantle. The parameter range spanned by the prior model space is represented by the grey shaded area, along with the 1-D reference models *PREM* (black, solid), *ak135f* (red, dashed), *STW105* (blue, dotted-dashed), *QL6* (magenta, solid) and *PREF* (Cammarano et al. (2005), 99 models, green) for V_S^{Voigt} and V_P^{Voigt} (top-left panel), ρ (top-right panel), Q_μ (bottom-left panel) and Q_κ (bottom-right panel). The horizontal scale for the bottom panels is logarithmic.

4.3 Methodology

We use artificial neural networks to solve the non-linear Bayesian inverse problem. Neural networks can approximate an arbitrary non-linear function, using a set of examples of corresponding input-output pairs. These examples are presented to a network in a so-called *training* process, during which the free parameters of a network are modified to approximate the function of interest. The particular class of neural network we use here, the MDN, takes seismological observations as input and outputs the parameters governing a conditional probability distribution (Section 2.4.4). We closely follow the methodology outlined in Chapter 2. We refer the reader to this chapter and references therein to Bishop (1995) and, for instance, Meier et al. (2007b) for details.

Neural network training is sensitive to the random initialisation of the network parameters. Therefore, it is common practice to train several neural networks with different initialisations, and subsequently choose the network which performs best on a given synthetic test data set, e.g. Bishop (1995). In Chapter 3, we trained 30 independent networks; the network which performed best on the test set was used to draw inferences from the observed data.

Here, we extend the methodology of Chapter 3 by introducing *ensembles* of MDNs, as described in Section 2.4.8, which were used by—for instance—Cornford et al. (1999) and Käufl et al. (2014). We use MDN ensembles in this chapter and in Chapters 5 to 8. Ensembles, or committees, of networks can result in better generalisation, i.e. achieve a better prediction accuracy on unseen data, e.g. Bishop (1995). The ensemble output is formed by a weighted average of the members, where the individual weights are determined by each network's performance on the same test set.

4.4 Data

4.4.1 Normal mode splitting functions

We use centre frequencies and mode quality factors derived from 185 self-coupled spheroidal mode splitting functions up to 10 mHz, the majority of which were measured by Deuss et al. (2013). This recent catalogue contains modes sensitive to V_P and inner core structure. We supplement this catalogue with similar measurements for Stoneley modes (Koelemeijer et al., 2013) and fundamental modes ${}_0S_{22} - {}_0S_{30}$ and the mode ${}_2S_{17}$ (Koelemeijer, 2014). Note that both centre frequencies and mode quality factors are only sensitive to radial (1-D) structure and thus only depend on the degree-zero splitting function coefficients c_{00}. Other splitting function coefficients are not used in our analysis, since they relate to lateral variations in Earth structure.

We use the Mineos package (Masters et al., 2011) to calculate exact normal mode frequencies and quality factors for all 100 000 synthetic 1-D earth models. Self-gravitation is taken into account for frequencies below 30 mHz and a reference period of 1 s is used for the attenuative dispersion correction. Since Mineos does not compute the

centre frequency and quality factor of mode ${}_0S_4$ for ∼2.5% of the synthetic models due to inherent computational limitations, we exclude this mode from our data set. The synthetic data for the normal modes thus consist of 184 free oscillation centre frequencies and quality factors.

We corrupt the synthetic data by adding Gaussian noise with zero mean and a standard deviation given by the uncertainty estimate accompanying each measurement (Deuss et al., 2013; Koelemeijer et al., 2013; Koelemeijer, 2014). The measurement errors were estimated using a cross-validation approach; it should be noted that this approach may not fully account for any systematic uncertainties. Further, *PREM* was used as a reference model in the iterative damped least-squares inversion of normal-mode spectra for splitting function coefficients, e.g. Deuss et al. (2013). As such, the coefficients could in theory include a bias towards *PREM*. However, Deuss et al. (2013) specify that the centre frequencies and quality factors—or the coefficient c_{00} from which these quantities are derived—are the most robust parameters in their inversion, to which no damping is applied; therefore, we assume that the bias is minimal.

4.4.2 Body wave travel times

In addition to free oscillation centre frequencies and quality factors, we used body wave travel times to perform a joint inversion of normal mode and travel time data for radial Earth structure. Similar to de Wit et al. (2013), we used first-arrival travel time data from the EHB bulletin for the years 2001 to 2008, as collected by the International Seismological Centre (ISC) and reprocessed by Engdahl et al. (1998), for the phases *Pn*, *P*, *PP*, *PKPab*, *PKPbc* and *PKIKP* (*PKPdf*). This data set was augmented by measurements from the EHB bulletin (years 2001–2008) for the *Sn* and *S* phases (Table 4.1). Synthetic first-arrival travel time curves were computed using the TauP package (Crotwell et al., 1999).

We processed the travel times following the procedure outlined in Section 3.3.4. Since the travel time curve for each phase was rather smooth, a large (linear) correlation exists between the travel time at different epicentral distances. Therefore, the travel time curves were sampled at 2° intervals. This reduced the number of free network parameters and thus made network training faster. As in de Wit et al. (2013), we assumed that this downsampling did not result in a significant loss of information on the earth model parameters, given the high correlation between the measurements. The resulting 186-D travel time vector was a concatenation of data for the Pn (8 travel time measurements), P (32), PP (62), PKPab (14), PKPbc (4), PKIKP (29), Sn (9) and S (28) phases.

The measurement errors for the centre frequencies were estimated by Deuss et al. (2013) using a cross-validation approach. This is in contrast to the conservative noise estimates in the travel time data in de Wit et al. (2013), which were based on the scatter in the available measurements in the EHB bulletin, i.e. the maximum difference between the data for each seismic phase and epicentral distance interval (Section 3.3.3). This spread originates from measurement errors, phase misidentifications, uncertain-

Table 4.1: Epicentral distance range for the seismic phases used in this study and by Kennett et al. (1995) ('Kennett95') and de Wit et al. (2013) ('DeWit2013', Section 3.3).

Distance [°]	Pn	P	PP	PKPab	PKPbc	PKIKP	Sn	S
This study	3:18	25:88	50:173	147:173	147:153	122:179	2:19	25:80
DeWit13	3:18	25:88	50:173	145:174	145:155	122:179	—	—
Kennett95	—	25:99	53:180	156:178	151:153	118:180	—	25:80

ties in the estimated source depth and lateral heterogeneities (3-D structure) in the Earth. To align the the two data sets and the associated noise estimates, we defined a new noise model for the travel time data. For a given epicentral distance, the average of the travel time data may be more representative of 1-D Earth structure, as the contribution of (incoherent) 3-D structure to the measurement scatter is averaged out. The uncertainty in this average is given by the sample variance, which we computed for the EHB travel time data for each phase and distance. For most travel time measurements, these new noise estimates are one to two orders of magnitudes smaller than the conservative error levels in de Wit et al. (2013).

4.5 Results

4.5.1 Network configuration

For all results presented in this study, we train MDNs with 40 hidden units and a Gaussian mixture consisting of 15 Gaussian kernels. The number of free parameters in an MDN N_w is given by

$$N_w = (I+1) \cdot J + (J+1) \cdot K, \tag{4.1}$$

where I, J and K are the number of input, hidden and output units, respectively (Bishop, 1995). For a 1-D target parameter and 15 Gaussian kernels, an MDN has 45 output parameters (the means, the standard deviations and the relative importance of the Gaussian kernels). In combination with a 184-D input and 40 hidden units, such an MDN has 9245 free parameters. Networks are trained using the Scaled Conjugate Gradient (SCG) algorithm (Møller, 1993) for a maximum of 5000 iterations.

As in de Wit et al. (2013), we employ early stopping, which means that network training is halted when the error of a separate validation set reaches a minimum. We use 80% of the 100 000 patterns in the synthetic data set for training, 15% for the validation set and the remaining 5% for the test set. We train an ensemble of 48 networks for each target. For each network realisation, the synthetic data are randomly divided over training, validation and test sets to enhance the generalisation capability of the ensemble.

4.5.2 Network target parameters

As we pointed out earlier, we can choose any (combination of) parameter(s) in the radial earth model as a target parameter for the MDNs. We aim at asking specific questions, i.e. we test hypotheses, rather than inferring a complete earth model that best fits the data. Our focus lies on discontinuities in the earth model; we investigate their depths and the amplitudes of the corresponding jumps in velocity and density. In particular, we address the ICB depth, the associated density contrast and the existence of the "220" discontinuity in the 1-D earth model. Further, we study the average velocities and density in the D" region and test the hypothesis of a density excess in the lower(most) mantle. In addition, we infer the mean velocities and density in the upper mantle. In the remainder of this section, we show the results for these parameters in order of decreasing depth. Finally, we investigate the bulk and shear attenuation.

Influence of radial anisotropy

We consider two different classes of radial anisotropy in our earth models (Appendix A.2): (i) radial anisotropy in the uppermost mantle (similar to *PREM*) and (ii) radial anisotropy in the whole mantle and inner core. While spheroidal modes are sensitive to all the radial anisotropy parameters, we would only expect to image them together with toroidal mode data. Therefore, we show results for the parametrisation in which the uppermost mantle is characterised by radial anisotropy (similar to *PREM*). These inferences are conditioned on the assumption that the rest of the mantle and inner core are isotropic. We use the second (fully anisotropic) parametrisation to address possible trade-offs with deeper anisotropic parameters in Section 4.6.1. Note that the outer core is set to be isotropic in both parametrisations.

4.5.3 Inferences on radial Earth structure

For networks trained on discontinuity depths, velocities, density and η, we select the centre frequencies as input. MDNs trained on attenuation use the mode quality factors as input. We evaluate the performance of each network ensemble by comparing the target value for 5000 test set samples with the Maximum A Posteriori (MAP) estimate of the ensemble output. The MAP estimate represents the parameter value that is assigned the highest probability in the posterior pdf. As an additional check, we investigate the prediction accuracy of the trained ensemble for *PREM*.

Further, we quantify the information content of the data for each target parameter by computing the Kullback-Leibler divergence D_{KL} in bits (Section 2.5) between the 1-D marginal posterior and prior pdfs, e.g. MacKay (2003). D_{KL} measures the information gain, or relative entropy, for a particular model parameter upon observing the data, e.g. Meier et al. (2007a); Käufl et al. (2014). For reference, consider a 1-D Gaussian distribution with mean μ and standard deviation σ; when comparing this distribution to one with the same mean and standard deviation $\frac{1}{2}\sigma$, the information gain is 1.16 bits. If we do not extract unique information from the data, i.e. informa-

Table 4.2: Posterior statistics for the seven discontinuity depths in kilometres, in terms of the MAP estimate θ and asymmetric 2σ model error bars, corresponding to $1/e^2$ levels in the unit normalised 1-D marginal posterior pdfs. The corresponding *PREM* and *ak135* values are given for comparison. The last column shows the information gain D_{KL} in bits (Section 2.5).

Discontinuity	*PREM*	*ak135*	θ	$\theta \pm 2\sigma$		D_{KL} [bits]
ICB	5149.5	5153.5	5160.1	5154.7	5165.7	4.7
CMB	2891.0	2891.5	2888.6	2886.5	2890.7	13.8
D" (top)	2741.0	2740.0	2722.4	2721.0	2761.0	0.0
"660"	670.0	660.0	663.7	650.3	676.2	2.5
"410"	400.0	410.0	386.2	370.0	413.8	0.7
"220"	220.0	—	200.9	200.0	240.0	0.0
Moho	24.4	35.0	36.2	20.8	47.3	3.4

tion not contained in our prior pdf, the information gain $D_{KL} = 0$. This could occur if the data have no sensitivity to a region within the Earth. In such a case, the MDN output will resemble the prior pdf (de Wit et al., 2013).

Discontinuity depths

Table 4.2 summarises the posterior statistics for the seven discontinuities by the MAP estimate θ and its asymmetric 2σ error bars, corresponding to $1/e^2$ levels in the unit normalised 1-D marginal posterior pdfs. Our analysis shows that normal mode data constrain the depths of the ICB, the core-mantle boundary (CMB), "660" and Moho, as indicated by the information gain, which is 2.5 bits or more for these parameters (Table 4.2). The depths of the top of the D" layer, the "410" and the "220" are not constrained, i.e. $D_{KL} < 1.0$ bits. As an example, Figure 4.2 shows the results for the test set, *PREM* and the observed data for the ICB. Prediction accuracy is high for the patterns in the test set and for *PREM*.

The inversion of the normal mode measurements results in clear deviations from the *PREM* reference values. This is to be expected for the "660", which is commonly found to be at a depth of 660 km in (radial) earth models but is fixed to 670 km in *PREM*, e.g. Deuss et al. (2013). For this parameter, $\theta = 663.7$ km, although *PREM* is included at the 2σ level ($\theta \pm 2\sigma = 650.3 - 676.2$ km). The ICB ($\theta = 5160.1$ km) lies significantly deeper than in *PREM* (5149.5 km) and *ak135* (5153.5 km), which are both outside the $\pm 2\sigma$ range (Figure 4.2, Table 4.2). We discuss the inference on the ICB depth in more detail in Section 4.6.3. The data indicate a strong preference for a CMB at a shallower depth ($\theta \pm 2\sigma = 2886.5 - 2890.7$ km) than in *PREM* and *ak135*. The posterior pdf for the Moho ($\theta = 36.2$ km) is in accord with the continental structure of *ak135* (35.0 km), but also includes the *PREM* value (24.4 km) at the 2σ level.

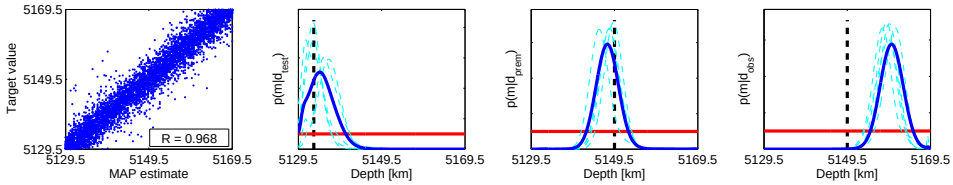


Figure 4.2: Ensemble of MDNs trained on the depth of the ICB. The left-hand panel shows the performance on the 5000 test set samples, represented by the correspondence between the MAP estimate and the target value and quantified by the correlation coefficient R. The other panels show the 1-D marginal posterior pdf (blue line), the prior pdf (red) and the target value (black, dashed) for a test set pattern (second panel from the left), *PREM* (third panel) and the observed data (right-hand panel). Note that the *PREM* value in the right-hand panel is shown as a reference and does not represent a target for the observed data. The normalised pdfs for five individual networks in the ensemble are shown in cyan.

The density contrast at the ICB

The ICB marks the deepest phase transformation in the Earth. The growth of the inner core forms a source of energy that powers the geodynamo, with the amount of energy delivered through compositional convection strongly dependent on the density contrast across the ICB, e.g. Masters and Gubbins (2003). Therefore, knowledge of the density contrast across this boundary ($\Delta\rho^{\mathrm{ICB}}$) is crucial to our understanding of the generation of the Earth's magnetic field. The density jump can be inferred from various seismic observables, but estimates of this parameter in a spherically symmetric Earth vary significantly in the literature. One approach uses normal modes that are sensitive to the inner core to infer the amplitude of the contrast, e.g. Masters and Gubbins (2003), who find 0.82 ± 0.18 g cm^{-3}. *PREM* is largely based on normal mode data and has $\Delta\rho^{\mathrm{ICB}} = 0.60$ g cm^{-3}, similar to the value in *ak135f* (0.56 g cm^{-3}, Montagner and Kennett (1996)).

A second technique considers the amplitude of the core reflected phase *PKiKP*, and its ratio with respect to the amplitude of *PcP*. Using *PKiKP*/*PcP* amplitude ratios, Cao and Romanowicz (2004) find a range of values 0.6 – 0.9 g cm^{-3} with a preferred value of 0.85 g cm^{-3}. Koper and Dombrovskaya (2005) infer a lower value of 0.52 ± 0.24 g cm^{-3}, as do Shearer and Masters (1990), who find a best fitting value of 0.55 g cm^{-3}. Approximate upper limits to the density jump, as derived from body wave studies, are 1.0 g cm^{-3} (Shearer and Masters, 1990) and 1.1 g cm^{-3} (Tkalčić et al., 2009).

Gubbins et al. (2008) point out that the two data types are sensitive to Earth structure on different length scales, which could explain discrepancies between estimates of the density contrast. Body waves are sensitive on a length scale of several kilometres, whereas the normal modes used in our study have radial wavelengths on the order of hundreds of kilometres. Gubbins et al. (2008) reconcile the different density contrasts from body wave and normal mode studies and estimates of inner core heat flux by introducing a thermochemical boundary layer at the base of the Earth's outer core.

Table 4.3: Similar to Table 4.2, but for $\Delta V_P^{\mathrm{ICB}}$, $\Delta V_S^{\mathrm{ICB}}$ and $\Delta\rho^{\mathrm{ICB}}$ at the ICB and the difference between the average V_P, V_S and density in ∼200 km wide regions above and below the ICB.

	PREM	*ak135f*	θ	$\theta \pm 2\sigma$		D_{KL} [bits]
$\Delta V_P^{\mathrm{ICB}}$	0.67	0.75	0.56	0.39	0.73	3.4
$\Delta V_P^{\mathrm{ICB\ (Wide)}}$	0.77	0.80	0.66	0.56	0.76	7.0
$\Delta V_S^{\mathrm{ICB}}$	3.50	3.50	3.47	3.43	3.56	0.3
$\Delta V_S^{\mathrm{ICB\ (Wide)}}$	3.53	3.53	3.48	3.45	3.58	0.5
$\Delta\rho^{\mathrm{ICB}}$	0.60	0.56	0.73	0.52	0.94	2.3
$\Delta\rho^{\mathrm{ICB\ (Wide)}}$	0.70	0.67	0.82	0.67	0.97	3.9

The flexibility of our method allows us to investigate the density contrast across the ICB on different length scales, although we are fundamentally limited by the resolving power (or wavelengths) of the normal modes. We train ensembles of MDNs on two different target parameters: (i) the density contrast at the boundary ($\Delta\rho^{\mathrm{ICB}}$), given by the difference in density between the two points representing the ICB in our model parametrisation, and (ii) the density contrast over a wider region spanning a few hundred kilometres. For all synthetic models and *PREM*, we calculate the average density in two ∼200 km wide regions above and below the ICB. The difference between the averages forms the new density jump ($\Delta\rho^{\mathrm{ICB\ (Wide)}}$).

For the first case, network predictions are accurate, as indicated by the performance on test set samples and *PREM* (Figure 4.3). The MAP estimate for the observed data (0.73 g cm^{-3}) is higher than in *PREM* and *ak135f*. The 2σ error levels span a range of 0.52 – 0.94 g cm^{-3} (Table 4.3). For the second target parameter, network prediction accuracy increases and the width of the posterior pdfs decreases. Again, the observed data assign most probability to values higher than in *PREM* and *ak135f*. On the 2σ level, the density jump lies in the range 0.67 – 0.97 g cm^{-3}, with an MAP estimate of 0.82 g cm^{-3}, which is comparable to the result of Masters and Gubbins (2003) (0.82 $\pm$ 0.18 g cm^{-3}) and in agreement with the upper limits of Shearer and Masters (1990) (∼1.0 g cm^{-3}) and Tkalčić et al. (2009) (∼1.1 g cm^{-3}).

Our uncertainty estimates naturally encompass the discrepancies between the estimates of earlier studies. Thus, we cannot differentiate between these estimates, given the constraint provided by the data we used here; their respective differences relate to the non-uniqueness of the inverse problem and the differences in the data used.

We also investigate the jumps in V_P and V_S across the ICB. The jump in V_P is resolved by the data, while the V_S contrast is poorly constrained ($D_{KL} < 1.0$ bits, Table 4.3). Whereas the density jump is probably larger than in *PREM* and *ak135f*, the V_P contrast is likely smaller.

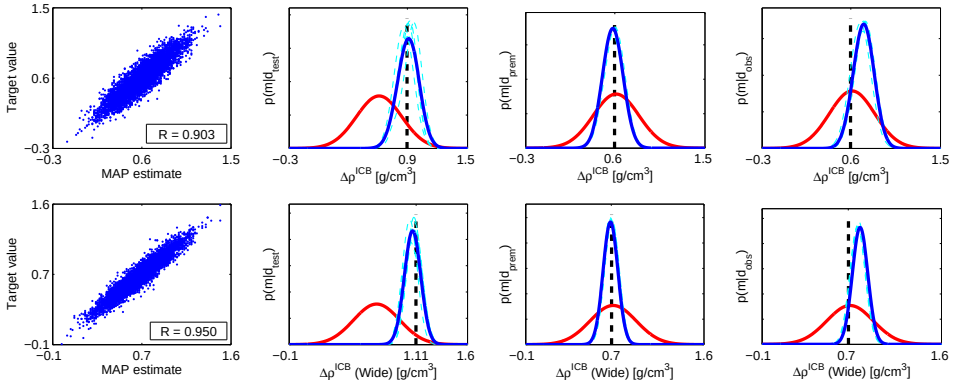


Figure 4.3: Ensembles of MDNs trained on the density contrast $\Delta\rho^{\mathrm{ICB}}$ at the ICB (first row) and the difference between the average density in ∼200 km wide regions above and below the ICB (second row). The panels are similar to those shown in Figure 4.2.

Table 4.4: Similar to Table 4.2, but for $\bar{V}_P$, $\bar{V}_S$ and $\bar{\rho}$ in the D" layer. Values are given as percentage deviations from *PREM*.

	PREM	θ [%]	$\theta \pm 2\sigma$ [%]		D_{KL} [bits]
$\bar{V}_P$ [%]	13.70 [km/s]	-0.9	-1.7	-0.0	6.9
$\bar{V}_S$ [%]	7.27 [km/s]	-1.3	-2.1	-0.5	8.1
$\bar{\rho}$ [%]	5.53 [g/cm^3]	1.5	0.4	2.5	6.0

Average velocity and density in D"

We train networks for the average V_P and V_S ($\bar{V}_P$ and $\bar{V}_S$, respectively) and density ($\bar{\rho}$) in the D" layer. The region is 150 km wide in *PREM*, but its thickness varies throughout the training data set due to the change in depth of the two enclosing discontinuities. The data provide a strong constraint on the parameters ($D_{KL} \geq 6.0$ bits, Table 4.4). The MAP estimates for $\bar{V}_P$ and $\bar{V}_S$ are respectively 0.9% and 1.3% lower than in *PREM*, which lies outside the 2σ range for both parameters. The mean density in the D" region is 1.5% higher than in *PREM*, in agreement with for instance Beghein et al. (2006), who find a mean density excess of ∼1.5% with respect to *PREM*, and *ak135f* (Montagner and Kennett, 1996), in which the average density in the D" layer deviates by 3.7% from *PREM*.

Excess density in the lowermost mantle

Kellogg et al. (1999) propose a compositionally distinct layer of ∼500 km at the bottom of the mantle and estimate that such a layer would be stable for a density excess of ∼1% (with respect to an isochemical mantle). From a linear analysis of normal mode data, Masters and Gubbins (2003) infer a possible excess of 0.4% with respect to a

Table 4.5: Similar to Table 4.2, but for $\bar{\rho}$ in three layers in the lower mantle, both including and excluding the D" region (2891–2741 km). Values are given as percentage deviations from *PREM*.

Depth [km]	*PREM* [g/cm^3]	θ [%]	$\theta \pm 2\sigma$ [%]		D_{KL} [bits]
2891–2376	5.44	0.8	0.4	1.1	10.0
2891–1792	5.29	0.6	0.3	0.9	10.7
2891– 670	5.00	0.4	0.1	0.7	10.4
2741–2376	5.40	0.2	-0.2	0.6	9.6
2741–1792	5.25	0.2	-0.1	0.6	11.0
2741– 670	4.96	0.1	-0.1	0.4	10.8

value of 5.45 g cm^{-3} (the median value of the models in their analysis), but they note that this value is within observational uncertainties. Alternatively, the whole lower mantle, ranging from the CMB to the "660", may be slightly more dense, which would imply that the excess density in the lowermost $\sim$500 km of the mantle is less than 0.4%. We investigate whether a density excess in the lower mantle could exist and, if so, whether this density excess is distributed throughout the lower mantle or whether the data can be explained by a strong density excess in the D" region and a *PREM*-like lower mantle.

First, we consider $\bar{\rho}$ in three layers of variable thickness, ranging from the CMB and upwards: (i) 2891–2376 km, (ii) 2891–1792 km, and (iii) the whole lower mantle (2891–670 km), i.e. ranging up to the "660". The uncertainties in the density estimates are relatively low for these thick layers and the data prefer a positive density anomaly (Table 4.5).

Second, we invert for $\bar{\rho}$ in three layers that exclude the D" region, i.e. layers ranging from the top of the D" region and upwards: (i) 2741–2376 km, (ii) 2741–1792 km, and (iii) 2741–670 km. For all three layers, the MAP estimates indicate a preference for a weak yet positive density anomaly (θ = 0.1% or 0.2%), but the data do not uniquely constrain the existence of a density anomaly nor its sign on the 2σ level (Table 4.5).

Thus, with the data used here we cannot uniquely determine whether a density excess with respect to *PREM* is distributed throughout the lower mantle. If we compare the pdfs for the target parameters including and excluding the D" region, the data clearly prefer a density excess in the D" layer, as was evident from Table 4.4.

The upper mantle

We train networks for $\bar{V}_P$, $\bar{V}_S$ and $\bar{\rho}$ in three layers in the upper mantle: (i) the transition zone (TZ), which is bordered by the "660" and "410" discontinuities, (ii) the region between the "410" and the "220" ("410-220") and (iii) the uppermost mantle between the "220" and the Moho ("220-Moho"). All nine parameters are resolved by the data, although to varying degrees (Table 4.6). In particular, the data contain much

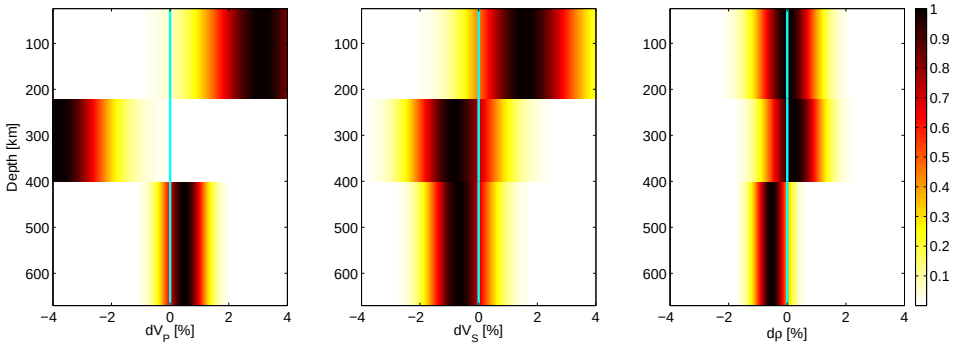


Figure 4.4: 1-D marginal posterior pdfs for $\bar{V}_P^{Voigt}$ (left-hand panel), $\bar{V}_S^{Voigt}$ (middle panel) and $\bar{\rho}$ (right-hand panel) in the upper mantle, expressed as percentage deviations with respect to *PREM*. The probability for each 1-D pdf is rescaled so that the maximum equals 1. Asymmetric 1σ and 2σ error bars correspond to the $1/e^{1/2}$ (0.61) and $1/e^2$ (0.14) contours, respectively.

Table 4.6: Information gain D_{KL} in bits for $\bar{V}_P$, $\bar{V}_P$ and $\bar{\rho}$ in the TZ, the "410-220" region and the "220-Moho" region (Figure 4.4).

	TZ	"410-220"	"220-Moho"
$\bar{V}_P$	8.2	3.5	5.5
$\bar{V}_S$	4.7	7.3	3.4
$\bar{\rho}$	9.7	7.4	7.5

information on density ($D_{KL} \geq 7.4$) and the parameters in the TZ ($D_{KL} \geq 4.7$).

The TZ is 270 km wide in *PREM*, but its thickness varies throughout the training data set with the change in depth of the two enclosing discontinuities. The MAP estimates for the mean $\bar{V}_P$ and $\bar{V}_S$ deviate respectively 0.5 and -0.7% from *PREM*, but the data do not uniquely determine the existence of a deviation from *PREM* in either velocity on the 2σ level (Figure 4.4). For the average TZ density, most of the probability mass in the pdf lies at values lower than in *PREM*, with a most probable estimate of -0.5%, in contrast to a positive density anomaly found by Beghein et al. (2006). However, we note that the posterior pdf inferred by Beghein et al. (2006) is relatively wide, i.e. extends down to -2% with respect to *PREM* on the 2σ level; our uncertainty estimate is narrower and falls within this range. For the "410-220" region, we find a negative deviation from *PREM* for V_P ($\theta \pm 2\sigma$ = -6.0% – -1.4%). A possible deviation from *PREM* is not constrained on the 2σ level for $\bar{V}_S$ and $\bar{\rho}$.

The uppermost mantle, which is enclosed by the "220" and the Moho, is radially anisotropic in our parametrisation. Therefore, we analyse the Voigt averages of V_P and V_S, the density and the anisotropic parameter η. The Voigt averages can be interpreted as the isotropic representation of the wave velocities in an anisotropic medium and can be approximated by $V_P^{Voigt} = \sqrt{(V_{PV}^2 + 4V_{PH}^2)/5}$ and $V_S^{Voigt} =$

Table 4.7: Similar to Table 4.2, but for the jumps in V_P, V_S and ρ across the "660" and "410" discontinuities.

	PREM	*ak135f*	θ	$\theta \pm 2\sigma$		D_{KL} [bits]
ΔV_P^{660} [km/s]	0.49	0.59	0.52	0.24	0.79	2.1
ΔV_S^{660} [km/s]	0.37	0.35	0.40	0.24	0.55	2.0
$\Delta\rho^{660}$ [g/cm^3]	0.39	0.31	0.43	0.34	0.51	4.5
ΔV_P^{410} [km/s]	0.23	0.33	0.60	0.23	0.95	2.0
ΔV_S^{410} [km/s]	0.16	0.21	0.17	-0.03	0.38	3.0
$\Delta\rho^{410}$ [g/cm^3]	0.18	0.42	0.14	0.02	0.26	6.1

$\sqrt{(2V_{SV}^2 + V_{SH}^2)/3}$ for V_P and V_S, respectively, e.g. (Panning and Romanowicz, 2006). Note that this approximation is valid under the assumption of small anisotropy, i.e. $\eta \approx 1$, which is valid for the anisotropy considered here. The region is 195.6 km wide in *PREM*, but its thickness varies throughout the training data set with the change in depth of the two enclosing discontinuities. For the observed data, the MAP estimates for $\bar{V}_P^{Voigt}$ and $\bar{V}_S^{Voigt}$ are respectively 3.1% and 1.6% higher than in *PREM*, which lies outside the $\pm 2\sigma$ range for $\bar{V}_P^{Voigt}$ (Figure 4.4). The higher V_P^{Voigt} and V_S^{Voigt} are in agreement with reference models such as *ak135* and *STW105*, which compensate for the absence of a discontinuity at 220 km with velocities in the overlying region that are higher than in *PREM* (Figure 4.1). We will investigate the relation with the "220" further in Section 4.5.3. The average density in this region is in agreement with *PREM*. This disagrees with a likely negative density anomaly inferred by Beghein et al. (2006), but again we note that their pdf is more conservative, i.e. represents a larger uncertainty, than the 1-D marginal pdf we obtained here. The average anisotropic parameter $\bar{\eta}$ in the "220–Moho" region is weakly constrained by the data ($D_{KL} = 1.0$ bits), but the $\pm 2\sigma$ error levels span a wide range of 0.91–0.99 ($\bar{\eta} = 0.94$ in *PREM*).

Further, we train ensembles of MDNs on the jumps in V_P, V_S and density across the "660" and "410" discontinuities. For the "660", the posterior pdfs for V_P and V_S are centred on *PREM* (Table 4.7). The density jump is resolved best ($D_{KL} = 4.5$ bits) and tends towards slightly higher values. This is in agreement with our finding that the average density in the TZ may be slightly lower than in *PREM* (Figure 4.4). For the "410", the jumps in V_S and density are in accord with *PREM*. The MAP estimate for the V_P contrast is relatively large ($\theta = 0.60$ km/s), although *PREM* (0.23 km/s) lies just within our uncertainty bounds.

Existence of a discontinuity at 220 km

The hotly-debated Lehmann discontinuity at 220 km depth is another good target for our flexible method. The discontinuity has been observed by many workers using several distinct data types (see e.g. Deuss et al. (2013) for a review). Consensus

Table 4.8: Similar to Table 4.2, but for $\Delta V_{P\ (Voigt)}^{220}$, $\Delta V_{S\ (Voigt)}^{220}$ and $\Delta\rho^{220}$.

	PREM	θ	$\theta \pm 2\sigma$		D_{KL} [bits]
$\Delta V_{P\ (Voigt)}^{220}$ [km/s]	0.56	0.10	-0.44	0.61	1.1
$\Delta V_{S\ (Voigt)}^{220}$ [km/s]	0.20	0.07	-0.17	0.31	1.6
$\Delta\rho^{220}$ [g/cm^3]	0.08	0.09	-0.09	0.27	3.4

seems to have been reached on its regional existence, but its nature and whether the discontinuity extends globally are still debated. The controversy is illustrated by the presence, e.g. *PREM*, and absence, e.g. *ak135*, of the discontinuity in existing reference earth models.

We construct target parameters similar in fashion to those used to investigate the velocity and density contrasts at the "660" and "410" and invert for the jumps in V_P^{Voigt}, V_S^{Voigt} and density at the "220". Our aim is to investigate the probability that the contrasts at the "220" in these parameters is positive, as they are in *PREM*. We show a test set pattern in Figure 4.5 with a near-zero jump, which we interpret as the absence of a discontinuity, to demonstrate the ability of the trained network ensemble to make correct predictions for such an earth model.

We find that the data prefer contrasts in V_P^{Voigt} ($\theta = 0.10$ km/s) and V_S^{Voigt} ($\theta = 0.07$ km/s) at the "220" that are smaller than in *PREM* (Table 4.8). However, the information gain is relatively low ($D_{KL} \leq 1.6$ bits) and the 1-D posterior pdfs include both zero (no jump) and *PREM*-like values (Figure 4.5). The smaller jump in V_P^{Voigt} is in agreement with the higher $\bar{V}_P^{Voigt}$ in the "220-Moho" region and the lower $\bar{V}_P$ in the "410-220" region (Figure 4.4). Similarly, the smaller contrast in V_S^{Voigt} corresponds to a higher $\bar{V}_S^{Voigt}$ in the "220-Moho" region. By contrast, the density contrast is similar to *PREM*, although the data do not exclude a (near-) zero amplitude on the 2σ level. The probability of a positive jump can be extracted from the marginal pdfs and is 0.63 ($\Delta V_{P\ (Voigt)}^{220}$), 0.72 ($\Delta V_{S\ (Voigt)}^{220}$) and 0.84 ($\Delta\rho^{220}$). Thus, the normal mode data cannot determine with certainty the presence or absence of a discontinuity at 220 km in the radial Earth structure, particularly for V_P^{Voigt} and V_S^{Voigt}. If a discontinuity exists, the jump in density and both velocities is probably small ($\sim$0.1 g/cm^3, $\sim$0.1 km/s).

Attenuation

We train MDNs that take the normal mode Q measurements as input and produce bulk and shear attenuation parameters, $1/Q_\kappa$ and $1/Q_\mu$, respectively, as output. The 13 Q_κ and Q_μ parameters are independently drawn from prior distributions that are uniform on a logarithmic scale (A.3). Figure 4.6 shows the 1-D posterior pdfs for the 13 Q_μ and Q_κ parameters, as well as the 1-D earth models *PREM*, *ak135f* (Montagner and Kennett, 1996) and *QL6* (Durek and Ekström, 1996). In addition, we show the most probable values obtained from a sampling-based inversion of surface wave and

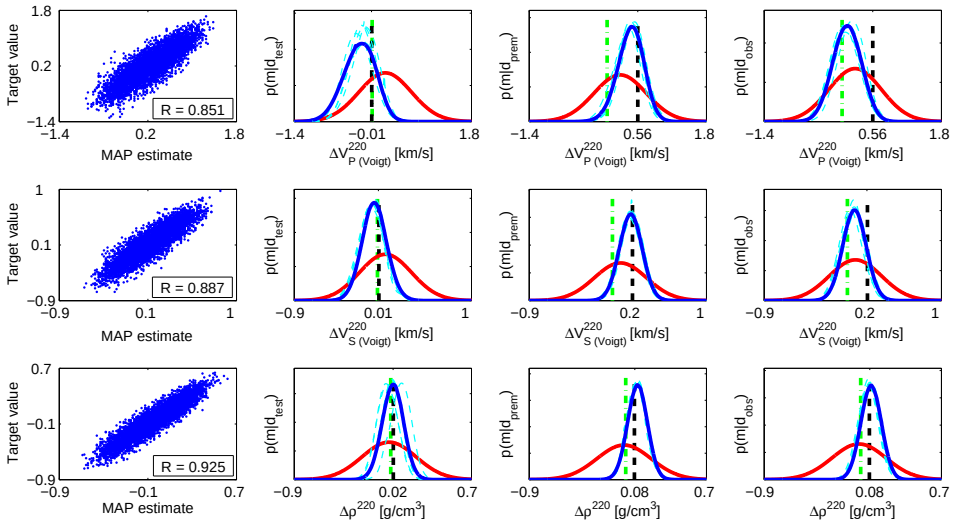


Figure 4.5: Ensembles of MDNs trained on $\Delta V_{P\ (Voigt)}^{220}$ (first row), $\Delta V_{S\ (Voigt)}^{220}$ (second row) and $\Delta\rho^{220}$ (third row). The green dashed-dotted line indicates no jump, such as in *ak135*. The panels are similar to those shown in Figure 4.2.

normal mode attenuation measurements (Resovsky et al., 2005). Note that our inferences and the mentioned 1-D earth models are based on the assumption of frequency-independent attenuation, despite the consensus that $Q_\mu \propto \omega^\alpha$ with α between 0.1 and 0.5 (e.g. Romanowicz and Mitchell (2007)), which may result in a bias in estimates of Q_μ (Lekić et al., 2009).

The information content is $D_{KL} \geq 3.1$ bits for all Q_μ parameters (Table 4.9). This indicates that the data are sensitive to the attenuation structure in both inner core and mantle, given the particular model parametrisation we haven chosen. For most Q_μ parameters, we find that our MAP estimates deviate from *PREM* in a similar fashion to the most probable values reported by Resovsky et al. (2005). Notable exceptions are the high-velocity lid and crust, for which we find a lower Q_μ, and the middle layer in the lower mantle, for which the data prefer a higher Q_μ that is similar to the value in *ak135f*.

In comparison to the 1-D models shown in Figure 4.6, we parametrised the inner core shear attenuation using an additional layer. The shear Q in the upper half of the inner core is in agreement with existing models ($\theta = 89$), but we observe a preference for a lower shear Q in the lower half of the inner core ($\theta = 43$). We estimate the compressional attenuation using these values for Q_μ, our MAP estimate for Q_κ of 2243, $V_P = 11.1$ km s^{-1}, $V_S = 3.6$ km s^{-1} and the relation (Anderson and Hart, 1978)

$$Q_\alpha^{-1} = \frac{4}{3}\left(\frac{V_S}{V_P}\right)^2 Q_\mu^{-1} + \left[1 - \frac{4}{3}\left(\frac{V_S}{V_P}\right)^2\right] Q_\kappa^{-1}. \tag{4.2}$$

We find $Q_\alpha \approx 510$ and $Q_\alpha \approx 274$ in the upper and lower half of the inner core, respectively. This contrasts with the consensus in the literature that both compressional and shear attenuation decrease with depth in the inner core (see Romanowicz and Mitchell (2007) for a review). Andrews et al. (2006) showed that measurements of inner core Q could be biased by neglecting mode coupling. This could influence the data we use here, which only contains self-coupled modes, and requires further investigation.

The existence of an innermost inner core, with approximate radius between 300 and 600 km, has been suggested previously based on studies of anisotropy, e.g. Ishii and Dziewoński (2002); Beghein and Trampert (2003), and attenuation, e.g. Li and Cormier (2002); Cormier and Stroujkova (2005), in the inner core, although Lythgoe et al. (2014) find that an innermost inner core is not required to explain *PKIKP* travel times. A review of studies on the inner core structure and dynamics can be found in e.g. Alboussière and Deguen (2012); Deguen (2012). Admittedly, we cannot address the structure of the inner core in detail within the limitations of our two-layer parametrisation. However, we verified that the MDNs make accurate predictions for *PREM* and test set models with similar Q_μ values in the two inner-core layers. By contrast, the normal mode data used here strongly prefer a non-homogeneous structure.

The data contain less information on Q_κ, most notably for the bulk attenuation in the inner core and upper mantle (Table 4.9). Our results indicate that Q_κ in the mantle and the outer core is lower than in *PREM*. In the upper mantle, our MAP estimate for Q_κ ($\theta = 1338$) is similar to model *QL6*. In contrast to Resovsky et al. (2005), we find a relatively low Q_κ in the inner core ($\theta = 2243$) that is similar to *PREM*. However, we note that the posterior pdf is strongly asymmetric and spans several orders of magnitude ($\theta + 2\sigma = 195\,349$, Figure 4.6).

4.6 Discussion

4.6.1 Trade-offs with anisotropy

Since spheroidal mode data alone do not fully constrain radial anisotropy, we showed results for the *PREM*-like parametrisation, in which only the uppermost mantle is characterised by radial anisotropy. However, it is instructive to investigate whether trade-offs between (anisotropic) parameters exist, in light of the information contained in the spheroidal modes. To this end, we train MDNs on the same target parameters as in Section 4.5, but for the second parametrisation with an anisotropic inner core and mantle, and compare the resulting 1-D posterior pdfs with those for the isotropic parametrisation (Figures 4.7 and 4.8). To quantify their similarity, we compute the percentage of overlap between the pdfs; a low overlap indicates that trade-offs related to potential anisotropy influence the result of the inversion. The discrepancies between the pdfs illustrate where we need to constrain radial anisotropy in the model. As an example, we discuss the average velocities and density in the D" region and the CMB depth and analyse possible trade-offs between anisotropic

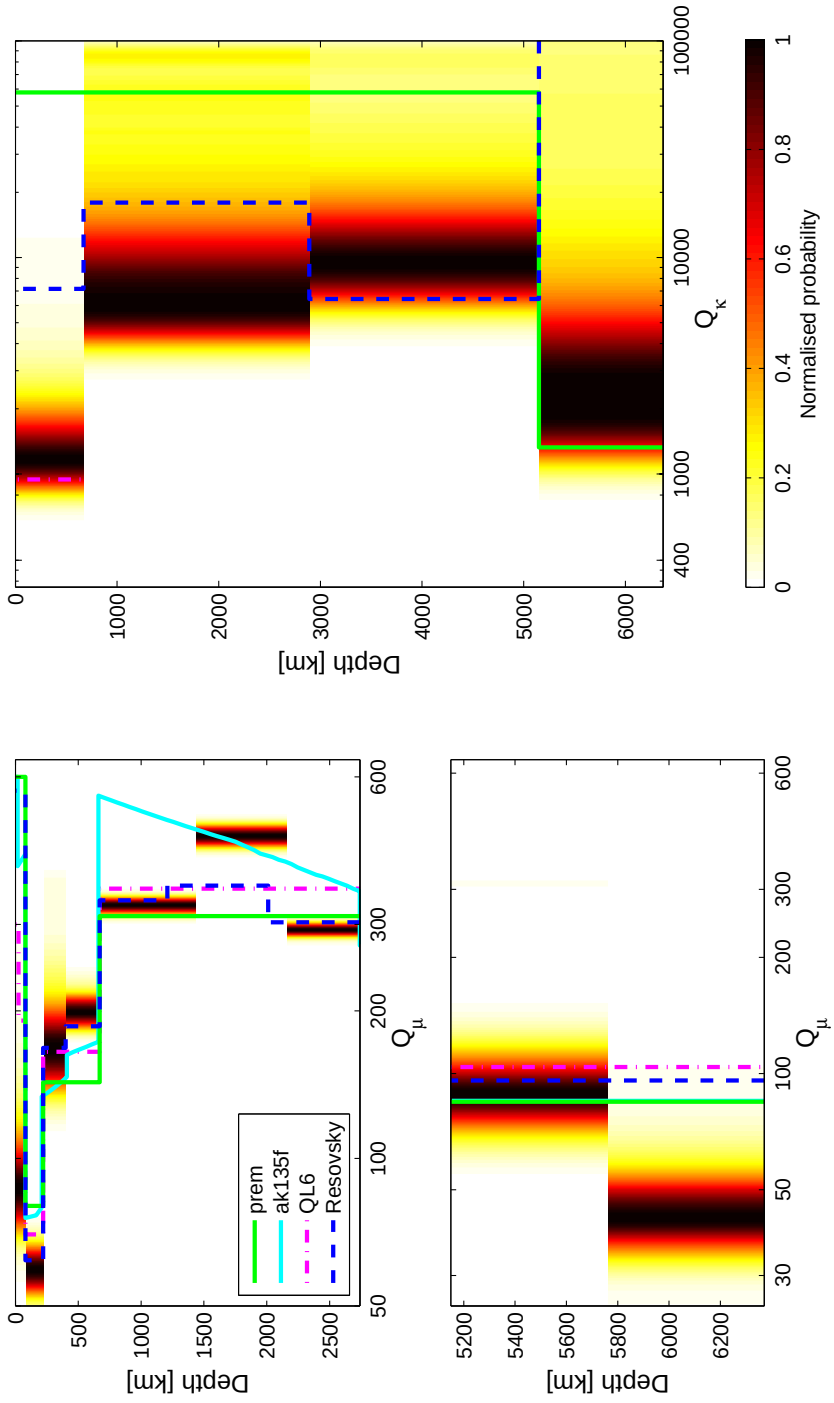


Figure 4.6: 1-D marginal posterior pdfs for Q_μ (left-hand panels) and Q_κ (right-hand panel). The probability for each 1-D pdf is rescaled so that the maximum equals 1. 1σ and 2σ error levels correspond to the $1/e^{1/2}$ (0.61) and $1/e^2$ (0.14) contours, respectively. Several 1-D earth models and the most probable values from Resovsky et al. (2005) are added for comparison.

Table 4.9: Information gain D_{KL} in bits for the 13 bulk and shear attenuation parameters.

Q_μ	Region	D_{KL} [bits]
	Inner core (IC)	
	6371–5760 km	3.4
	5760–5150 km	8.3
	Outer core (OC)	—
	Lower mantle (LM)	
	2891–2157 km	13.2
	2157–1428 km	12.3
	1428– 670 km	13.1
	Transition zone (TZ)	11.5
	410–220	6.3
	Low-velocity zone (LVZ)	9.4
	High-velocity lid + crust	5.1
Q_κ	Region	D_{KL} [bits]
	Inner core (IC)	2.2
	Outer core (OC)	6.4
	Lower mantle (LM)	6.3
	Upper mantle (UM)	2.8

parameters.

For the D" region, we find a strong discrepancy between the pdfs for the two parametrisations (Figure 4.7). Additional trade-offs result in larger uncertainties and hence a lower information gain, especially for the density (Table 4.10). This may not be surprising, as the D" region is in general believed to be (radially) anisotropic, at least in terms of V_P and the anisotropic parameter η, e.g. Montagner and Kennett (1996); Beghein et al. (2006). The sign of a possible deviation from *PREM* is not constrained on the 2σ level for both velocities and density (red lines, Figure 4.7). Further, we observe that the MAP estimates for $\bar{V}_P^{Voigt}$ and $\bar{\rho}$ are opposite in sign compared to the results for the isotropic parametrisation.

We further investigate whether the weaker constraint is the result of trade-offs related to anisotropic parameters. Therefore, we construct 2-D marginal pdfs for two model parameters m_1 and m_2 using the decomposition, e.g. Tarantola (2005),

$$\sigma(m_1, m_2|\mathbf{d}) = \sigma(m_2|m_1, \mathbf{d})\sigma(m_1|\mathbf{d}), \tag{4.3}$$

where the l.h.s. 2-D marginal pdf is given by the product of the 1-D marginal pdf $\sigma(m_1|\mathbf{d})$ and the conditional pdf $\sigma(m_2|m_1, \mathbf{d})$, i.e. the pdf for m_2 conditioned on m_1. All pdfs in Equation 4.3 are conditioned on the observed data $\mathbf{d}$. For both pdfs in

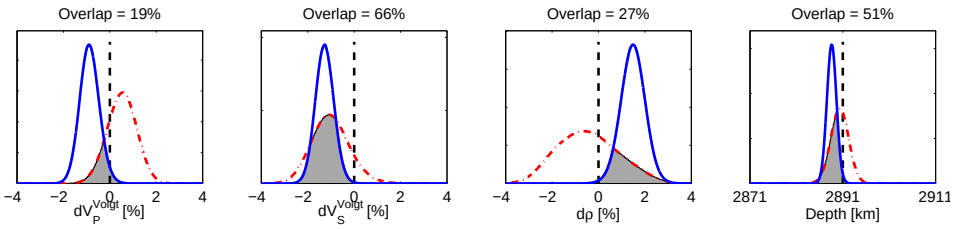


Figure 4.7: 1-D marginal posterior pdfs for the partly (blue, solid) and fully (red, dashed-dotted) anisotropic model parametrisation for $\bar{V}_P^{Voigt}$ (left-hand panel), $\bar{V}_S^{Voigt}$ (middle-left panel), $\bar{\rho}$ (middle-right panel) in the D" region and the CMB depth (right-hand panel). *PREM* is shown as a reference (black, dashed) and the percentage of overlap between the two pdfs is shown above each panel.

Table 4.10: Information gain D_{KL} in bits for $\bar{V}_P^{Voigt}$, $\bar{V}_S^{Voigt}$ and $\bar{\rho}$ in the D" region and the CMB depth for the partly and fully anisotropic model parametrisation (Figure 4.7).

	Anisotropy	
Parameter	Partly	Fully
$\bar{V}_P^{Voigt}$	6.9	2.3
$\bar{V}_S^{Voigt}$	8.1	1.4
$\bar{\rho}$	6.0	0.3
CMB	13.8	8.7

the r.h.s. product, a separate MDN is trained. Note that it is straightforward to train MDNs on conditional pdfs; this merely requires the conditional model parameter(s) to be appended to the input pattern (Section 3.5).

Figure 4.9 shows an example of 2-D pdfs for $\bar{\eta}$ versus both $\bar{V}_P^{Voigt}$ and $\bar{V}_S^{Voigt}$ in the D" region. $\bar{\eta}$ is unresolved by the data, as is evident from the 1-D marginal pdf (left-hand panels). In both cases, the conditional and the 2-D marginal pdfs show a clear correlation between $\bar{\eta}$ and the velocities. The trade-offs in the 2-D pdfs indicate that the data do not uniquely constrain the individual parameters. If we impose isotropy in the D" region *a priori*, i.e. $\eta = 1$, we would infer a different $\bar{V}_P^{Voigt}$ and $\bar{V}_S^{Voigt}$ than if anisotropy is present.

Constraining radial anisotropy

For some of our model parameters, we do not obtain a similar pdf for the two parametrisations (Figures 4.7 and 4.8), which highlights the need to constrain radial anisotropy in the whole earth model. We can potentially resolve these (anisotropic) parameters better, and thereby reduce trade-offs between parameters, if we augment the spheroidal mode data with complementary measurements, such as toroidal modes

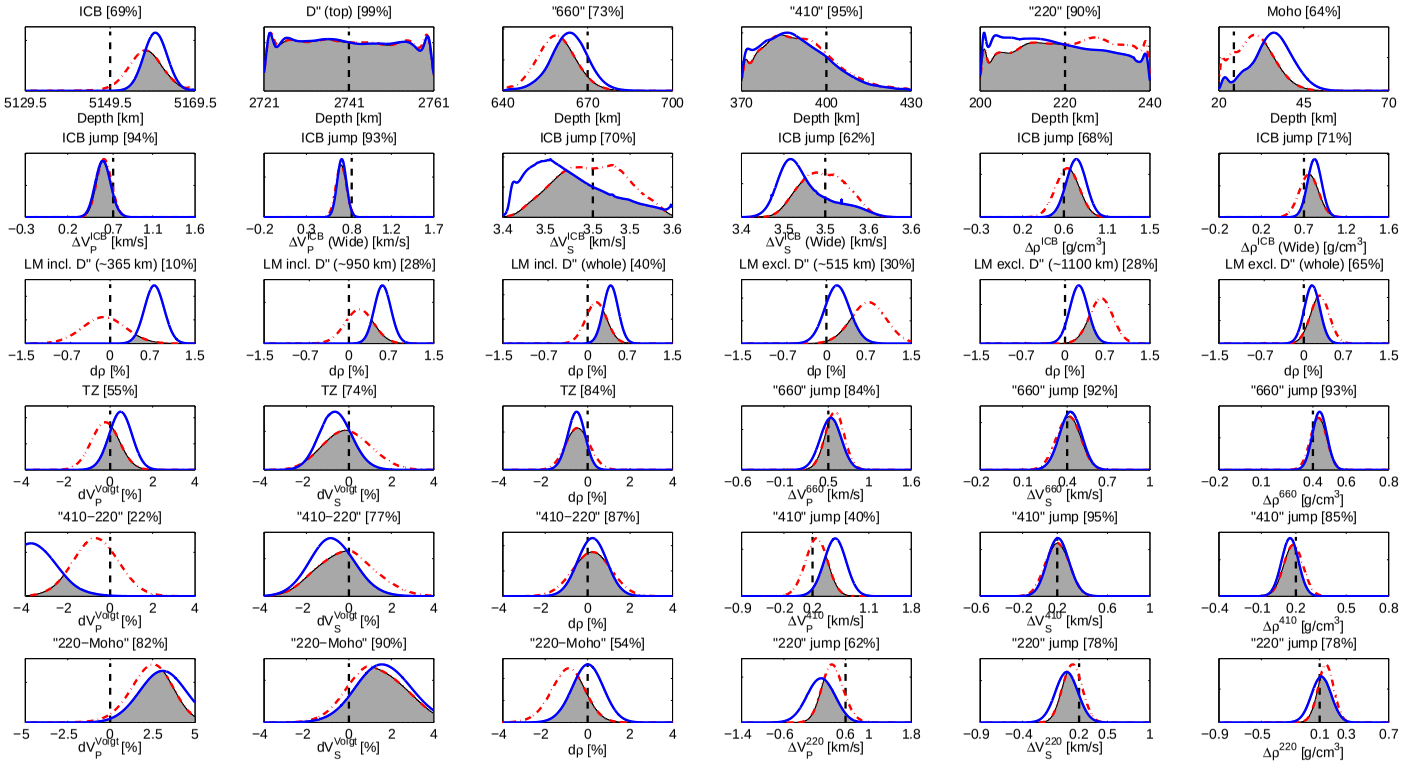


Figure 4.8: 1-D marginal posterior pdfs for the partly (blue, solid) and fully (red, dashed-dotted)) anisotropic model parametrisation for all parameters addressed in the main text, except for parameters related to D'' (Figure 4.7). *PREM* is shown as a reference (black, dashed) and the percentage of overlap between the two pdfs is shown above each panel. Note that the outer core is isotropic in both parametrisations.

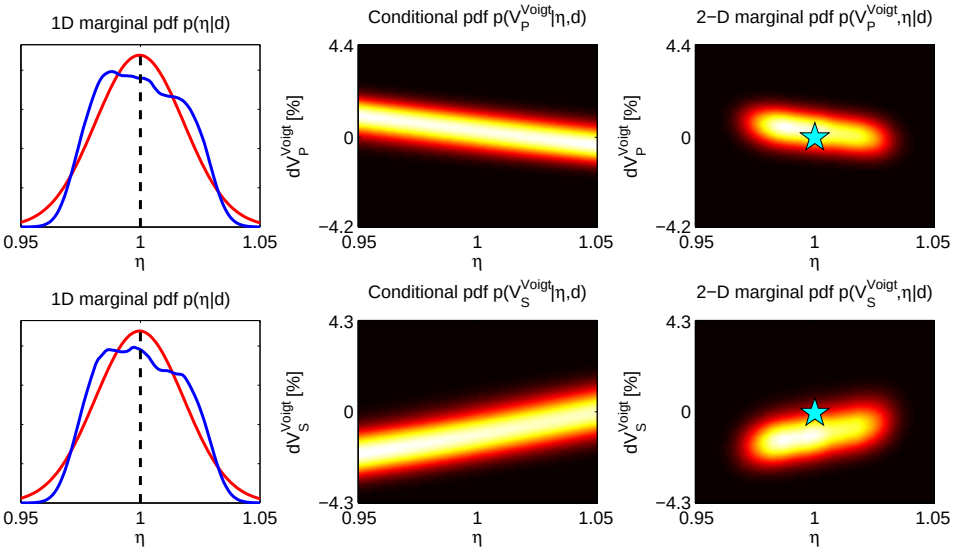


Figure 4.9: Construction of 2-D marginal posterior pdfs via Equation 4.3 for $\bar{V}_P^{Voigt}$ (top row) and $\bar{V}_S^{Voigt}$ (bottom row) versus $\bar{\eta}$ in the D'' region. The three panels in each row show the 1-D marginal (blue) and prior (red) pdfs (left-hand panel), the conditional pdf of the velocity given $\bar{\eta}$ and the observed data **d** (middle panel) and the 2-D marginal pdf of $\bar{\eta}$ and the average velocity for the observed data **d** (right-hand panel). Lighter colours denote higher probabilities. The corresponding *PREM* values are denoted by the black line (left-hand panel) and the cyan star (right-hand panel).

and surface waves. We did not use such data here and limited our inversion to spheroidal mode data. However, the input to a neural network can easily be extended by such additional data; the same flexibility applies as we have already highlighted for the target parameters.

Finally, we note that our radially anisotropic parametrisation in the inner core ignores a part of the existing knowledge of this region. This choice was deliberate, since we used centre frequencies and quality factors that are only sensitive to the radial (degree-zero) structure. Inner core anisotropy can be better described by a cylindrical symmetry with the symmetry axis aligned with the Earth's rotation axis, e.g. Morelli et al. (1986); Woodhouse et al. (1986). However, such cylindrical symmetry manifests itself in splitting function coefficients of higher degrees, which we did not use in this work. We checked whether the assumption of radial anisotropy in the inner core biased the results of the second (anisotropic) parametrisation by constructing a new training set with a radially anisotropic mantle and an isotropic inner (and outer) core. We trained MDNs on this new training set and found that the resulting pdfs were similar, with overlap $\geq 87\%$, to the pdfs for the parametrisation with a radially anisotropic inner core (not shown here). Therefore, we concluded that our analysis of trade-offs, as presented above, was not biased by either an isotropic or a radially anisotropic parametrisation in the inner core.

4.6.2 Joint inversion with travel time data

All results shown in this study are based on normal mode centre frequencies and quality factors. We also performed joint inversions of normal mode data and body wave travel times. For the latter data type, we adopted a setup similar to de Wit et al. (2013), who used first-arrival travel time data from the EHB bulletin (Engdahl et al., 1998) for the phases *Pn*, *P*, *PP*, *PKP* for the years 2001 to 2008. We supplemented this data set with data for the *Sn* and *S* phases (Section 4.4.2). For the joint inversion, the input to the network consisted of 184 free oscillations and 186 travel time measurements. The 186-D travel time vector was a concatenation of data for the Pn (8 travel time measurements), P (32), PP (62), PKPab (14), PKPbc (4), PKIKP (29), Sn (9) and S (28) phases.

However, we analysed the information content of the data and found that the travel time measurements do not provide additional information on Earth structure to the inversion of the normal mode data for the chosen parametrisation. We computed the information gain D_{KL} for an inversion of the normal mode data, an inversion of the travel time data and a joint inversion of the two data sets. As an example, Figure 4.10 shows the information gain for the three different inversions for the depth of four discontinuities. With the exception of the Moho depth, and the velocities near this discontinuity, we found that the travel time data have a low information content compared to the normal mode data for most model parameters. Consequently, the joint inversion did not yield a better constraint on the model parameters than a separate inversion of the normal modes. We performed this analysis for a synthetic data set in which the bulk and shear attenuation structures were fixed to that of *PREM*, as body wave travel times have very little sensitivity to the attenuation parameters.

Based on this analysis, we decided to focus on the results without travel time data. By removing the travel times, we reduced the size of the (input of the) neural network, the required number of training patterns and thus computation time, whilst retaining the information on the model parameters. This does not mean that the information in travel times is never complementary to the information contained in the normal mode data; their relative contributions to our inversions relate to the choices we have made in our setup. A joint inversion of normal modes and travel times may be beneficial if one is able to construct a more complete noise model for the travel times and include additional seismic phases. In addition, more information on Earth structure may be available when using travel time picks of higher quality than the data in the ISC (and EHB) catalogues, as these bulletins contain measurements of varying precision.

4.6.3 Shift in ICB depth

We evaluated the ICB depth for the fully anisotropic parametrisation (Figure 4.8) and found that the pdfs for the two parametrisation overlap by 69%. The pdf for the anisotropic case is wider and corresponds more to *PREM* (5149.5 km) and *ak135* (5153.5 km), but most of the probability mass is still assigned to greater depths ($\theta =$ 5157.8 km and $\theta \pm 2\sigma = 5149.9 - 5165.7$ km). For the isotropic case, our MAP esti-

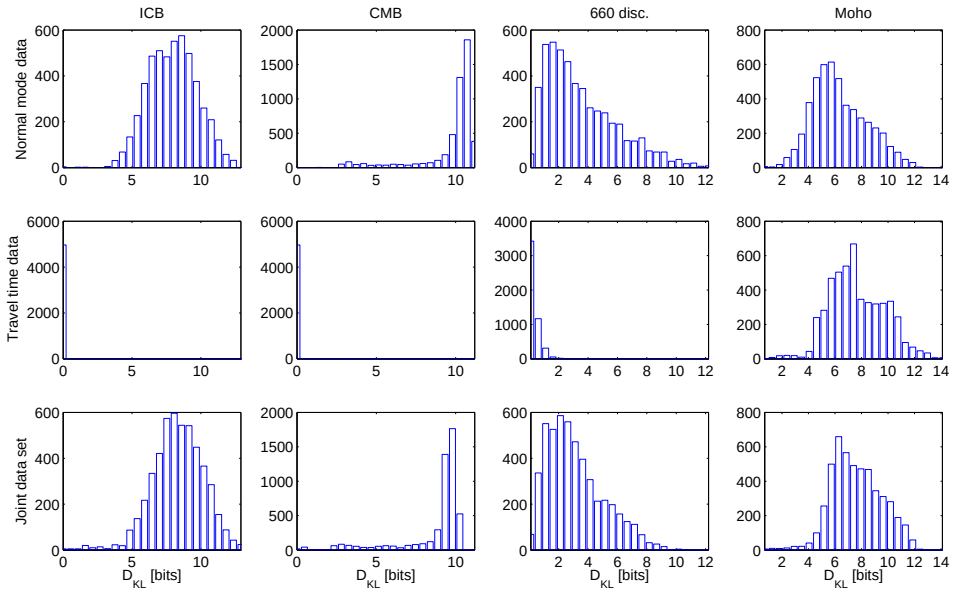


Figure 4.10: Information gain, as measured by the Kullback-Leibler divergence D_{KL} (Equation 2.33), for networks trained on the normal mode data (first row), the travel time data (second row) and the joint data set (third row). D_{KL} (in bits) is shown for the depths of the ICB, the CMB, the 660 km discontinuity and the Moho for the $\sim$5000 patterns in the test set.

mate is 5160.1 km compared to the values in *PREM* and *ak135*, or earlier estimates, e.g. 5155.0 ± 1.0 km (Bolt, 1977) and 5153.9 km in *PEM* (Dziewoński et al., 1975). We found that the most likely ICB depth is still larger than in *PREM*, if the standard deviation of the data noise, as used during network training, is increased by a factor of $\sim$5 (not shown here). As an aside, we note that we perform a similar check for lower mantle anisotropy in Section 5.4.2. Given the robustness with respect to an anisotropic parametrisation and the assumed data noise level, we conclude that the signal for this intriguing result comes from the data and is not an artefact of the method.

Could such a shift in ICB depth be detected by body waves sensitive to the inner core? We investigate the effect on the travel times of two inner core sensitive phases *PKIKP* and *PKiKP* for a modified version of *PREM* with the ICB depth at 5160 km, denoted $PREM_{ICB}$. Travel times for *PKIKP* and *PKiKP* are typically studied in an epicentral distance range of 130 to 143 degrees to avoid interference between the two phases and the outer core sensitive *PKPab* and *PKPbc* phases, e.g. Waszek et al. (2011); Waszek and Deuss (2013). We compute the travel time difference between *PREM* and $PREM_{ICB}$ in this distance range at one degree intervals using TauP and a surface source. The average difference is 0.38 s ($\sigma = 0.06$ s) for *PKIKP* and 0.33 ($\sigma = 0.08$ s) for *PKiKP*.

For the *PKIKP* data we considered in this study, the estimated noise, i.e. the sample

standard deviation, varies between 0.18 s and 0.43 s in the 130 to 143 degree distance range. Thus, the effect of this shift in ICB depth on the travel times is on average of the same order of magnitude as the assumed data noise, albeit larger for some distances. The ICB depth may be difficult to constrain accurately using these body wave phases, considering that the data are affected by more than just a simple shift in ICB depth, i.e. considering additional trade-offs between other model parameters. Note that we did not use *PKiKP* travel times in this study (Table 4.1); we simply consider TauP synthetic data for *PKIKP* and *PKiKP* for the straightforward analysis performed here.

One way to improve the sensitivity of the body-wave data to the ICB could be to simply sample measumements at a higher density than used here and in Chapter 3, where the epicentral distance interval between travel time measurements was relatively wide (2°). An additional advantage of using complete travel time curves is the possibility to include any triplications due to caustics, which result from strong increases or contrasts in velocity and are thus well-suited to constrain the depth of seismic discontinuities (Stein and Wysession, 2003). Incorporating travel times for reflected phases, such as *PKiKP*, may enhance the resolvability of the ICB depth, provided that the error in the measurements is smaller than the effect of the ICB depth shift on the travel times. Finally, care must be taken as travel time data are heterogeneously distributed, since stations are mainly restricted to continents and earthquakes to plate boundaries. Depending on the amount of data available for a given phase, and—more importantly—the distribution of source and receiver locations, this can become relevant when assessing 1-D Earth structure. For instance, the PKP and PKIKP phases are often recorded along a few distinct ray paths, e.g. from Tonga/Fiji to Eurasia or Sandwich Island to Alaska, and thus do not sample the surface of the ICB homogeneously, e.g. Romanowicz et al. (2003).

4.6.4 Gradients

To understand the Earth's thermochemical structure and dynamics requires an accurate knowledge of the sensitivities of velocity and density with respect to temperature and composition, e.g. Deschamps et al. (2007). The gradients in the radial distribution of these elastic parameters can provide important constraints. For instance, several phase transitions are necessary to explain the gradients in the upper mantle of seismological models, e.g. Cammarano et al. (2005); Cobden et al. (2009).

An insightful first-order inference would be to constrain the sign of the gradient in a certain region. We investigated the potential of the normal mode data to constrain the depth gradients in velocity and density in our model. However, we found that in general very little information on gradients is directly available, especially if the gradient is evaluated over relatively narrow regions of $\sim$200 – 300 kilometres.

4.6.5 A note on the number of synthetic samples

Compared to model space searches with Monte Carlo methods, e.g. Sambridge (1999a); Bodin and Sambridge (2009), which commonly involve $\sim 10^6$ models, we have a rel-

atively low number of samples (100 000). However, with the setup presented here, we are still able to draw insightful inferences on the Earth's radial structure for reasons addressed here. First, the success of the training process and accuracy of the trained networks can be verified by testing network performance on the 5000 test set samples and synthetic data for *PREM* (see Figures 4.2, 4.3 and 4.5). Second, before network training commences the MDN output is initialised to resemble the prior pdf for the target parameter. If no systematic relation between input (seismological data) and output (earth model parameter) is found during training, no information is extracted from the data. Consequently, the final MDN output will resemble the prior, the information gain $D_{KL} = 0$ bits (Section 2.5) and two possibilities exist.

First, the data may simply not be sensitive to the given earth model parameter, compared to the assumed data noise. Second, the set of training samples, which we generated from the prior model space, may not include all regions of the model space of non-zero probability, which is a potential issue of any sampling-based method. As an additional check, we trained networks with a larger set of models (200 000 samples) for some of the target parameters investigated in Section 4.5. We found that for this larger training set, uncertainties decrease slightly for some parameters (the width of the posterior pdf decreases; D_{KL} is a bit higher), but the bulk of probability mass is assigned to similar values as for the set with 100 000 earth models (not shown here). More importantly, we find that in no case the pdf is wider for the larger training set, which would indeed be undesirable. In that sense, the method is conservative and therefore allows us to make insightful inferences on Earth structure. We performed a similar check, with a similar outcome, for a joint inversion of spheroidal and toroidal modes in Section 5.4.1.

4.7 Conclusions

We used artificial neural networks to obtain 1-D marginal posterior pdfs for parameters in a radial earth model, thereby solving seismological inverse problems. We have illustrated the flexibility of the method, which offers the freedom to invert for an arbitrary combination of model parameters and allows us to test specific hypotheses.

The 1-D distributions can be used to quantify uncertainties in the model parameters. In addition, they provide a basis to measure the information contained in the normal mode splitting function measurements, using the Kullback-Leibler divergence. The information content for the various data sets showed that the free oscillations constrain most parameters better than the travel time data. By removing the travel times, the MDNs had fewer free parameters and thus solving the inverse problem was computationally more efficient, whilst the information on the earth model was retained.

The results of our inversions can be summarised as follows:

1. The spheroidal mode data constrain the depths of the ICB, CMB, "660" and Moho. The data strongly prefer an ICB that lies deeper (5154.7–5165.7 km)

than in existing reference models; the result is robust with respect to a radially anisotropic inner core, although less pronounced (5149.9–5165.7 km). The effect on the travel time of inner core-sensitive seismic phases is comparable to the estimated noise in such measurements.

2. The most probable value for the density contrast at the ICB (0.73 g cm^{-3}) is larger than in *PREM* (0.60 g cm^{-3}) and *ak135f* (0.56 g cm^{-3}); the 2σ error levels span a range of 0.52–0.94 g cm^{-3}, which encompasses all previous estimates in the literature. With the data used here, we cannot differentiate between these estimates; their respective differences reflect the non-uniqueness of the inverse problem.

3. We observed a negative deviation with respect to *PREM* for both V_P and V_S in the D" layer, in contrast to a positive anomaly found for the average density. However, these deviations are not robust with respect to a fully anisotropic parametrisation. The data cannot uniquely determine whether the possible density excess is restricted to the D" region or whether it is distributed throughout the lower(most) mantle.

4. In the upper mantle, the strongest deviations from *PREM* were observed for V_P^{Voigt} and V_S^{Voigt} in the "220-Moho" region and V_P in the "410-220" layer, although the latter result is not robust with respect to the anisotropic parametrisation.

5. We found that the data cannot uniquely determine the presence or absence of a discontinuity at the "220". If present, the V_P^{Voigt} and V_S^{Voigt} contrasts are likely smaller than in *PREM*, while the density jump is similar to *PREM*.

6. The MAP estimates for most shear attenuation parameters in the mantle deviate from *PREM* in a similar fashion to results from more recent studies. The data strongly prefer a non-homogeneous shear attenuation in the inner core, enforcing the hypothesis that a distinct innermost inner core may exist.

7. The bulk attenuation is stronger than in *PREM*, save for the inner core, for which most probability is assigned to a Q_κ higher than in *PREM*.

We have also addressed the influence of radial anisotropy on the inversions. We compared the results with the posterior pdfs for a parametrisation in which the inner core and mantle were anisotropic. This enabled us to analyse possible trade-offs between (anisotropic) parameters by constructing conditional and 2-D pdfs. The largest discrepancies were observed in regions that are believed to be anisotropic, such as the D" region. This illustrates the need to constrain anisotropy in the (deep) mantle and suggests the addition of complementary data, such as toroidal modes.

Acknowledgements

We are grateful for the constructive comments of two anonymous reviewers. We thank Paula Koelemeijer for providing additional normal mode measurements. Ralph de Wit, Paul Käufl and Andrew Valentine are funded by the Netherlands Organisation for Scientific Research (NWO) under the grant ALW Top-subsidy 854.10.002. Computational resources for this work were provided by the Netherlands Research Center for Integrated Solid Earth Science (ISES 3.2.5 High End Scientific Computation Resources). We acknowledge the International Seismological Centre for making the EHB bulletin available.

5

Joint inversion of spheroidal and toroidal modes for average radial mantle structure

Abstract

We invert spheroidal and toroidal mode centre frequencies for radial mantle structure using a fully anisotropic parametrisation. Spheroidal and toroidal modes are sensitive to P-SV and SH-motion, respectively, and thus are complementary data types. The information gain for the joint data set is higher than for an inversion using only spheroidal mode data. The biggest increase is observed for shear-wave velocity and anisotropy, in agreement with the enhanced sensitivity to SH motion of the toroidal modes. We checked that the inferences are not biased, at least to first order, by the amplitude of the assumed data noise or the number of synthetic training samples. We analyse 1-D marginal posterior pdfs for radially averaged upper mantle structure (the lower mantle is discussed in Chapter 6). Other than in the "220–Moho" region, we only find evidence for P-wave anisotropy in the transition zone. In summary, our results show both agreement and disagreement with previous studies on upper mantle radial anisotropy in spherically symmetric earth models. Future work should reconcile these studies, or explain the cause for these discrepancies, if we wish to succesfully constrain radial upper mantle structure. This may require the use of complementary data types, such as surface wave phase velocities, body wave travel times and long-period waveforms.

5.1 Introduction

In Chapter 4, we investigated the constraint on radial Earth structure offered by a catalogue of recently measured spheroidal mode centre frequencies (Deuss et al., 2013; Koelemeijer et al., 2013; Koelemeijer, 2014). Two classes of parametrisations were studied: (i) an isotropic parametrisation with a radially anisotropic uppermost mantle, similar to *PREM* (Dziewoński and Anderson, 1981), and (ii) a parametrisation with radial anisotropy in the inner core and the whole mantle. The results for the second class highlighted the existence of anisotropy-related trade-offs between model parameters and stressed the need to supplement the spheroidal mode data, which could improve the constraint on anisotropic parameters. Toroidal modes are primarily sensitive to SH-motion and thus are complementary to the spheroidal modes, which are sensitive to P-SV motion. Therefore, we combine the spheroidal mode centre frequencies, as used in Chapter 4, with similar measurements for toroidal modes (Reference Earth Model web pages, 2001).

In this chapter and in Chapter 6, we use the second, i.e. the fully anisotropic, parametrisation and perform a joint inversion of the spheroidal and toroidal mode data. As in previous chapters, we use Mixture Density Networks (MDNs, Section 2.4.4) to obtain marginal posterior probability density functions (pdfs) for earth model parameters. Our flexible method is very suitable for hypothesis testing, as it enables us to assess the probability of a certain statement or hypothesis. For instance, we can infer the average of any parameter of interest over an arbitrary depth range (Chapter 4). Here, we study the Voigt average isotropic wave velocities, density and parameters describing radial anisotropy in nine layers in the mantle.

This chapter is outlined as follows. First, we briefly describe the data, followed by a quantitative comparison of the information content of the spheroidal mode data and the joint data set. We make this comparison for the average radial structure in nine mantle layers. Furthermore, we investigate the information gain for all target parameters shown in Chapter 4, except for attenuation, since no quality factor data are available for the toroidal modes. Second, we assess the radially averaged isotropic and anisotropic structure in three upper mantle layers and compare these to previous studies of spherically symmetric upper mantle structure (Montagner and Kennett, 1996; Beghein et al., 2006; Kustowski et al., 2008; Visser et al., 2008). In Chapter 6, we focus on radial anisotropy in the lower mantle and its interpretation in terms of the Earth's thermochemical structure. Third, we continue the discussion initiated in Section 4.6.5, which considered the influence of the size of the training set on the network output. Finally, we investigate the robustness of inferences with respect to the amplitude of the assumed data noise.

5.2 Setup

5.2.1 Model parametrisation

The parametrisation of the 1-D earth models, in terms of velocity, density, attenuation and parameters describing radial anisotropy, is described in Appendix A. In this chapter and Chapter 6, we use a parametrisation that allows for radial anisotropy in the inner core and mantle. A radially anisotropic medium can be described by hexagonal symmetry with a vertical (radial) symmetry axis, density and the five independent Love coefficients A, C, N, L and F (Love, 1927). Three parameters are commonly used to describe the radial anisotropy: the P-wave anisotropy ($\phi = \frac{C}{A} = \frac{V_{PV}^2}{V_{PH}^2}$), the shear-wave anisotropy ($\xi = \frac{N}{L} = \frac{V_{SH}^2}{V_{SV}^2}$) and $\eta = \frac{F}{A-2L}$, which corresponds to anisotropy at intermediate incidence angles (Appendix A.2). For an isotropic medium, $\eta = \phi = \xi = 1$.

5.2.2 Normal mode data

We use centre frequencies derived from self-coupled splitting function measurements for spheroidal and toroidal modes, which have a complementary sensitivity to P-SV and SH motion, respectively. The centre frequencies represent the degree-zero component of the mode splitting functions and are thus only sensitive to radial (1-D) Earth structure (Section 4.4.1).

Spheroidal modes

Similar to de Wit et al. (2014), we use 184 self-coupled spheroidal mode centre frequencies up to 10 mHz, the majority of which were measured by Deuss et al. (2013) (Figure 5.1). This extensive recent catalogue contains new observations of modes which are sensitive to V_P and inner core structure. We supplement this catalogue with similar measurements for Stoneley modes (Koelemeijer et al., 2013), fundamental modes ${}_0S_{22} - {}_0S_{30}$ and ${}_2S_{17}$ (Koelemeijer, 2014).

Toroidal modes

We use "best estimate" mean frequencies for toroidal modes, as made available on the Reference Earth Model (REM) web pages (http://igppweb.ucsd.edu/~gabi/rem.dir/surface/tmodes.list). We selected toroidal modes with radial orders $n = 0$–5 and angular orders similar to those of the spheroidal modes, i.e. $l = 1$–30 (Figure 5.1). The resulting data set included 125 toroidal mode measurements with frequencies up to 8 mHz, almost half of which were measured by Widmer (1991). By adding the 125 toroidal modes to the 184 spheroidal mode measurements, we almost doubled the dimensionality of the input to the neural networks. Consequently, the training times for the neural network increased. A total of 287 toroidal mode measurements (for

higher angular orders) is available on the REM website, but we decided to limit the number of toroidal modes to align the two data sets in terms of the angular orders and their frequency content. An additional advantage is that the total input dimensionality ($184 + 125 = 309$) is relatively low (compared to $184 + 287 = 471$), and so are the number of free network parameters, the required number of training samples and thus computation time. In contrast to the spheroidal mode data set, no estimates of quality factors are available for the toroidal modes. Therefore, we cannot investigate the information content of the joint data set for the attenuation parameters.

5.2.3 Synthetic data

We use the Mineos package (Masters et al., 2011) to calculate exact normal mode frequencies for 100 000 synthetic 1-D earth models, which we generated randomly from the prior model distribution (Appendix A). Self-gravitation was taken into account for frequencies below 30 mHz and a reference period of 1 s was used for the dispersion correction due to attenuation. The synthetic data for the normal modes thus consisted of 309 (184+125) free oscillation centre frequencies. The synthetic data were corrupted by adding Gaussian noise with zero mean and a standard deviation given by the uncertainty estimate for each measurement, as reported by Deuss et al. (2013); Koelemeijer et al. (2013); Koelemeijer (2014) and on the REM website (http://igppweb.ucsd.edu/~gabi/rem.dir/surface/tmodes.list).

We note that the measurement errors were estimated using a cross-validation approach, which may not fully account for any systematic uncertainties. The measurement errors are used during network training to make the neural networks insensitive to noise-level fluctuations. Clearly, if the error estimates are underestimated, network training will map noise in the input signal into the prediction for the earth model parameters. As in previous chapters, we have assumed a simple Gaussian noise model for each measurement, since there is no information available on possible correlations in the data noise. However, we can assess the sensitivity of the MDN output to the amplitude of the estimated measurement errors. As a first-order test, we increase the noise amplitude and analyse the effect on the inferences for lower mantle anisotropy in Section 5.4.2.

5.2.4 Network configuration

In Chapter 4, we trained MDNs with a 184-D input vector, 40 hidden units and a mixture of 15 Gaussian kernels as output. For the joint inversion of the spheroidal and toroidal modes, the input to the networks consists of 309 measurements. To ensure that the neural network can extract additional information—if available—from this larger input vector, we increase the number of hidden units to 50. This number reflects a compromise between increasing flexibility during network training and restricting the size of the network and the required number of training samples. We verified that the precise size of the hidden layer is not of paramount importance to network prediction accuracy. This results in MDNs that consist of 17 795 free param-

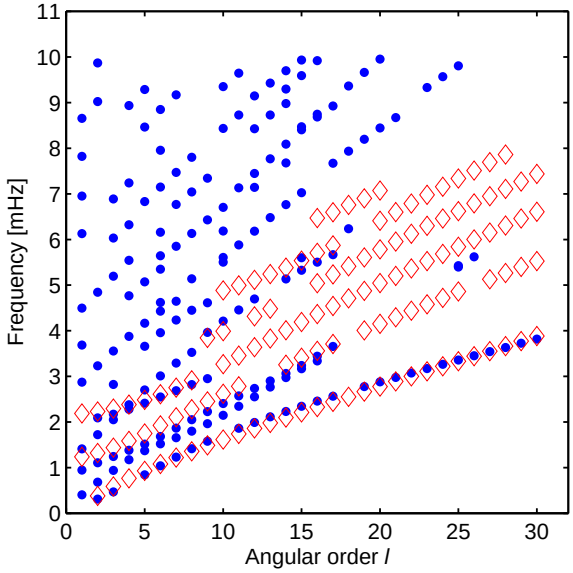


Figure 5.1: Measured centre frequencies for the 184 spheroidal (blue dots) and 125 toroidal (red diamonds) modes, which form the input to the neural networks.

eters (Equation 4.1). Thus, the number of network parameters for the joint inversion is roughly twice as large as for the MDNs trained on the spheroidal mode data in Chapter 4 (9245 free parameters). Since the number of synthetic examples is the same (100 000) for both setups, the ratio of training samples to network parameters is twice as low for the joint inversion. We will further explore this issue in Section 5.4.1.

For each target parameter considered in this chapter and Chapter 6, we construct an ensemble of 48 networks (Section 2.4.8). A network ensemble can result in better generalisation, i.e. achieve a better prediction accuracy on unseen data (Bishop, 1995). The ensemble output is formed by a weighted average of the members, where the individual weights are determined by each network's performance on the same test set (Käufl et al., 2014; de Wit et al., 2014). Each individual network is trained using the Scaled Conjugate Gradient (SCG) algorithm (Møller, 1993) for a maximum of 5000 iterations. As in de Wit et al. (2013, 2014), we employ early stopping, which means that network training is halted when the error of a separate validation set reaches a minimum. 80% of the 100 000 patterns in the synthetic data set is used for training, 15% for the validation set and the remaining 5% for the test set. For each network realisation, the synthetic data are randomly divided over training, validation and test sets to enhance the generalisation capability of the ensemble.

5.3 Results

5.3.1 Network target parameters

We study the radially averaged η, ϕ, ξ, ρ and the Voigt average isotropic velocities V_P and V_S (Equations A.4 and A.5) in nine mantle layers. From hereonwards, we always use this exact formulation for the Voigt average velocities, as opposed to the approximate formulations used in Section 4.5.3 (Figure 4.4). Note that this approximation to the Voigt average V_P and V_S is valid under the assumption of small anisotropy, i.e. $\eta \approx 1$, and would be valid here as well, but the exact formulation comes at no additional computational cost. The upper mantle consists of three layers, enclosed by the Moho, "220", "410" and "660" discontinuities. The bulk of the lower mantle, Bullen's D′ region (Bullen, 1949), is divided into five layers of roughly equal thickness, which have approximate depth ranges 670–1027, 1027–1456, 1456–1884, 1884–2313 and 2313–2741 in kilometres. The D" region forms the sixth layer (2741–2891 km). The depths of the three discontinuities enclosing the D′ and D" regions, i.e. the top of the lower mantle (670 km), the top of the D" region (2741 km) and the core-mantle boundary (CMB, 2891 km), are allowed to vary by $\pm$20–30 km in the earth model parametrisation (Table A.2). For each synthetic model, the depths of the remaining boundaries of the five layers in D′ are determined by linearly interpolating between the new depths of the discontinuities at 670 and 2741 km. Note that all other parameters in the model, i.e. parameters describing bulk and shear attenuation, core and upper mantle structure, are allowed to vary in the prior model distribution (Figure A.1). We emphasise that we do not need to impose constant-value layers in the earth models, separated by discontinuities fixed *a priori*, to be able to assess the radially averaged structure in the mantle layers. The MDNs can be trained directly on such radial averages, while the underlying 1-D earth models in the training set are realistic, i.e. smooth, similar to existing reference earth models.

Further, we train MDNs on the joint data set for all target parameters investigated in Chapter 4, except for the targets related to the inner core, since our toroidal mode data are insensitive to this part of the Earth. Moreover, inner-core–sensitive toroidal modes are extremely difficult or even impossible to observe, although they could be excited due to cross-coupling with spheroidal modes (Dziewoński and Woodhouse, Dziewoński and Woodhouse). Dahlen and Tromp (1998) note that for a spherically symmetric non-rotating elastic isotropic (SNREI) Earth, such toroidal modes only exist in theory; the amount of energy required to excite such modes cannot be generated by earthquakes in the crust or upper mantle, nor can they be registered by seismometers along the Earth's surface.

5.3.2 Analysis of information content

For each target parameter, we calculate the information gain, as quantified by the Kullback-Leibler divergence D_{KL} (Section 2.5). We train MDNs on parameters in the anisotropic earth models and compare D_{KL} for the inversion of the spheroidal modes

(S_I) and the inversion of the joint spheroidal and toroidal mode data set (S_{II}). Table 5.1 shows D_{KL} for the two data sets for the averages of the six seismic parameters (V_P, V_S, ρ, η, ϕ and ξ) in nine layers in the mantle (Section 5.2.1). Tables 5.2 to 5.4 show D_{KL} for all other target parameters that were considered in de Wit et al. (2014), i.e. discontinuity depths, jumps across discontinuities and the average density in the lower mantle, except for attenuation.

D_{KL} is higher for S_{II} for most parameters in Tables 5.1 to 5.4; thus, the toroidal modes are complementary to the spheroidal mode measurements. For a few target parameters, D_{KL} is lower for S_{II}, e.g. for the depth of the "410" (Table 5.2) and the jumps in V_P and density across the "660" (Table 5.3). For such parameters, less information could be extracted, probably due to the lower ratio of the number of training samples versus the number of network parameters. We will discuss the effect of a larger training set on D_{KL} in Section 5.4.1. For the three anisotropic parameters, the increase in information upon addition of the toroidal modes is evident (Table 5.1). The most notable increase in D_{KL} is observed for shear wave anisotropy (ξ), in accord with the complementary sensitivity to SV and SH motion provided by the joint data set. Both ϕ and ξ are strongly constrained for S_{II} in all layers ($D_{KL} \geq 5.3$ bits), except for the D" region. η is only resolved in the five layers in the D' region, with the strongest constraint reached in the four relatively thick ($\sim$ 435 km) layers (LM_{II-V}).

Finally, it is important to note that, although D_{KL} increases, the marginals are similar for S_I and S_{II} for the parameters considered in Chapter 4, i.e. the discontinuity depths and the velocity and density contrasts in across mantle discontinuities. Therefore, we do not consider the marginals for these parameters further in this chapter, and focus on the radial averages in the upper mantle layers. For the average density in the lowermost mantle, the transition from an isotropic to anisotropic model parametrisation resulted in a significant change in the marginals (Figure 4.8). We find that the addition of the toroidal mode measurements also influences the marginal pdfs and thus the inference, albeit less strongly than the anisotropic parametrisation (Table 5.4). We will investigate the average density and anisotropy in the lower mantle in detail in Chapter 6.

Table 5.1: D_{KL} (in bits) for target parameters representing averages in nine mantle layers. For each parameter, the two columns show D_{KL} for the two data sets S_I and S_{II}.

Layer	Depth [km]	$\bar{V}_P$		$\bar{V}_S$		$\bar{\rho}$	
		S_I	S_{II}	S_I	S_{II}	S_I	S_{II}
"220–Moho"	24.4–220	2.6	2.1	1.4	4.2	5.8	7.0
"410–220"	220–400	1.9	2.0	0.6	1.4	3.9	4.3
TZ	400–670	4.9	5.2	1.0	2.8	7.8	8.1
LM_I	670–1027	6.5	6.8	2.4	4.6	5.0	5.8
LM_{II}	1027–1456	8.9	8.9	5.5	7.3	7.2	8.5
LM_{III}	1456–1884	9.1	9.5	7.3	7.9	7.9	9.6
LM_{IV}	1884–2313	9.5	9.7	6.5	10.1	9.5	10.1
LM_V	2313–2741	8.7	8.2	5.3	9.1	8.4	8.7
D"	2741–2891	2.3	2.8	1.4	2.4	0.3	0.3
Layer	Depth [km]	$\bar{\eta}$		$\bar{\phi}$		$\bar{\xi}$	
		S_I	S_{II}	S_I	S_{II}	S_I	S_{II}
"220–Moho"	24.4–220	0.3	0.3	6.3	7.3	0.6	7.9
"410–220"	220–400	0.1	0.1	7.5	7.2	0.3	5.3
TZ	400–670	0.2	0.6	10.7	10.6	0.5	10.0
LM_I	670–1027	1.2	1.4	9.1	9.3	1.9	9.2
LM_{II}	1027–1456	3.5	5.5	6.9	11.2	4.1	11.2
LM_{III}	1456–1884	8.2	8.7	10.7	10.8	6.3	12.1
LM_{IV}	1884–2313	9.0	9.8	10.8	10.9	7.7	11.8
LM_V	2313–2741	6.5	7.1	9.7	10.3	5.8	10.1
D"	2741–2891	0.1	0.1	1.8	1.9	0.6	2.6

Table 5.2: D_{KL} (in bits) for target parameters representing discontinuity depths. For each parameter, the two columns show D_{KL} for the two data sets (see Table 5.1).

Discontinuity	D_{KL} [bits]	
	S_I	S_{II}
CMB	8.7	11.0
D" (top)	0.0	0.0
"660"	2.8	2.9
"410"	0.6	0.4
"220"	0.0	0.0
Moho	4.5	5.0

Table 5.3: D_{KL} (in bits) for target parameters representing jumps across mantle discontinuities. For each parameter, the two columns show D_{KL} for the two data sets (see Table 5.1).

Discontinuity	$\bar{V}_P$		$\bar{V}_S$		$\bar{\rho}$	
	S_I	S_{II}	S_I	S_{II}	S_I	S_{II}
"660"	1.9	1.7	0.5	1.1	3.7	3.5
"410"	0.8	1.2	0.3	0.8	2.2	2.1
"220"	1.2	1.8	0.7	1.7	1.8	2.2

Table 5.4: D_{KL} (in bits) for target parameters representing density averages in the lower(most) mantle. For each parameter, the two columns show D_{KL} for the two data sets (see Table 5.1).

Depth [km]	D_{KL} [bits]	
	S_I	S_{II}
2891–2376	4.1	4.4
2891–1792	7.1	7.9
2891– 670	7.5	8.5
2741–2376	7.5	8.2
2741–1792	9.3	10.3
2741– 670	9.2	9.8

5.3.3 Inferences on upper mantle structure

We now compare the pdfs for S_I and S_{II} for the average velocities, density and three anisotropic parameters in the upper mantle (Figure 5.2 and 5.3). In Chapter 6, we focus on these six seismic parameters in the six lower mantle layers.

Velocities and density

de Wit et al. (2014) compared the results for the isotropic and anisotropic parametrisation by computing the overlap between the respective 1-D pdfs (Figures 4.7 and 4.8). Here, we focus on the differences between the pdfs for inversions of the two data sets (Figure 5.2). A discrepancy between the pdfs for S_I and S_{II} does not necessarily result from trade-offs between model parameters, as was the case for the comparison of the two classes of parametrisation (Section 4.6.1). Rather, a low overlap may simply relate to the additional information offered by the toroidal mode data. We observe that the pdfs for S_{II} are slightly narrower for most parameters, in agreement with the higher D_{KL} (Table 5.1). Further, the 1-D pdfs agree well for both velocities and density; the bulk of the probability mass is assigned to similar values for most target parameters. Some differences are visible for V_S in the "410–220" and "220–Moho" regions and the density in the "220–Moho" region, but the overlap in the pdfs is still relatively high ($> 60\%$). In summary, the results for the velocities and density are mostly consistent for S_I and S_{II}, the main difference being a slightly better constraint due to the joint data set.

For S_{II} (Figure 5.2, bottom row), the marginals for V_P and V_S in the "410–220" region and transition zone agree with the reference models *PREM* and *ak135f* (Kennett et al., 1995; Montagner and Kennett, 1996). Both *PREM* and our marginals are in disagreement with the density structure in *ak135f*, but Montagner and Kennett (1996) note that the density in the upper mantle of their model should be treated with caution. In the "220–Moho" region, both V_P and V_S may be higher than in both *PREM* and *ak135f*, while the average density is likely lower. The Maximum A Posteriori (MAP) estimate for the average density in the "220–Moho" region corresponds to a density deficit of 1.6% with respect to *PREM*. This agrees with a negative density deviation found in a model space search using normal mode and surface wave phase velocity data (Beghein et al., 2006), although their posterior pdf is more conservative and includes *PREM* and positive density deviations as explanations for the data.

Anisotropic parameters

As became evident from the analysis of the information content, the most prominent difference between the 1-D pdfs for S_I and S_{II} is the much tighter constraint, mainly on ξ, that is offered by the joint data set (Figure 5.3). Consequently, we find that the overlap between the pdfs is low ($\sim$30–50%) for ξ in all three layers. It is straightforward to extract the probability of anisotropy from the 1-D pdfs. Table 5.5 summarises the probability that η, ϕ and ξ are smaller than in *PREM* for the joint in-

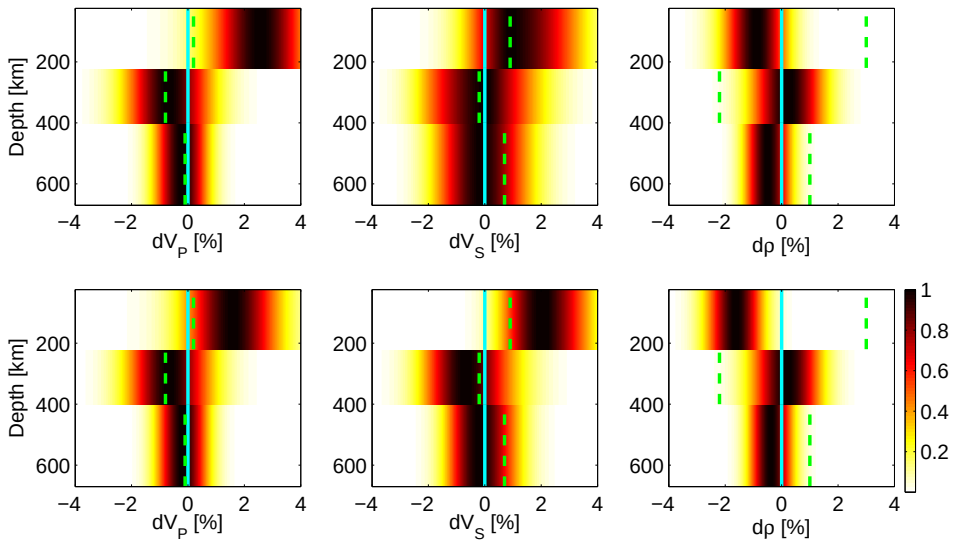


Figure 5.2: 1-D marginal posterior pdfs for S_I (top row) and S_{II} (bottom row). Shown are $\bar{V}_P$ (left-hand panel), $\bar{V}_S$ (middle panel) and $\bar{\rho}$ (right-hand panel) in the upper mantle, expressed as percentage deviations with respect to *PREM* (cyan line), as is the model *ak135f* (Kennett et al. (1995); Montagner and Kennett (1996), green-dashed line). The probability for each 1-D pdf is rescaled so that the maximum equals 1. Asymmetric 1σ and 2σ error bars correspond to the $1/e^{1/2}$ (0.61) and $1/e^2$ (0.14) contours, respectively.

version (S_{II}). No significant anisotropy is observed in the upper mantle, other than in the "220–Moho" region and for ϕ in the TZ. The anisotropy in the "220–Moho" region may be slightly stronger than in *PREM*, with probability 0.86 (ϕ) and 0.78 (ξ, $1.0 - 0.22 = 0.78$). The same is true for η, with probability 0.74, but this parameter is largely unconstrained by the data (Table 5.1) and thus the posterior pdf includes a wide range of possible values. We infer a value for ξ close to 1 in both the "410–220" region and the TZ, which disagrees with Montagner and Kennett (1996); Visser et al. (2008) but is largely in agreement with Beghein et al. (2006); Kustowski et al. (2008). In contrast to the latter study, we do not find strong evidence for P-wave anisotropy in the "410–220" region ($p(\phi > 1) = 1 - 0.33 = 0.67$), but we note that our inferred $\pm 2\sigma$ error level for this parameter is conservative and covers a range of ϕ =0.976–1.036 (-2.4–3.6%), i.e. a few percent of either positive or negative P-wave anisotropy is admissible. Possible changes in sign of the anisotropy above (and below) the transition zone, as reported by Montagner and Kennett (1996), are not constrained by the data we used here. However, we find $\pm 2\sigma$ =-2.5–0.7% in the transition zone, and the corresponding probability $p(\phi < 1) = 0.88$ (Table 5.5) is in line with the negative TZ P-wave anisotropy in their model.

Table 5.5: Probability that the average η, ϕ and ξ in three layers in the upper mantle are smaller than *PREM* for the joint inversion (S_{II}). Note that for "410–220" and the TZ, *PREM* is isotropic ($\eta = \phi = \xi = 1$).

Layer	Depth [km]	$\bar{\eta}$	$\bar{\phi}$	$\bar{\xi}$
"220–Moho"	24 – 220	0.74	0.86	0.22
"410–220"	220 – 410	0.55	0.33	0.62
TZ	410 – 670	0.56	0.88	0.39

5.4 Discussion

5.4.1 Information content and the number of synthetic samples

In Section 4.6.5, we considered the influence of the size of the training set on network performance. We doubled the training set to 200 000 patterns and found that the network output was similar to the output for the original training set of 100 000 samples. Uncertainties decreased slightly, i.e. D_{KL} was higher, for some parameters, but most of the probability mass was assigned to similar values as for the set with 100 000 earth models. We argued that our method is conservative, as for none of the target parameters the pdf was wider, or equivalently D_{KL} lower, for the larger training set.

For the joint inversion performed in this chapter, the ratio of synthetic examples to network parameters is approximately twice as small as it is for the inversion of spheroidal modes. Therefore, we performed a similar test by training MDNs on a set of 200 000 samples. We investigated whether network prediction accuracy improves and whether the information extracted from the observed data (D_{KL}) increases.

For a selection of target parameters, Table 5.6 shows D_{KL} for the two training sets. The difference for the well-resolved ϕ and ξ in the TZ is negligible, indicating that the MDN performance does not improve when using more samples and that the information available on these parameters has been extracted from the data. The associated marginals are similar: for instance, the probability of $\xi < 1$ is 0.45 (*ST200*) and 0.39 (*ST100*), i.e. for both training sets the pdf for ξ in the TZ is approximately centred on *PREM* (Table 5.5). By contrast, D_{KL} increases for the poorly resolved η in the TZ, although the improved constraint is only moderate ($D_{KL} = 1.3$ bits). Clearly, the MDNs benefited from the increased number of prior samples for this parameter. Further, we find that the marginals for the two data sets differ slightly: the data prefer negative η anisotropy in the TZ with probability 0.76 for *ST200*, whereas this probability is 0.56 for *ST100*. For η in the five lower mantle layers, D_{KL} is similar for the two training sets, as are the MAP estimates and the width of the pdfs (not shown here).

D_{KL} increases for the six discontinuity depths (Table 5.6), except for the top of the D" region and the "220", which are unresolved by the data ($D_{KL} = 0.0$ bits). However, the increase in D_{KL} is relatively small and we find that the MAP estimates are similar for both training sets, e.g. for the CMB depth the peaks of the pdfs lie at 2890.6 km (*ST*100) and 2891.1 km (*ST*200). Thus, the only difference is the width of the pdfs,

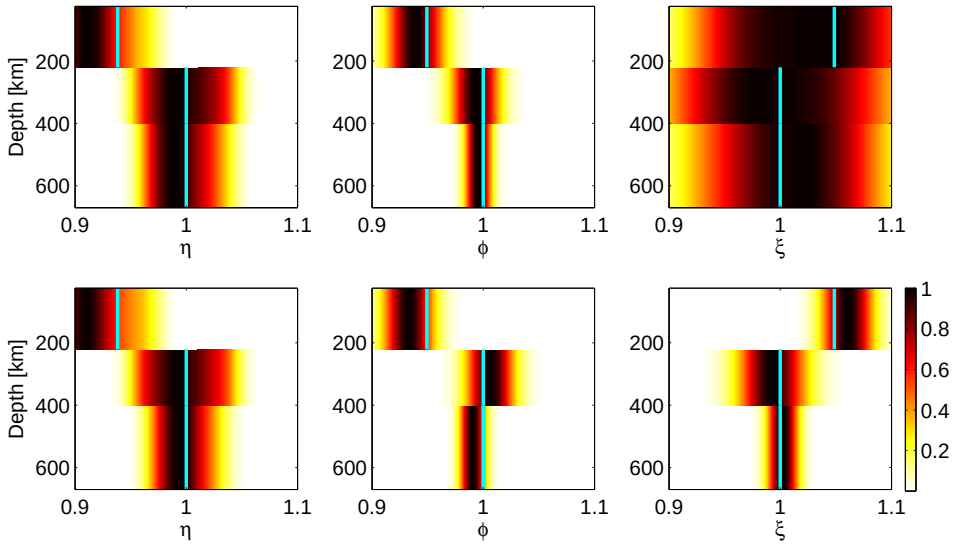


Figure 5.3: 1-D marginal posterior pdfs for S_I (top row) and S_{II} (bottom row). Shown are $\bar{\eta}$ (left-hand panel), $\bar{\phi}$ (middle panel) and $\bar{\xi}$ (right-hand panel) in the mantle. *PREM* is indicated by the cyan lines. The probability for each 1-D pdf is rescaled so that the maximum equals 1. Asymmetric 1σ and 2σ error bars correspond to the $1/e^{1/2}$ (0.61) and $1/e^2$ (0.14) contours, respectively.

which reflects the higher certainty on the target parameters when training networks on the larger data set *ST200*. For the D" region, D_{KL} increases slightly for ϕ and ξ, while η remains unresolved. This suggests that the lack of constraint is likely related to the lack of information on this region in the data and not the number of training patterns.

Arguably, if we find an improvement in network performance, we should train networks on an even larger training set, i.e. use more than 200 000 patterns, until some convergence of prediction accuracy and D_{KL} has been reached. Another argument to use more patterns is the classical rule of thumb that one needs approximately ten times as many training patterns as there are adjustable network parameters to achieve a prediction accuracy of ∼90%, e.g. Duda et al. (2001). Clearly, we are not adhering to this rule here, as we have 80 000 training patterns and 17 795 network parameters (Section 5.2.4), but we feel that this rule of thumb becomes less relevant in our current setup using MDN ensembles (see the discussion in Section 2.4.8). As we argued in Section 4.6.5 and in the above, any changes that do occur show that the output of the MDN ensembles is conservative. The widths of the 1-D marginal pdfs reflect both the uncertainty in the target parameters and, to first order, the uncertainty in the inverse mapping approximated by the networks. This is a desirable property, since we want to minimise the possibility that an (overly) optimistic inference, as represented by a relatively narrow pdf, rejects the true Earth as a possible explanation for the measured data. Therefore, we regard the pdfs obtained using 100 000 patterns as a final result,

Table 5.6: D_{KL} (in bits) for the joint inversion, using networks trained on 100 000 (*ST100*, left-hand column) and 200 000 (*ST200*, right-hand column) synthetic samples.

Target	*ST100*	*ST200*	Target	*ST100*	*ST200*
$\bar{\eta}$ TZ	0.6	1.3			
$\bar{\phi}$ TZ	10.6	10.9	CMB	11.0	11.9
$\bar{\xi}$ TZ	10.0	10.0	D" (top)	0.0	0.0
$\bar{\eta}$ LM_I	1.4	1.5	"660"	2.9	3.4
$\bar{\eta}$ LM_{II}	5.5	5.8	"410"	0.4	0.9
$\bar{\eta}$ LM_{III}	8.7	8.8	"220"	0.0	0.0
$\bar{\eta}$ LM_{IV}	9.8	10.3	Moho	5.0	5.4
$\bar{\eta}$ LM_V	7.1	7.5			
$\bar{\eta}$ D"	0.1	0.1	$\bar{V}_P$ D"	2.8	3.2
$\bar{\phi}$ D"	1.9	2.6	$\bar{V}_S$ D"	2.4	3.2
$\bar{\xi}$ D"	2.6	2.9	$\bar{\rho}$ D"	0.3	0.6

strengthened by the observation that the inferences are similar when using a larger training set.

5.4.2 Information content and data noise

We noted earlier that the noise estimates, as reported in the normal mode catalogues, may not fully account for systematic uncertainties (Sections 4.4.1 and 5.2.3). If systematic errors are neglected, or if the amplitude of the noise is underestimated, the result of an inversion may be biased. Clearly, such a problem pertains to any inference based on noisy observations. As a first-order analysis, we study the robustness of MDN output with respect to the amplitude of the measurement noise. Thus, this simple exercise does not consider possible correlations between the noise in individual measurements. As an example, we analyse the effect on the results for the average η in two lower mantle layers,

We reiterate that the reported measurement uncertainties are used as the standard deviations of zero-mean Gaussian distributions, from which we generate random noise that is added to the synthetic data. By doing so, the neural networks become insensitive to noise-level differences between the synthetic training samples, i.e. prevents the networks from overfitting details in the training data (Section 2.4.6). We multiply the error estimates for each mode measurement with a factor $A \geq 1$ and train MDNs on the average η in the five lower mantle layers LM_{I-V}. $A = 1$ corresponds to the reported measurement uncertainties and thus to D_{KL} shown in Table 5.1 (S_{II}) for these five target parameters. Naturally, D_{KL} is lower for for higher noise levels ($A > 1$), but we find that η is still resolved for most layers, even for $A = 10$ (Table 5.7). Furthermore, we find that the MAP estimates (the peaks of the pdfs) are

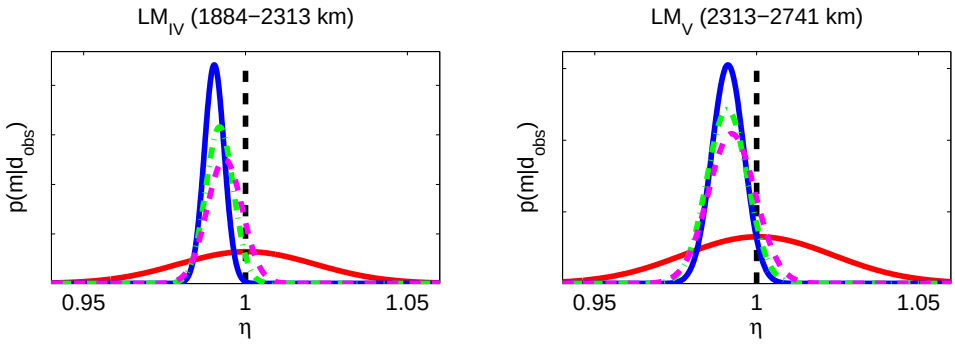


Figure 5.4: 1-D marginal posterior pdfs for η in LM_{IV} (left-hand panel) and LM_{V} (right-hand panel) for noise level amplification factors $A = 1$ (blue, solid line), $A = 5$ (green, dashed-dotted line) and $A = 10$ (magenta, dashed line). Also shown are the prior pdf (red, solid line) and the isotropic *PREM* model ($\eta = 1$, black, dashed line).

Table 5.7: D_{KL} (in bits) for the average η in five lower mantle layers for the joint data set and various noise amplification factors A.

Layer	Depth [km]	$A = 1$	$A = 5$	$A = 10$
LM_{I}	670 –1027	1.4	1.0	0.7
LM_{II}	1027 –1456	5.5	3.3	2.4
LM_{III}	1456 –1884	8.7	6.6	4.6
LM_{IV}	1884 –2313	9.8	7.3	5.5
LM_{V}	2313 –2741	7.1	5.4	4.1

largely insensitive to the amplitude of the noise levels. As an example, we show the marginal pdfs for η in the two deepest lower mantle layers (LM_{IV-V}) for the three values of A (Figure 5.4). Clearly, the pdfs become wider for a higher noise level, but in all cases we would infer that the data prefer $\eta < 1$ for these two parameters. In Chapter 6, we assess possible radial anisotropy in the lower mantle in more detail.

5.5 Conclusions

We performed a joint inversion of complementary spheroidal and toroidal mode centre frequencies using a fully anisotropic parametrisation. We made a quantitative comparison of the information gain for the joint data set and the inversion based on the spheroidal mode data. For most earth model parameters, the constraint was larger for the joint inversion. The biggest increase was observed for shear-wave velocity and anisotropy, as a result of the additional sensitivity to SH motion offered by the toroidal modes. In this chapter, we focused on the quantification of information content and analysed 1-D marginal posterior pdfs for upper mantle parameters. We investigate

the radial structure of the lower mantle in Chapter 6.

The marginal posterior pdfs for wave velocities and density are similar for the two data sets (S_I and S_{II}). The biggest difference occurs for the average density in the "220–Moho" region, which is likely lower than in *PREM*. With the joint data set, we were able to constrain the radially averaged ϕ and ξ in three upper mantle layers, while η was largely unconstrained. Other than in the "220–Moho" region, which is known to be anisotropic, we only found evidence for anisotropy in ϕ in the transition zone. The anisotropy in the "220–Moho" region may be slightly stronger than in *PREM*. An isotropic structure for ξ (and η) in both the "410–220" region and the TZ contrasts with Montagner and Kennett (1996); Visser et al. (2008) but supports the findings by Beghein et al. (2006). Possible changes in sign of the anisotropy around the transition zone (Montagner and Kennett, 1996) are not constrained by our joint data set.

In any study based on the sampling of a model space, one has to investigate the number of samples that is necessary to successfully draw inferences on the model in question. Continuing the discussion in Section 4.6.5, we analysed the difference in information content for training sets of two different sizes. We concluded that our method is conservative, since all pdfs were either similar or slightly wider for the smaller training set. In that sense, the MDN ensemble output reflects both the uncertainty in the earth model parameter and, to first order, the uncertainty in the inverse mapping approximated by the neural networks.

As a final exercise, we investigated the effect of the assumed measurement errors on the marginal posterior pdfs. If the noise level is underestimated, the networks may map data noise into our estimates for the earth model parameters. We increased the amplitude of the assumed noise and investigated the robustness of η anisotropy in the lower mantle. As such, we did not consider possible correlations between the errors in individual measurements. However, taking such correlations into account in a complete noise model for the centre frequencies is not trivial. We found that the normal mode data constrained η in the deep lower mantle, even for noise levels that were one order of magnitude larger than the reported measurement uncertainties. Furthermore, the bulk of the probability mass was assigned to similar values for η, regardless of the noise amplification factor.

In summary, our results partly agree and partly disagree with previous studies on upper mantle radial anisotropy in spherically symmetric earth models (Montagner and Kennett, 1996; Beghein et al., 2006; Kustowski et al., 2008; Visser et al., 2008). We illustrated that our inferences are, at least to first order, not biased by the amplitude of the assumed data noise or the number of synthetic training samples. This leaves differences in the types of data used, such as normal modes, body waves and surface waves, as a possible cause for the discrepancies between our and other upper mantle studies. Future work should reconcile these studies, or explain why these discrepancies exist, if we wish to succesfully constrain radial upper mantle structure. In our case, this may require the addition of complementary data types to the joint normal mode data set, such as surface wave phase velocities (Beghein et al., 2006; Visser et al., 2008), body wave travel times and long-period waveforms (Kustowski et al., 2008).

6

Robust constraints on average radial lower mantle anisotropy and consequences for composition and texture

Abstract

Seismic anisotropy has been observed in the upper mantle (<660 km depth) and the lowermost ∼150–250 km of the mantle (the D" region), while the remainder of the lower mantle is believed to be isotropic. Here, we use centre frequencies for spheroidal and toroidal normal modes together with a neural network-based technique to infer probability density functions for the average radial anisotropy in the lower mantle. We show, for the first time, a robust observation that the average lower mantle is anisotropic below 1900 km depth, challenging the consensus that this part of the mantle is isotropic. The mass density also shows a well-constrained positive deviation from existing models at the same depths. Our results are compatible with an average lower mantle that is about 100–200K colder than commonly-assumed adiabats and that consists of a mixture of about 60–65% perovskite and 35–40% ferropericlase containing 10–15% iron. The observed anisotropy constrains the orientation of the two minerals and opens a new window to study the nature of mantle flow.

The content of this chapter has been submitted to *Earth and Planetary Science Letters* as part of: de Wit, R. W. L. and J. Trampert, 2015. Robust constraints on average radial lower mantle anisotropy and consequences for composition and texture.

6.1 Introduction

Seismic anisotropy, the direction-dependence of elastic wave propagation, can be a key indicator of mantle flow, deformation and consequently mantle dynamics (Montagner, 1994; McNamara et al., 2002; Panning and Romanowicz, 2004). It is commonly interpreted as lattice-preffered orientation (LPO) or shape-preferred orientation (SPO) of the mineral crystals that constitute the mantle (Karato, 2008; Fichtner et al., 2013). LPO refers to the alignment of intrinsically anisotropic minerals, such as olivine, while SPO relates to (long-wavelength) apparent anisotropy that is observed as a result of a specific configuration of isotropic material, e.g. a stack of thin alternating layers with contrasting elastic properties, melts or cracks (Backus, 1962).

Seismic anisotropy has been observed in the upper mantle (above the 660 km discontinuity) and the lowermost ∼150–250 km of the mantle (Montagner and Kennett, 1996; Panning and Romanowicz, 2004; Beghein et al., 2006; Visser et al., 2008; Chang et al., 2014). By contrast, the current consensus is that the remainder of the lower mantle is isotropic, although both experimental and modelling studies have shown that lower mantle minerals are intrinsically anisotropic (Meade et al., 1995; Mainprice et al., 2000). Karato et al. (1995) explained the absence of lower mantle anisotropy by superplastic flow, since the associated diffusion creep does not lead to the development of LPO of mantle minerals. There are also no viable candidates known for SPO in the lower mantle.

Most seismically anisotropic earth models suffer from several limitations. First, seismological inverse problems are notoriously non-unique. Second, scaling between chosen model parameters is often imposed to simplify the seismological inverse problem and reduce the number of free parameters, but may lead to biased models (Beghein et al., 2006; Panning and Romanowicz, 2006; Kustowski et al., 2008). Finally, regularisation is commonly applied to stabilise the inverse problem, which can have a significant effect on the final solution (Beghein and Trampert, 2003; de Wit et al., 2012). These issues call for a quantitative assessment of model uncertainties. Nonetheless, most models come without error bars, which makes it impossible to quantify the discrepancies between existing models.

Normal mode theory provides a means to relate the Earth's free oscillations to Earth structural parameters (Dahlen and Tromp, 1998). Spheroidal and toroidal modes have a complementary sensitivity to P-SV and SH motion, respectively, and a joint inversion of these two types of free oscillations can provide a strong constraint on the anisotropic structure. Therefore, we use centre frequencies derived from self-coupled splitting function measurements for 184 spheroidal (Deuss et al., 2013; Koelemeijer et al., 2013; Koelemeijer, 2014) and 125 toroidal (Reference Earth Model web pages, 2001) modes (Section 5.2.2). Splitting function measurements naturally separate structure with radial symmetry from 3-D variations. Centre frequencies only depend on the degree-zero splitting function coefficients c_{00} and are thus only sensitive to radial (1-D) Earth structure. We do not use splitting coefficients of higher degrees, since they correspond to lateral variations in Earth structure. Therefore, our data are only sensi-

tive to radial anisotropy, which can be represented by three parameters, describing the P-wave anisotropy ($\phi = \frac{C}{A}$), the shear-wave anisotropy ($\xi = \frac{N}{L}$) and the anisotropy at intermediate incidence angles ($\eta = \frac{F}{A-2L}$). A, C, N, L and F are the five independent Love coefficients (Love, 1927) commonly used to characterise an anisotropic medium with a radial symmetry axis.

We assess anisotropy in the mantle in a fully quantitative manner, i.e. we solve the inverse problem and quantified uncertainties without imposing any scaling between parameters. We adopt a Bayesian framework, in which any inference made about a model is the result of the conjunction of our current (*prior*) knowledge and the ability of the model to explain the observations (Tarantola and Valette, 1982). The updated (*posterior*) knowledge on the model—that is, after observing the data—represents the new degree of belief in the model, expressed by a probability density function (pdf). In this study, we only consider the marginal posterior pdfs for single earth model parameters. Such a 1-D marginal posterior pdf, hereafter referred to as a 'marginal', represents the information on a single model parameter, given the data and the possible variations in all other model parameters. We employ machine learning techniques to learn relationships between data and model based on samples of the prior model space. To obtain marginals, we use a Mixture Density Network (MDN, (Bishop, 1995; de Wit et al., 2013; Käufl et al., 2014)), which takes the normal mode data as input and outputs the marginal for the earth model parameter of interest (Section 2.4.4). The specifics of the network configuration are given in Section 5.2.4. We emphasise that our method is not based on a classical misfit function, i.e. it does not explicitly measure how well a complete earth model explains the data.

Instead, our inversion method is designed to provide a flexible tool for hypothesis testing, which allows us to assess the probability of a certain statement or hypothesis. The flexibility enables us to focus on averages of any parameter of interest over an arbitrary depth range (de Wit et al., 2014). We invert the normal mode data for the radially averaged η, ϕ, ξ, density (ρ) and the Voigt average isotropic P-wave (V_P) and S-wave (V_S) velocities in six lower mantle layers (Section 5.3.1). The deepest layer represents the D" region, which is well-known to be anisotropic (see Nowacki et al. (2011); Chang et al. (2014) for reviews) and for which our results are in agreement with previous studies. The remainder of the lower mantle, which extends from approximately 670 to 2741 km depth, is represented by five layers of roughly equal thickness. We focus on the elasticity in this part of the mantle and show that this region is anisotropic as well. Note that all other parameters in the model, i.e. parameters describing bulk and shear attenuation, core and upper mantle structure, are also allowed to vary in our prior model distribution (Figure A.1).

6.2 Results

6.2.1 Inferences on lower mantle structure

Before analysing any marginal, we assess the MDNs with synthetic test data that were not used during the training process (not shown here, see for instance Figure 4.2). Radial (1-D) seismological models, such as the Preliminary Reference Earth Model (*PREM*, (Dziewoński and Anderson, 1981)), are commonly used as a reference for 3-D tomographic models or to constrain the Earth's radial temperature and composition profiles (Kennett, 2006; Cobden et al., 2009; Cammarano et al., 2011). We also apply the trained MDNs to synthetic data for *PREM*, which are not used to train the networks. We find that the MDNs made accurate predictions for the test samples and for *PREM*, for which all parameters lay within one standard deviation of the most probable value in the marginals (Figure 6.1). As an additional measure of robustness, we quantify the constraint provided by the data for each target parameter (Section 5.3.2). This measure of information content, or gain, indicates that the six seismic parameters in the five lower mantle layers are well-resolved ($D_{KL} > 3.7$ bits), except for η in the shallowest layer (LM_1), which is moderately constrained ($D_{KL} = 1.3$ bits, see the results for the joint spheroidal and toroidal mode data set (S_{II}) in Table 5.1).

The accurate predictions for *PREM* and the information gain indicate that network training was successful and that we can apply the MDNs to the observed normal mode data (Figure 6.2). The most prominent feature is the small, yet robust, η anisotropy in the deeper layers (below ∼1900 km). Our observation agrees with Montagner and Kennett (1996), who found negative anisotropy ($\eta < 1$) below ∼2000 km, with a maximum amplitude of 1.5–2.0%, jumping to a positive anisotropy in the D" region. However, these authors assumed the lower mantle anisotropy to be insignificant, due to its relatively low amplitude and a lack of uncertainty analysis. This highlights the advantage of our Bayesian approach, which allowed us to assess the significance of the observed anisotropy. The probability of negative anisotropy in the two layers between 1884 and 2741 km is very high (≥ 0.96) and the most probable values (the peaks of the two marginals) correspond to 0.9–1.0% of negative anisotropy (Table 6.1). The probability that this anisotropy is stronger than 0.5% is 0.91 (1884–2313 km) and 0.78 (2313–2741 km). This is, to our knowledge, the first robust observation of anisotropy in this part of the mantle and contrasts with the consensus that the lower mantle is isotropic.

As noted above, our method produces marginal distributions for individual model parameters and does not output a complete earth model, which could be used to calculate the overall fit to the data. Nevertheless, we conduct a simple misfit-based analysis to verify that the inferred η anisotropy in the lower mantle is indeed more compatible with the centre frequency measurements than an isotropic lower mantle. We consider *PREM* and set its isotropic lower mantle structure for η to the most likely values, i.e. the peaks of the marginals, for η in our five lower mantle layers. We use

an L_2-norm to calculate the data misfit $\psi(\mathbf{m})$ for a given earth model $\mathbf{m}$ as

$$\psi(\mathbf{m}) = \sum_{i=1}^{N} \frac{(d_i - d_i^{syn}(\mathbf{m}))^2}{\sigma_i^2}, \tag{6.1}$$

where d_i is one of the N measured spheroidal and toroidal centre frequencies, $d_i^{syn}(\mathbf{m})$ is the synthetic centre frequency computed for an earth model $\mathbf{m}$ and σ_i is the corresponding estimate of measurement uncertainty. We find that the misfit is ∼9% lower for the *PREM* model with the updated η structure, indicating that the data are better explained by a lower mantle that is anisotropic in η. Note that this misfit analysis forms an additional test, which is meant to verify the robustness of our results for η, and is not part of our inversion approach.

We also find a preference for negative, albeit very weak, P-wave anisotropy ($\phi < 1$), with probability around 0.8 throughout most of the lower mantle (Figure 6.2). No significant shear-wave anisotropy (ξ) was observed, in agreement with previous studies. Further, there is a clear positive density anomaly in the two layers below ∼1900 km: the peaks of the marginals indicate densities 0.3% and 0.7% higher than in *PREM*. This supports an earlier hypothesis (Kellogg et al., 1999) that an average excess density exists in the bottom ∼500–1000 km of the mantle, which could not unambiguously be determined in earlier studies (Masters and Gubbins, 2003; de Wit et al., 2014). The observed deviations from *PREM* in both V_S and V_P match the 1-D reference model *ak135* (Kennett et al., 1995) very closely. Since *ak135* was constructed using body-wave travel time measurements, we infer that our results for the isotropic P- and S-wave velocities are compatible with both normal mode and travel time data.

We do not impose scaling between the anisotropic parameters, but we can extract such scaling relations from our results and compare them to commonly-used negative scaling factors in the literature (Montagner and Anderson, 1989; Panning and Romanowicz, 2006). For each layer, we draw 10 000 random numbers from the marginals for η, ϕ and ξ and compute the corresponding ratios $d\ln\eta/d\ln\xi$ and $d\ln\phi/d\ln\xi$. The resulting distributions of the ratios allow us to calculate the probability that the scaling is in a certain range. We find that a significant fraction of the probability corresponds to positive scaling relations: for the six layers, $p(d\ln\eta/d\ln\xi > 0)$ ranges between 0.23 and 0.56, while $p(d\ln\phi/d\ln\xi > 0)$ varies between 0.33 and 0.71 (Table 6.2). This suggests that care must be taken when fixed (negative) scaling relations are used to construct (seismological) mantle models. The sign of the scaling relation obviously imposes a strong assumption on the underlying mechanism for anisotropy. We believe that the scaling used in existing studies is responsible for not detecting any lower mantle anisotropy to date. In general, seismic data are mostly sensitive to ξ; thus, employing the wrong scaling to ϕ and η will not reveal their anisotropy given that ξ shows no sign of anisotropy.

6.2.2 Constraints on thermochemical structure

An isotropic lower mantle has been the underlying assumption of most models of Earth's composition and dynamics. If, by contrast, the lower mantle is slightly anisotropic,

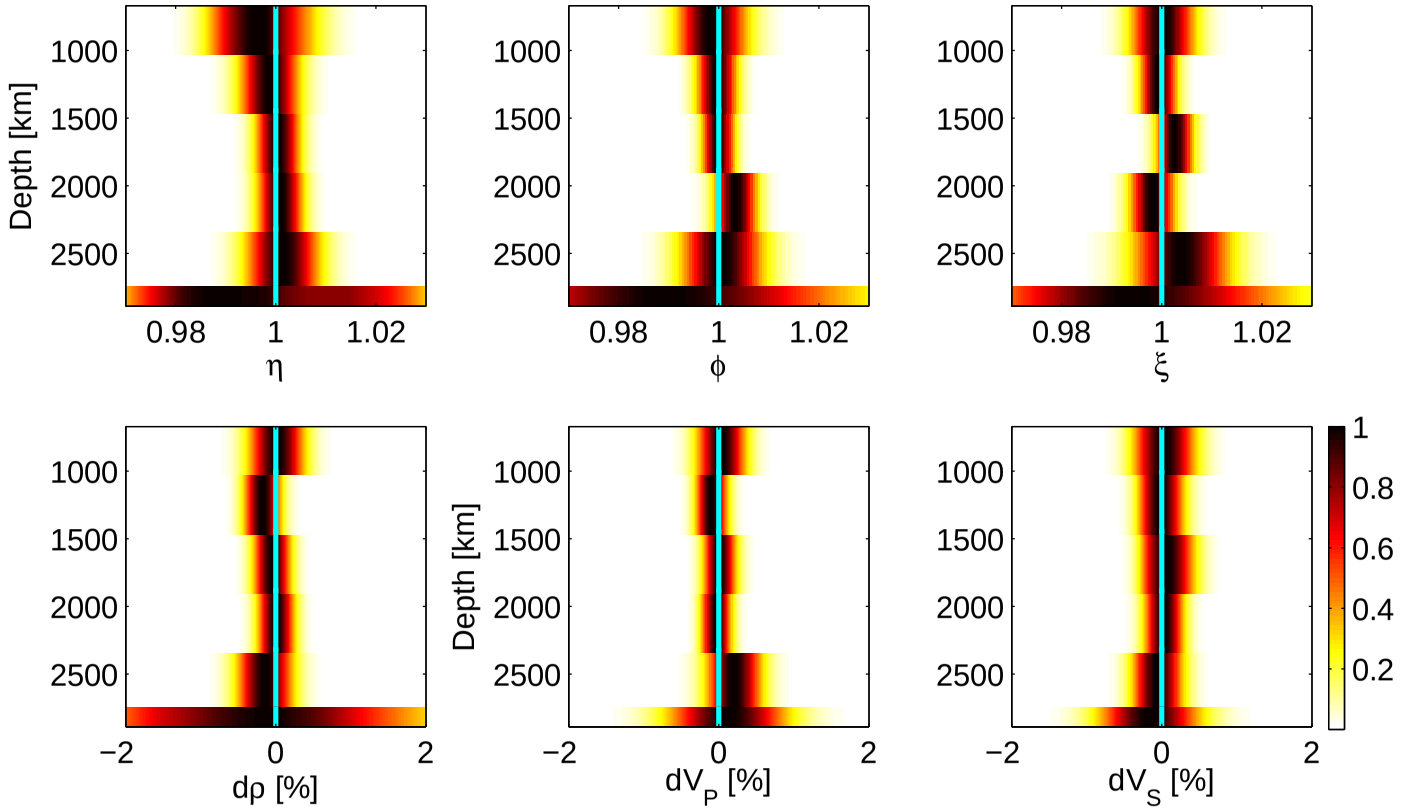


Figure 6.1: 1-D marginal posterior pdfs for the inversion of synthetic *PREM* data for the averages of the six seismic parameters in the six lower mantle layers (Section 5.3.1). The bottom layer in each panel represents the D" region. *PREM* (cyan line) is isotropic in the lower mantle and is given as a reference. The velocities and density are expressed as percentage deviations with respect to *PREM*. The probability for each 1-D pdf is rescaled so that the maximum equals 1. Asymmetric 1σ and 2σ error bars correspond to the $1/e^{1/2}$ (0.61) and $1/e^{2}$ (0.14) contours, respectively. For all parameters, *PREM* lies within 1σ of the most probable value (the peak of the pdf).

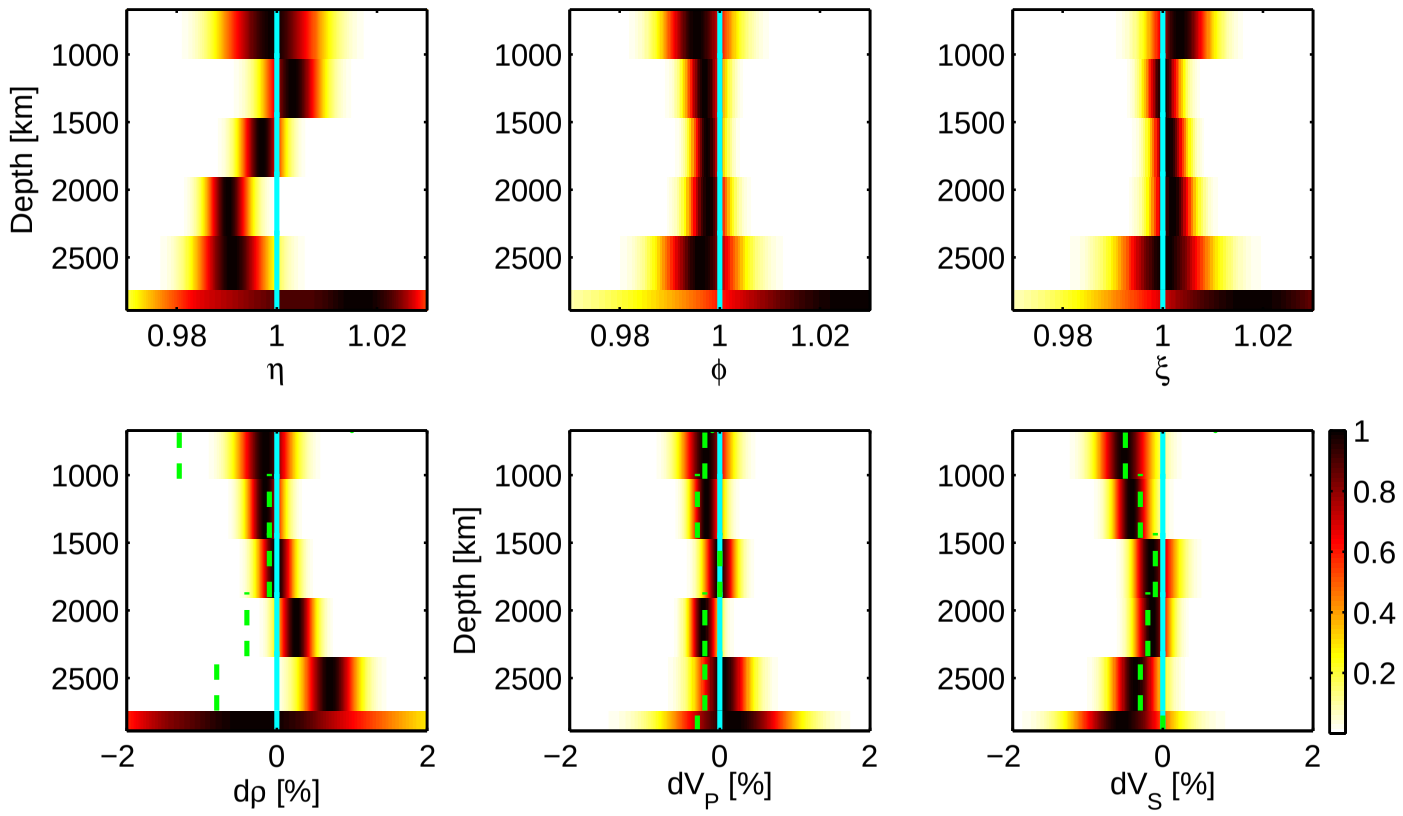


Figure 6.2: 1-D marginal posterior pdfs for the averages of the six seismic parameters in the six lower mantle layers (Section 5.3.1). The bottom layer in each panel represents the D'' region. *PREM* (cyan line) is isotropic in the lower mantle and is given as a reference. The velocities and density are expressed as percentage deviations from *PREM*, as is the model *ak135f* (Kennett et al. (1995); Montagner and Kennett (1996), green-dashed line). The probability for each 1-D pdf is rescaled so that the maximum equals 1. Asymmetric 1σ and 2σ error bars correspond to the $1/e^{1/2}$ (0.61) and $1/e^{2}$ (0.14) contours, respectively.

Table 6.1: Probability of the anisotropy b in the average η, ϕ and ξ being negative, i.e. $p(b < 1)$, and stronger than -0.5%, i.e. $p(b < 0.995)$, for the five lower mantle layers and the D" region. Note that the depths of discontinuities and the layer boundaries are allowed to vary between earth models (Appendix A).

Region	Depth [km]	$p(b<1)$			$p(b<0.995)$		
		$\bar{\eta}$	$\bar{\phi}$	$\bar{\xi}$	$\bar{\eta}$	$\bar{\phi}$	$\bar{\xi}$
LM_I	670 –1027	0.56	0.82	0.23	0.28	0.48	0.03
LM_{II}	1027 –1456	0.22	0.87	0.51	0.02	0.24	0.03
LM_{III}	1456 –1884	0.85	0.85	0.33	0.25	0.14	0.01
LM_{IV}	1884 –2313	1.00	0.79	0.23	0.91	0.19	0.01
LM_V	2313 –2741	0.96	0.73	0.46	0.78	0.38	0.21
D"	2741 –2891	0.40	0.15	0.20	0.31	0.11	0.14

Table 6.2: Probability that the scaling relations $d\ln\eta/d\ln\xi$ and $d\ln\phi/d\ln\xi$ are positive in the five lower mantle layers and the D" region, in contrast to the negative values commonly assumed in the literature, e.g. $d\ln\eta/d\ln\xi = -2.5$ and $d\ln\phi/d\ln\xi = -1.5$ (Panning and Romanowicz, 2006). Note that the depths of discontinuities and the layer boundaries are allowed to vary between earth models (Appendix A).

Region	Depth [km]	$p(\frac{d\ln\eta}{d\ln\xi} > 0)$	$p(\frac{d\ln\phi}{d\ln\xi} > 0)$
LM_I	670 –1027	0.47	0.33
LM_{II}	1027 –1456	0.49	0.51
LM_{III}	1456 –1884	0.39	0.38
LM_{IV}	1884 –2313	0.23	0.34
LM_V	2313 –2741	0.46	0.48
D"	2741 –2891	0.56	0.71

it is crucial to understand its potential effect. We investigate whether the observed elasticity can be explained by a simple lower mantle model, given the currently available estimates of elasticity derived from mineral physics. We restrict our analysis to a polycrystal aggregate of iron-bearing perovskite and ferropericlase (Appendix B). Laboratory and first-principles modelling studies show that both minerals are anisotropic under lower mantle conditions (Karki et al., 1997, 2000; Oganov et al., 2001; Wentzcovitch et al., 2004). We vary the fractions of perovskite (X_{Pv}) and iron (X_{Fe}), the iron partitioning coefficient (K_D) and the temperature and considered all possible combinations of these parameters (Table B.1), resulting in a total of 22 491 different thermochemical models. Further, we rotate both minerals individually and imposed radial anisotropy, or vertical transverse isotropy, on the resulting polycrystal (Walker and Wookey, 2012). The corresponding elasticity tensor is hexagonally symmetric and it is straightforward to extract η, ϕ, ξ, ρ, V_P and V_S (Mainprice, 2007).

We require the polycrystal aggregate to simultaneously fit our marginals for η, ϕ, ξ, density and the Voigt average V_P and V_S. For each of the 22 491 thermochemical models, we compare the six parameters with the (asymmetric) 2σ error levels in the six marginals for each of the five layers in the lower mantle (Figure 6.2). Whenever one of the six parameters for the polycrystal aggregate lay outside the $\pm 2\sigma$ range, the corresponding thermochemical model is discarded.

A first key observation is that we can in fact find thermochemical models, and associated crystal orientations, that fit all seismic observations simultaneously. We find a strong constraint on the composition and the temperature (Figure 6.3), which are mainly determined by the isotropic velocities and density. Perovskite content is primarily sensitive to V_P structure (Deschamps et al., 2007), which we observed to be similar to *PREM* (Figure 6.2). For all five layers the perovskite content has to be lower than $\sim$75%, in agreement with Deschamps and Trampert (2004); Verhoeven et al. (2009), who used *PREM* velocities, but contradicting piclogitic models of the lower mantle, which have $\sim$90% perovskite (Murakami et al., 2012). The iron content is higher than 10% in all layers and increases with depth: in the deepest layer the range of accepted X_{Fe} is 14–19%. This relates to the elevated density in the deep lower mantle (Figure 6.2) and explains the difference with Deschamps and Trampert (2004); Verhoeven et al. (2009), who used the density of *PREM*. In the top four layers, all accepted thermochemical models have temperatures similar to or up to 200 K lower than the Brown-Shankland geotherm (Brown and Shankland, 1981) (Figure 6.3). The correlation between temperature and V_S in the deep lower mantle ($>$ 2000 km depth) is low compared to that between temperature and density (Trampert et al., 2004). Therefore, the relatively low temperatures that we inferred here are likely related to the positive deviations in density in the deep mantle. We can fit the seismic observations for all three values considered for the partitioning factor K_D (Table B.1); we only show results for $K_D = 0.3$, which is a value typically assumed in the literature for aluminium-free systems (Deschamps and Trampert, 2004; Kobayashi et al., 2005; Lin et al., 2013).

The required rotations of the perovskite and ferropericlase crystals, which are mainly determined by the anisotropic parameters, indicate a tight constraint on the preferred orientation of the orthorhombic perovskite crystal (Figure 6.4). Independent of depth, the perovskite crystal needs to be rotated about the two horizontal principal axes (x_1 and x_2) by approximately 30–40 degrees to match the marginals inferred for η, ϕ and ξ. There is more freedom in the rotation of the ferropericlase crystal, due to its higher degree of symmetry (cubic) compared to that of perovskite (orthorhombic). A key observation is the difference between the accepted rotation angles for the two minerals. To explain the observed seismic anisotropy, the two minerals have to be rotated individually, i.e. about different angles, prior to the construction of the polycrystal aggregate. The anisotropy cannot be explained by rotating the polycrystal, i.e. by rotating the two minerals together. For ease of comparison, we also visualise the accepted orientations using the more conventional Bunge Euler angles ((Bunge, 1982), Figures B.2 and B.3).

Of course, we cannot hope to fully understand the geodynamic implications of

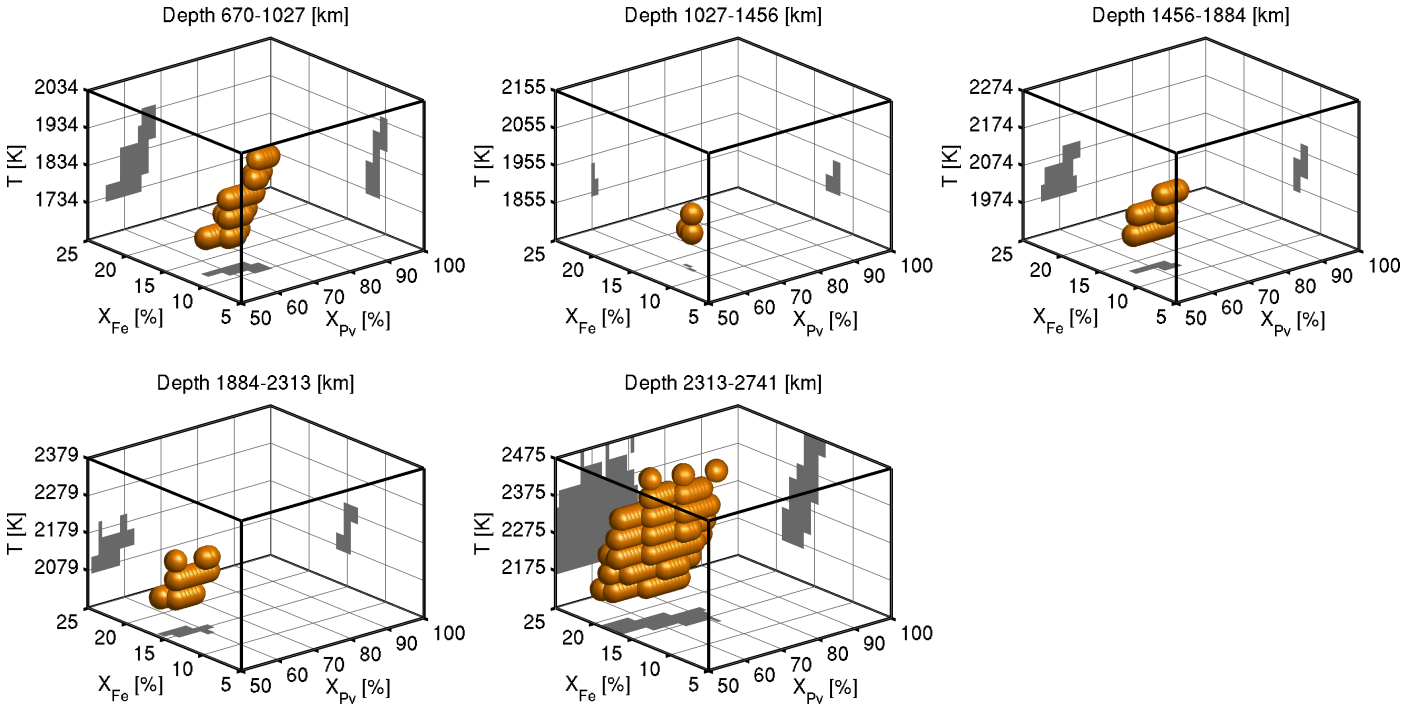


Figure 6.3: Constraints on perovskite content (X_{Pv}), iron content (X_{Fe}) and temperature (T) for $K_D = 0.3$ (Table B.1). The thermochemical models that fit all six seismic parameters (η, ϕ, ξ, ρ, V_P and V_S) within their uncertainties are shown in Dutch-orange spheres (voxels) and projected in grey. The five boxes represent the five lower mantle layers, with the corresponding depth range given above each box.

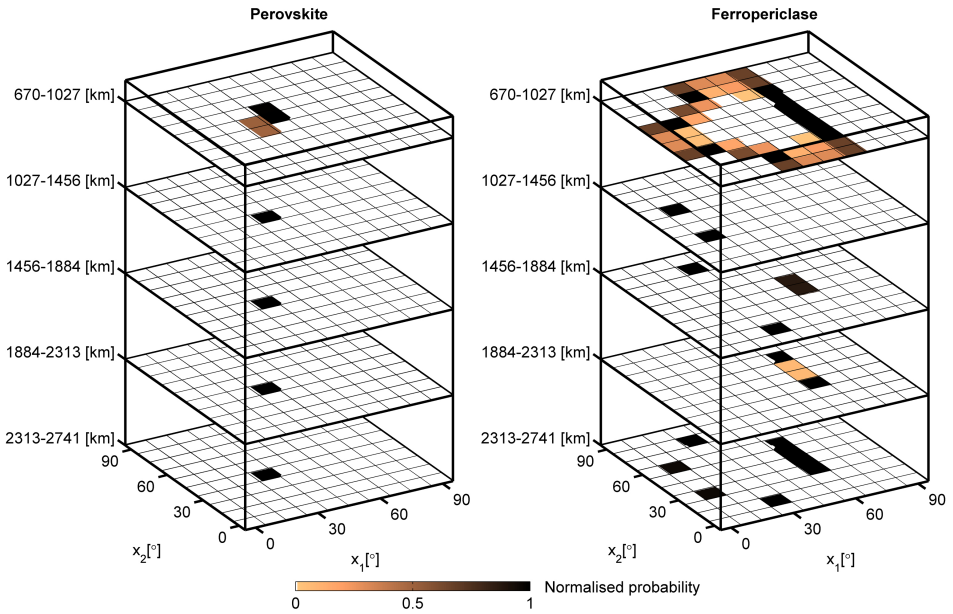


Figure 6.4: Constraints on the orientation of the perovskite and ferropericlase crystals, represented by 2-D histograms of rotation angles for all accepted thermochemical models in the five lower mantle layers. In each panel, the rotation angle about two principal (horizontal) axes of the elastic tensors (x_1 and x_2) is shown for the orthorhombic perovskite (left) and cubic ferropericlase (right). The order of rotation is $x_2 - x_1(-x_3)$, where x_3 represents the vertical axis, over which we averaged to impose radial anisotropy (Section B). Each of the ten 2-D histograms is normalised, so that the colour indicates the relative number of accepted rotations. Empty cells represent rotation angles for which none of the 22 491 thermochemical models fits all six seismic parameters within their uncertainties.

lower mantle anisotropy from a 1-D radial model, and an extension to the 3-D case is required. However, since the strongest anisotropy is in η, and not in the expected parameter ξ, it is instructive to investigate if existing mineral physics data are compatible with such an observation. Furthermore, our simple polycrystal model implicitly assumes that all perovskite and ferropericlase crystals in the lower mantle are aligned with the single rotated perovskite and ferropericlase crystals in our polycrystal, respectively. This corresponds to strong crystallographic texturing and may not represent the situation in the Earth's interior, in which individual crystals are not necessarily 100% aligned. However, we emphasise that we only intended to perform a first-order analysis which should be seen as a proof of concept. The simple thermochemical lower mantle model we inferred, and the associated preferred crystal orientation, merely provide a first of possibly many explanations for our seismic observations.

6.3 Conclusions

In summary, seismic normal mode data can constrain composition, temperature and the preferred orientation of mantle-forming crystals by combining the inferences for the six seismic parameters describing the average lower mantle. We successfully addressed common issues in previous studies of seismic anisotropy by using novel Bayesian machine learning techniques. We studied the radially averaged η, ϕ, ξ, ρ and the Voigt average isotropic V_P and V_S in six lower mantle layers. We showed, for the first time, that the average lower mantle is anisotropic below 1900 km depth. This robust observation of seismic anisotropy challenges the consensus that the bulk of the lower mantle is isotropic. In addition, the mass density shows a well-constrained positive deviation from existing models at the same depths. The 1-D marginals for V_S and V_P match *ak135* closely. Since *ak135* was constructed using body-wave travel time data, we conclude that our inferences for the isotropic P-wave and S-wave velocities are compatible with both normal mode and travel time data.

Given the currently available mineral physics data, we showed that the observed anisotropy can be the result of LPO, which is most likely caused by deformation through dislocation creep (Karato, 2008). So-called deformation mechanism maps (Karato, 1998) for perovskite and MgO suggest grain sizes of at most 10^{-2}–1 mm for deformation by diffusion creep under typical lower mantle stress levels; in the case of dislocation creep the grain sizes can be expected to be larger, i.e. on the order of 10^{-1}–10 mm as a minimum. Therefore, seismic anisotropy, such as observed here, should provide constraints on the grain size of minerals in the lower mantle and open a new window to constrain the flow in the deep mantle. However, meaningful geodynamic interpretations can only be made in a full 3-D analysis.

Acknowledgements

We are grateful to Hanneke Paulssen for comments on an early draft of the manuscript. We thank all people who made their normal mode measurements available in various publications and on the REM webpages, and thank Paula Koelemeijer for providing additional normal mode data. Ralph de Wit is funded by the Netherlands Organisation for Scientific Research (NWO) under the grant ALW Top-subsidy 854.10.002. Computational resources for this work were provided by the Netherlands Research Center for Integrated Solid Earth Science (ISES 3.2.5 High End Scientific Computation Resources). James Wookey and Andrew Walker developed the *MSAT* toolbox used to handle elastic tensors.

7

Evidence for radial anisotropy in Earth's inner core from normal modes

Abstract

Seismologists have established that the inner core is seismically anisotropic, but ambiguities remain in the details of its structure. Proof of radial anisotropy in the top of the inner core can help us discriminate between existing dynamic models and thereby improve our understanding of core dynamics and solidification processes. We assess the likelihood of radial anisotropy in the inner core in a spherically symmetric earth model using a joint data set of spheroidal and radial mode centre frequencies. In the top ∼115 km of the inner core, our inferences are compatible with radial anisotropy for which $V_{PV} > V_{PH}$, which could provide constraints on the dynamics of the shallow part of the inner core, but we cannot rule out an isotropic structure in this region. Further, we find strong positive anisotropy in ϕ in the deep inner core, increasing with depth, while the average density is likely lower than in both *PREM* and *ak135f*. Tradeoffs between parameters (V_P and ϕ) in the deep inner core prevent us from making unambiguous inferences on the radial anisotropy. Given the nature of the earth model parametrisation used here, two options exist. The inferred radial anisotropy either (i) is due to an inner core that is truly radially anisotropic or (ii) relates to the signal of cylindrical anisotropy in the degree-zero measurements. At this stage, we can only conclude that the centre frequencies require anisotropy in the (deep) inner core, in agreement with previous studies, and may provide further evidence for a distinct innermost inner core.

7.1 Introduction

Seismologists have unambiguously established that the inner core is seismically anisotropic, but the complete details of the structure are yet to be filled in. The first evidence for inner core anisotropy was given by PKIKP travel time residuals (Poupinet et al., 1983; Morelli et al., 1986) and the anomalous splitting of free oscillations (Woodhouse et al., 1986), who proposed an axisymmetric anisotropy with the symmetry axis aligned with the Earth's rotation axis. The average P-wave velocity (V_P) along this direction was inferred to be 1% faster than in the equatorial plane. Subsequent studies reinforced the hypothesis of inner core anisotropy with cylindrical symmetry, although there is still disagreement on its amplitude and variation with depth (see Tkalčić and Kennett (2008); Deuss (2014) for reviews). Tromp (1993) showed that such a cylindrical anisotropy could explain the anomalous splitting of most spheroidal modes available at the time, whilst being compatible with PKIKP travel time residuals. Further studies on PKP travel times, e.g. Creager (1992); Song and Helmberger (1993), established the view that the anisotropy was strong ($\sim$ 3%) and restricted to the top 300 km of the inner core.

By contrast, anisotropy was observed throughout the inner core, e.g. Vinnik et al. (1994), with the strongest anisotropy ($>$ 3%) in the innermost 300 km of the inner core (Su and Dziewonski, 1995). Using a model space search and free oscillation data, Beghein and Trampert (2003) found anisotropy throughout the inner core, consistent with body wave travel times, and suggested that the regularisation applied in previous inversions caused the apparent lack of anisotropy near the centre of the Earth. A change at depth in the slow direction of PKIKP travel times was explained by the existence of an innermost inner core, in which the anisotropy is stronger and the fast direction is no longer parallel to the Earth's rotation axis, e.g. Ishii and Dziewoński (2002); Cao and Romanowicz (2007). However, no consensus has been reached on its exact structure, as evidenced by the variation in estimates of its radius ($300 - 500$ km) and the angle of the slow direction relative to the Earth's rotation axis ($45 - 55^\circ$). Recently, Wang et al. (2015) analysed the autocorrelation of earthquake coda and proposed an innermost inner core in which the fast axis of the anisotropy aligns with the equatorial plane. Studies of inner core attenuation show similar evidence for a distinct innermost structure, e.g. Li and Cormier (2002); Cormier and Stroujkova (2005) and Figure 4.6 in this manuscript.

The upper $\sim 50 - 80$ km of the inner core has been proposed to be isotropic in an attempt to explain travel time measurements of inner-core sensitive body wave phases, e.g. Shearer (1994); Song and Helmberger (1995). Irving and Deuss (2011) found normal mode observations to be compatible with such an isotropic layer directly below the ICB. However, a recent reassessment of free oscillation centre frequencies may provide evidence that the top of the inner core is indeed anisotropic (K. Lythgoe and A. Deuss, personal communication, 2014). If present, the anisotropy in this region may be radial in nature, in contrast to the now well-established cylindrical symmetry in the deeper inner core. Interestingly, a spherically symmetric structure with radial

anisotropy in the outermost inner core would not be incompatible with PKIKP travel time data, since radial anisotropy does not cause a dependency on ray direction for such measurements. Consequently, radial anisotropy in the top of Earth's inner core can only be assessed using normal modes.

Seismic anisotropy originates from lattice-preffered orientation (LPO), the alignment of intrinsically anisotropic minerals, or shape-preferred orientation (SPO) of mineral crystals, e.g. (Karato, 2008). LPO is a likely candidate for inner core anisotropy, as the crystals of its main constituent, iron, are elastically anisotropic under core conditions (Stixrude and Cohen, 1995; Vočadlo et al., 2003; Tateno et al., 2010). Several models for inner core structure and the alignment of iron crystals have been proposed (see Deguen (2012) for a review). These include aspherical inner core growth (Yoshida et al., 1996; Deguen et al., 2011), internal flow driven by the Maxwell stress of Earth's magnetic field (Karato, 1999; Buffett and Wenk, 2001), internal convection due to thermochemical variations (Jeanloz and Wenk, 1988) and solidification texturing, aligned with the Earth's magnetic field (Karato, 1993) or with flow in the outer core (Bergman, 1997).

Solidification texturing is associated with the growth of the inner core due to the solidification of outer core material at the inner core boundary (ICB). As the direction of maximum heat flux in the core is radial, this model suggests that crystals align in the radial direction, which would result in a (spherically symmetric) radially anisotropic structure. The only other model that allows for radial anisotropy in the top of the inner core is the model proposed by Yoshida et al. (1996). In this model, heterogeneous growth, or more specifically, preferential growth in equatorial regions, causes relaxation of the ICB topography. Other geodynamical models do not allow for radial anisotropy in the outermost inner core. Thus, proof of radial anisotropy in this region can help us discern between existing dynamic models and thereby improve our understanding of core dynamics and solidification processes.

In this study, we assess the likelihood of radial anisotropy in both the shallow and deep inner core in a spherically symmetric earth model. Existing 1-D reference models, such as *PREM* (Dziewoński and Anderson, 1981) and *ak135f* (Kennett et al., 1995; Montagner and Kennett, 1996), contain an isotropic inner core structure. We use self-coupled spheroidal (Deuss et al., 2013; Koelemeijer et al., 2013; Koelemeijer, 2014) and radial mode centre frequency measurements (Masters, unpublished) and invert for the average velocity, density and parameters describing radial anisotropy in layers of varying thickness. Similar to de Wit et al. (2014), we adopt a radial (1-D) parametrisation and use the Mineos package (Masters et al., 2011) to calculate normal mode centre frequencies in a spherically symmetric Earth. As such, our analysis of inner core anisotropy is limited to degree-zero structure and we cannot assess recent hypotheses regarding hemispherical (degree-one) inner core structure, e.g. Deuss et al. (2010); Waszek et al. (2011); Lythgoe et al. (2014).

First, we perform a resolution analysis for parameters in the upper $\sim$300 km of the inner core and investigate the minimum layer thickness that is required for the free oscillation data to constrain the velocity, density and anisotropic parameters. Second, we infer the radial averages for these parameters in the top $\sim$115 km and assess the

probability of radial anisotropy in this region. Third, we invert for the average radial structure in the whole inner core, as represented by four layers of ∼300 km thickness. Finally, we construct 2-D marginal posterior pdfs to investigate trade-offs between parameters, similar to earlier analyses (Figures 3.12 and 4.9).

7.2 Setup

7.2.1 Normal mode data

The data consist of centre, or degenerate, frequencies obtained from self-coupled splitting function measurements for both spheroidal and radial modes. We invert the same 184 self-coupled spheroidal mode centre frequencies as were used in Chapters 4 to 6 (Figure 5.1), which measured by Deuss et al. (2013); Koelemeijer et al. (2013); Koelemeijer (2014). Radial modes are only sensitive to radial motion and correspond to an angular order $l = 0$. Radial mode centre frequencies have been measured by He and Tromp (1996) and by Masters (unpublished), who used a singlet stripping technique, and are available from the Reference Earth Model (REM) web pages (http://igppweb.ucsd.edu/~gabi/rem.dir/surface/tab.rf). Masters and Gubbins (2003) used the latter data set in combination with spheroidal modes to assess the Earth's density, for which radial modes are well-suited due to their relatively large sensitivity to density in the deep Earth (Dahlen and Tromp, 1998). A total of 14 radial mode measurements are available on the REM web pages. We exclude radial modes ${}_{12}S_0$, ${}_{13}S_0$, ${}_{14}S_0$ and ${}_{19}S_0$, since these were not measured by He and Tromp (1996) and their respective error estimates are relatively large, and ${}_{0}S_0$, since this mode has little sensitivity to the top (∼100 km) of the inner core (K. Lythgoe, personal communication, 2014). To obtain a better constraint on (anisotropic) parameters, it would be necessary to include toroidal modes, which exhibit SH-motion. However, inner-core–sensitive toroidal modes are extremely difficult to excite, let alone observe at the surface, although they could be excited due to cross-coupling (resonance) with spheroidal modes (Dziewoński and Woodhouse, Dziewoński and Woodhouse). For a spherically symmetric non-rotating elastic isotropic (SNREI) Earth, they are confined to theory (Dahlen and Tromp, 1998); the energy required to excite such modes cannot be generated by earthquakes in the crust or upper mantle, nor can they be recorded by seismometers along the Earth's surface. Therefore, our total data set comprises 193 centre frequency measurements for 184 spheroidal and 9 radial modes.

7.2.2 Synthetic data

Exact frequencies for the 193 normal modes are calculated using the Mineos package (Masters et al., 2011) for 100 000 synthetic 1-D earth models. As in Chapters 5 and 6, we use the fully anisotropic parametrisation, i.e. we allow for radial anisotropy in the inner core and whole mantle, and vary all parameters including discontinuity depths and attenuation (Appendix A). Self-gravitation is taken into account for frequencies

below 30 mHz and a reference period of 1 s was used for the dispersion correction due to attenuation. As in previous chapters, we add zero-mean Gaussian noise to the synthetic data; for the standard deviation we use the uncertainty estimate for each measurement, as estimated for the spheroidal modes (Deuss et al., 2013; Koelemeijer et al., 2013; Koelemeijer, 2014) and for the radial modes (http://igppweb.ucsd.edu/~gabi/rem.dir/surface/tab.rf).

7.2.3 Network configuration

We use a similar configuration as in Chapters 4 to 6 and train Mixture Density Networks (MDNs) to obtain 1-D marginal posterior probability density functions (pdfs) for earth model parameters (Section 2.4.4). The MDNs consist of a 193-D input vector, 50 hidden units and a mixture of 15 Gaussian kernels as output, which results in 11 995 free parameters per network (Equation 4.1). We use early stopping during network training, i.e. training is halted for each ensemble member when the error of a separate validation set reaches a minimum, and construct an ensemble of 48 MDNs for each target parameter (Section 2.4.8). The networks are trained using the Scaled Conjugate Gradient (SCG) algorithm Møller (1993) for a maximum of 5000 iterations. 80% of the 100 000 patterns in the synthetic data set is used for training, 15% for the validation set and the remaining 5% for the test set. To improve the generalisation capacity of the ensemble, we divide the synthetic data randomly over the training, validation and test sets for each of the 48 ensemble members.

7.3 Results

7.3.1 Information gain and resolving power

First, we assess the ability of the spheroidal and radial mode data to resolve the structure in the top of the inner core. We use the Kullback-Leibler divergence to measure the information content, which enables us to quantify the constraint provided by the normal mode data (Section 2.5). We train MDNs on the radially averaged Voigt average isotropic velocities V_P and V_S, density (ρ) and the three parameters describing radial anisotropy (η, ϕ and ξ) in layers of increasing thickness (Appendix A). For an isotropic medium, $\eta = \phi = \xi = 1$. The upper bound of the layers is fixed to the ICB, while the depth of the lower bound is increased by including more grid points (knots), which are spaced roughly 38 km apart in the inner core parametrisation (Table A.1). This results in five layers that are respectively $\sim$38, 76, 115, 153 and 305 km thick. Note that the precise depth range of the layers changes with the allowed variation in the ICB depth in the prior model distribution (Appendix A). For each model, the depth of the lower boundary of the layer was determined by linearly interpolating between the centre of the Earth and the new ICB depth.

The average P-wave velocity (V_P) and anisotropy (ϕ) below the ICB are constrained for all layer thicknesses ($D_{KL} \geq 1.7$ bits, Figure 7.1). The constraint increases strongly

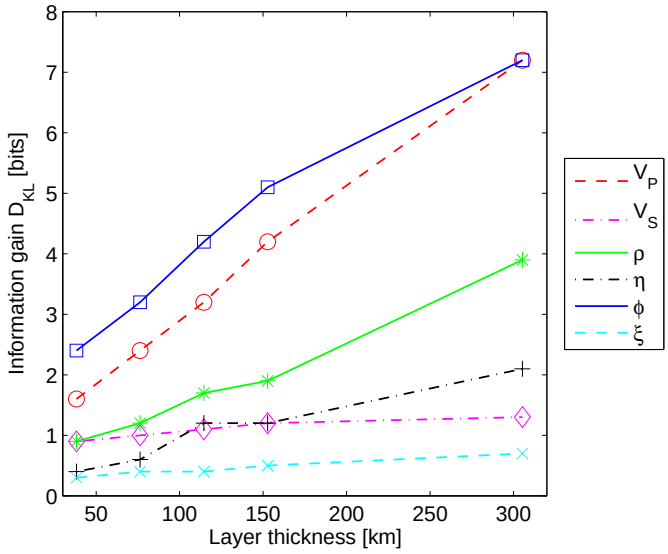


Figure 7.1: Information gain D_{KL} for the averages of all six physical parameters in layers of increasing thickness in the top ∼300 km of the inner core. The upper bound of each layer is formed by the ICB, while the depth of the lower bound is increased.

with layer thickness; we will discuss the cause(s) in Section 7.4.3. This effect is also observed for ρ and, albeit less pronounced, for V_S, η and ξ. The weak constraint on V_S and shear-wave anisotropy (ξ) is in agreement with the nature of the spheroidal and radial modes used here. Thus, to return to the question of resolving power: we can constrain V_P, ϕ and ρ, especially for layers thicker than 100 km. For V_S and η, such a thickness serves as a minimum, and even then we obtain only a weak to moderate constraint (D_{KL} ≈1–2 bits).

7.3.2 Inferences on the outermost inner core

Based on the results of our resolving power test, we assess the 1-D marginals posterior pdfs for the radial averages in the top ∼115 km of the inner core. The most prominent feature in the marginals are the negative deviation from *PREM* for ρ (Figure 7.2). The peak of the pdf, the Maximum A Posteriori (MAP) estimate θ, corresponds to an average density that is 1.2% lower than in *PREM*, while the asymmetric 2σ error bars span a range of -2.1–0.2% (Table 7.1). The marginal for ϕ indicates a slight preference for positive radial anisotropy with probability $p(\phi > 1) = 0.71$. This would suggest that the data prefer $V_{PV} > V_{PH}$, in agreement with another study of inner-core–sensitive radial and spheroidal modes (K. Lythgoe, personal communication, 2014). However, this result does not unequivocally prove the presence of radial anisotropy in the top of the inner core. Despite the high information gain for ϕ, the error estimate spans

a wide range ($\theta \pm 2\sigma$ =-1.5–2.6%, i.e. 0.985–1.026); as such, an isotropic structure or even negative anisotropy ($\phi < 1$) are compatible with the data. The same is true for η, for which $p(\eta > 1) = 0.64$. The marginal for V_S suggests that this parameter is smaller than in *PREM*, but we note that, by construction, the prior for the Voigt average V_S is relatively wide compared to for instance the prior for V_P (Figure 7.2), and the corresponding posterior pdf is also relatively broad ($\theta \pm 2\sigma$ =-2.0–1.4%).

The inferences on the discontinuity depths, as shown in Figures 4.2 and 4.8, indicated that the ICB may be deeper than in existing 1-D reference models. For completeness, we inverted the joint set of 184 spheroidal and 9 radial centre frequencies for the ICB depth. This yielded a marginal pdf similar to that for the anisotropic parametrisation and the spheroidal mode data set (Figure 4.8), i.e. a most likely ICB depth that is larger than in *PREM* and *ak135* (not shown here).

Another note on the number of synthetic samples

In Sections 4.6.5 and 5.4.1, we discussed whether or not the size of the training set (100 000 samples) was sufficient to extract most or all available information on the target parameters from the data. Although D_{KL} increased for a training set double in size (Table 5.6), we found that the corresponding 1-D marginals were similar for the two training sets. Here, we perform a similar analysis as in Section 5.4.1 to increase the prior sampling density. However, instead of increasing the number of synthetic samples, we decrease the size of the prior model space. Note that we have a rather conservative prior for ϕ and ξ; for instance, the prior range for ϕ is $\sim$0.9–1.1 (Figure 7.2). Since previous studies have not observed radial anisotropy in the top of the inner core, it is unlikely that anisotropy in ϕ would be as strong as $\pm$10%. At the ICB, the uniform prior distribution for η and the four velocity components (V_{PV}, V_{PH}, V_{SV} and V_{SH}) is $\pm$2% with respect to *PREM* (Table A.2). We reduced the uniform prior distribution for the four velocities at the ICB to $\pm$1% and constructed a new training set. As a result, the prior range for the radially averaged ϕ in the $\sim$115 km thick layer reduced to approximately 0.95–1.05. We trained MDNs on the same six parameters in the outermost inner core and compared the 1-D marginals for the two training sets (not shown here). Results were similar to the analysis in Section 5.4.1. The marginals for the training set with reduced prior range were (negligibly) narrower, but the bulk of the probability mass was assigned to similar values for the target parameters. Thus, again it seems that we have enough samples to extract the available information from the data. Simultaneously, this means that, given the normal mode data and our chosen earth model parametrisation, we cannot infer more on radial anisotropy in the top of the inner core than our marginals for ϕ (and η) are currently telling us (Figure 7.2).

7.3.3 Inferences on deep inner core structure

Given the strong constraint on V_P, ϕ and density in the outermost inner core, we investigate the ability of the normal mode data to constrain the deeper parts of the inner core. We invert for the radial averages of the six parameters in four layers of

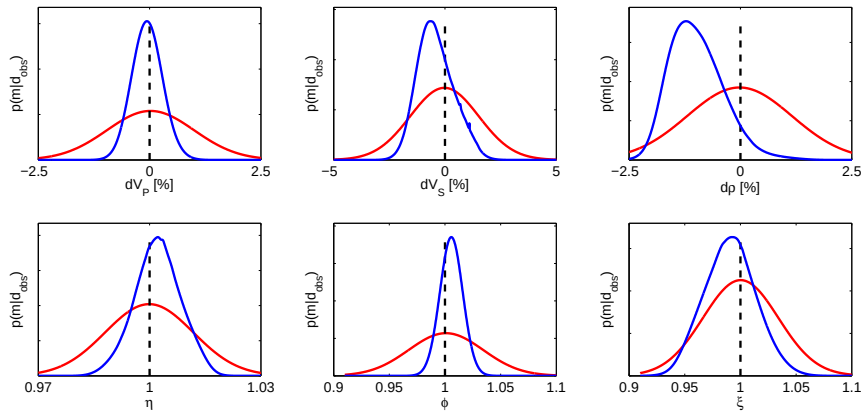


Figure 7.2: 1-D marginal posterior (blue) and prior (red) pdfs and *PREM* (black dashed line) for the averages of all six seismic parameters in the top ∼115 km of the inner core. The pdfs for velocities and density are expressed as percentage deviations from *PREM*.

equal thickness (∼ 305 km). Together, the four layers span the whole inner core and are denoted IC^4_{I-IV}, IC^4_I being the shallowest layer and corresponding to the 305 km thick layer considered in Figure 7.1. Again, we emphasise that we do not need to impose constant-value layers in the earth models, separated by fictitious discontinuities (see Section 5.3.1). The MDNs are trained directly on the radial averages in the four layers, while the underlying 1-D earth models in the training set are realistic, i.e. smooth, similar to for instance *PREM*. V_P and ϕ are resolved throughout the inner core ($D_{KL} \geq 3.9$ bits, Table 7.2). In general, the information content decreases with depth, with a minimum reached in the deepest layer for all six parameters. This agrees with the lack of sensitivity of normal modes to the centre of the Earth due to symmetry constraints. The 1-D marginals are shown in Figure 7.3 and in Table 7.2. The main results can be summarised as follows:

1. The data have a strong preference for a radially anisotropic inner core; the probability $p(\phi > 1)$, i.e. $p(V_{PV} > V_{PH})$, is 0.89 or higher for the top layer and the lower half of the inner core. Further, the strength of the radial anisotropy increases with depth. This result seems puzzling at first, as the deep inner core is assumed to be cylindrically anisotropic. We discuss this result further in Section 7.4.1.

2. The Voigt average V_P is likely larger than in *PREM* (and *ak135f*, which is very similar to *PREM*) and the deviation increases with depth.

3. All four layers show a negative deviation in density from *PREM*; the density is also probably lower than in *ak135f*.

4. η is moderately resolved in the upper half of the inner core ($D_{KL} \approx 2$ bits). For all four layers, the (peaks of the) pdfs are close to *PREM* (isotropy) and the

Table 7.1: Posterior statistics for the averages of six seismic parameters in the top ∼115 km of the the inner core (Figure 7.2), in terms of the MAP estimate θ and asymmetric 2σ model error bars, corresponding to $1/e^2$ levels in the unit normalised 1-D marginal posterior pdfs. All values are given as percentage deviations from *PREM*. Also given is the probability of positive anisotropy $p(a > 1)$ for each of the three anisotropic parameters.

IC_{top}	θ [%]	$\theta \pm 2\sigma$ [%]		$p(a > 1)$
$\bar{V}_P$	-0.1	-0.7	0.6	
$\bar{V}_S$	-0.7	-2.0	1.4	
$\bar{\rho}$	-1.2	-2.1	0.2	
$\bar{\eta}$	0.2	-1.0	1.5	0.64
$\bar{\phi}$	0.6	-1.5	2.6	0.71
$\bar{\xi}$	-0.7	-5.7	3.8	0.34

$\pm 2\sigma$ error bars allow for 1% of both negative and positive anisotropy. As before (Figure 7.1), V_S and ξ are weakly resolved and we thus do not consider these parameters any further.

7.4 Discussion

First, we discuss the strong P-wave anisotropy observed in the deep inner core and the relation to different styles of parametrisation. Further, we investigate possible trade-offs between ϕ and V_P (and density). Third, we go into more detail on the resolving power test shown in Figure 7.1.

7.4.1 Radial versus cylindrical anisotropy

The normal mode data prefer a degree-zero structure with radial P-wave anisotropy in the deep half of the inner core (Figure 7.3). The deep inner core is commonly assumed to be cylindrically anisotropic with the symmetry axis aligned with Earth's rotation axis. Such a model is fully captured by the zonal structure coefficients for the angular degrees 0, 2 and 4 (Tromp, 1995). Degrees 2 and 4 can be shown to linearly relate to three parameters describing the anisotropy, which are themselves defined in terms of the five Love parameters A, C, L, N and F, e.g. Tromp (1993); Irving et al. (2009). By contrast, the degree-zero coefficients, from which the centre frequency data are extracted, depend on all five Love coefficients. Since cylindrical anisotropy affects the degree-zero coefficients, it is perhaps unsurprising that we infer anisotropy throughout the inner core, as was observed earlier by for instance Vinnik et al. (1994); Su and Dziewonski (1995); Beghein and Trampert (2003). However, the nature of the

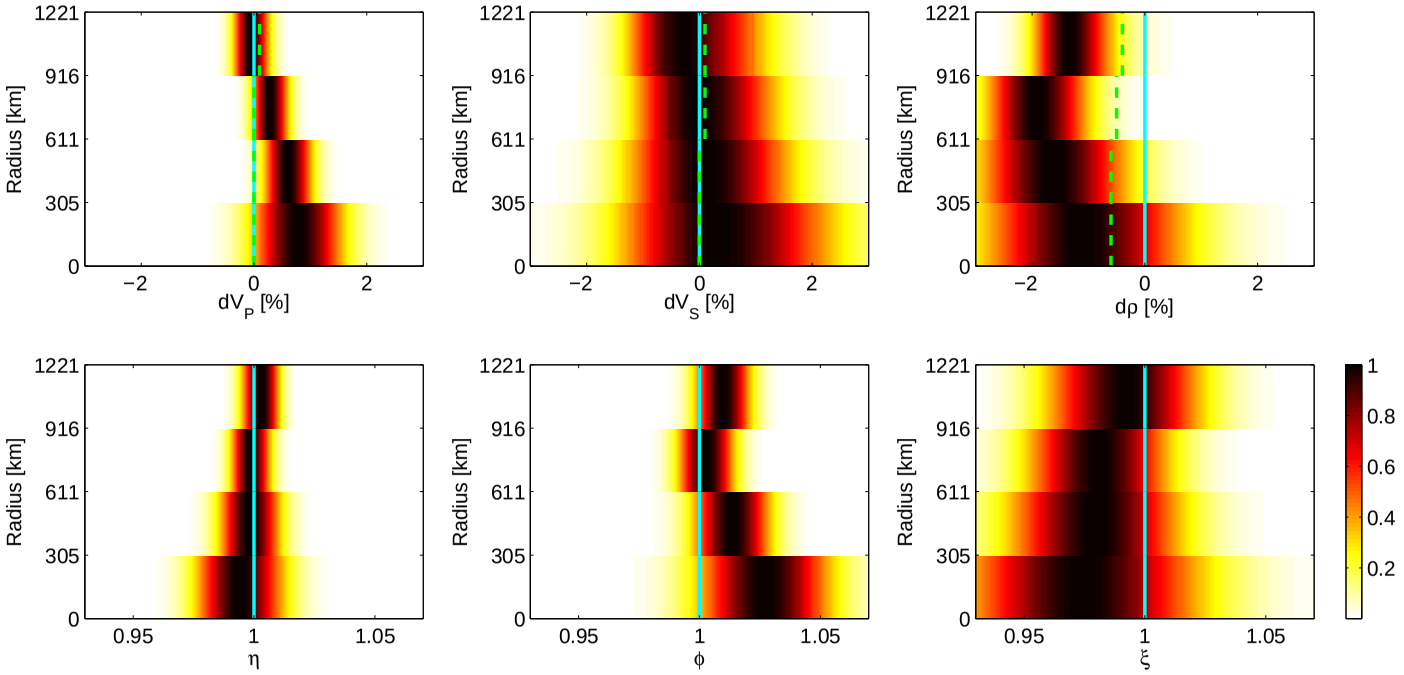


Figure 7.3: 1-D marginal posterior pdfs for the averages of all six seismic parameters in four layers in the inner core. The velocities and density are expressed as percentage deviations from *PREM*, as is the model *ak135f* (Kennett et al. (1995); Montagner and Kennett (1996), green-dashed line). The probability for each 1-D pdf is rescaled so that the maximum equals 1. Asymmetric 1σ and 2σ error bars correspond to the $1/e^{1/2}$ (0.61) and $1/e^{2}$ (0.14) contours, respectively.

Table 7.2: Posterior statistics for the radial averages of the Voigt V_P, V_S, density, η, ϕ and ξ in the inner core, represented by four layers, in terms of the MAP estimate θ and asymmetric 2σ model error bars, corresponding to $1/e^2$ levels in the unit normalised 1-D marginal posterior pdfs (Figure 7.3). Deviations are in percentage with respect to *PREM*. The fifth column lists the information gain D_{KL} in bits. The probability of positive anisotropy is shown in the last column for η, ϕ and ξ.

V_P	Radius [km]	θ [%]	$\theta \pm 2\sigma$ [%]		D_{KL} [bits]	
IC_I^4	916 – 1221	-0.1	-0.5	0.3	7.2	
IC_{II}^4	611 – 916	0.2	-0.2	0.7	7.9	
IC_{III}^4	305 – 611	0.6	0	1.2	7.3	
IC_{IV}^4	0 – 305	0.8	-0.3	1.9	4.1	
V_S	Radius [km]	θ [%]	$\theta \pm 2\sigma$ [%]		D_{KL} [bits]	
IC_I^4	916 – 1221	-0.4	-1.7	1.5	1.3	
IC_{II}^4	611 – 916	-0.1	-1.6	1.7	1.8	
IC_{III}^4	305 – 611	0.1	-1.7	2.1	1.5	
IC_{IV}^4	0 – 305	0.3	-2.0	3.1	1.0	
ρ	Radius [km]	θ [%]	$\theta \pm 2\sigma$ [%]		D_{KL} [bits]	
IC_I^4	916 – 1221	-1.4	-2.3	-0.3	3.9	
IC_{II}^4	611 – 916	-2.0	-3.2	-0.7	6.1	
IC_{III}^4	305 – 611	-1.6	-3.4	0.2	2.9	
IC_{IV}^4	0 – 305	-1.1	-3.4	1.4	0.9	
η	Radius [km]	θ [%]	$\theta \pm 2\sigma$ [%]		D_{KL} [bits]	$p(\eta > 1)$
IC_I^4	916 – 1221	0.3	-0.8	1.3	2.1	0.70
IC_{II}^4	611 – 916	-0.1	-1.2	1.0	2.5	0.40
IC_{III}^4	305 – 611	-0.2	-1.8	1.5	1.2	0.43
IC_{IV}^4	0 – 305	-0.6	-3.0	1.9	0.5	0.32
ϕ	Radius [km]	θ [%]	$\theta \pm 2\sigma$ [%]		D_{KL} [bits]	$p(\phi > 1)$
IC_I^4	916 – 1221	0.9	-0.6	2.5	7.2	0.89
IC_{II}^4	611 – 916	0.3	-1.3	1.8	7.6	0.63
IC_{III}^4	305 – 611	1.3	-0.9	3.5	6.4	0.89
IC_{IV}^4	0 – 305	2.7	-1.1	6.4	3.9	0.92
ξ	Radius [km]	θ [%]	$\theta \pm 2\sigma$ [%]		D_{KL} [bits]	$p(\xi > 1)$
IC_I^4	916 – 1221	-0.6	-5.1	3.5	0.7	0.35
IC_{II}^4	611 – 916	-1.8	-5.9	2.1	1.8	0.17
IC_{III}^4	305 – 611	-2.0	-6.9	2.8	1.5	0.20
IC_{IV}^4	0 – 305	-2.5	-9.1	4.1	0.9	0.22

anisotropy, in terms of its depth extent and amplitude, depends very much on the chosen parametrisation. Here, we allowed for radial anisotropy, since we only work with 1-D earth models, and our marginals suggest strong radial anisotropy in the deep inner core. If we would parametrise the anisotropy with cylindrical symmetry, the inferences drawn from the degree-zero data could be different, e.g. the strongest anisotropy may present itself at a different depth or in a different anisotropic parameter. Thus, at this stage we can only suggest that either (i) the deep inner core is radially anisotropic or that (ii) the inferred structure results from the imprint of a cylindrically anisotropic structure on the degree-zero measurements. The inferred structure in the deep half of the inner core may point to the existence of a distinct innermost inner core, as proposed in earlier studies of inner core anisotropy, e.g. Ishii and Dziewoński (2002); Beghein and Trampert (2003), and attenuation, e.g. Li and Cormier (2002); Cormier and Stroujkova (2005).

As another explanation, one could conceive that these centre frequencies are not purely sensitive to degree-zero Earth structure. A possible source of error in normal mode measurements results from neglecting cross-coupling (resonance) between modes that are close in frequency, which has been shown to be relevant for the (anisotropic) inner core, e.g. (Andrews et al., 2006; Irving et al., 2008, 2009). However, the spheroidal modes measured by Deuss et al. (2013) take cross-coupling into account whenever two modes are close in frequency. Furthermore, these authors investigated the robustness of their inversion and found the centre frequencies to be the most well-resolved parameters in their inversion. Whether or not the observed anisotropy is related to other possible issues with either the spheroidal or radial mode data is difficult to investigate further.

7.4.2 Trade-offs between target parameters

Instead of assessing the nature of the data, we can focus on the earth model itself. We assess the strong P-wave anisotropy in the deep inner core in light of possible trade-offs between the well-resolved V_P and ϕ and construct 2-D marginal posterior pdfs for these two parameters in the four layers (IC^4_{I-IV}). This is achieved by multiplying a 1-D marginal pdf $p(x|\mathbf{d})$ with a conditional pdf $p(y|x,\mathbf{d})$ to obtain the 2-D marginal pdf $p(y,x|\mathbf{d})$, as explained in Sections 3.5 and 4.6.1. The 1-D marginal and conditional pdfs are both conditioned on the normal mode data $\mathbf{d}$ and are both approximated by training a separate ensemble of 48 MDNs.

Figure 7.4 shows the conditional and 2-D marginal posterior pdfs for the average V_P and ϕ in the four $\sim$300 km thick layers (IC^4_{I-IV}). Evidently, there exists a strong negative trade-off between V_P and ϕ in the three deepest layers (bottom three rows). Given the data $\mathbf{d}$, an increase in the Voigt average V_P requires a decrease in ϕ. Further, it is clear from the 2-D marginals for the two deepest layers that a *PREM*-like value for both parameters simultaneously is incompatible with the data. This illustrates the additional information on correlations (covariances) between model parameters available in higher-dimensional pdfs. Alternatively, we infer that either one of the parameters could be similar to *PREM*, as was clear from the 1-D marginal pdfs. For

instance, the radially averaged ϕ in the lower half of the inner core can be close to *PREM* and thus isotropy. However, this does require the V_P to be significantly larger ($\sim$1–2%) than in *PREM* (and vice versa).

Interestingly, a negligible trade-off exists for the top layer (IC_I^4, Figure 7.4, top row). Inspired by this result, we also construct 2-D marginals for (i) V_P and ϕ and (ii) ρ and ϕ in the top $\sim$115 km of the inner core, i.e. the three best-resolved parameters in our original region of interest (Section 7.3.2). Similar to the result for IC_I, we find that there is little trade-off in the 2-D pdfs between ϕ and either V_P or ρ (Figures 7.5 and 7.6). Possible trade-offs of ϕ with the more poorly resolved parameters V_S, ξ and η were not investigated here, but may also contribute to the uncertainties in the 1-D marginals for ϕ.

7.4.3 Layer thickness and information gain

The data provide a stronger constraint on the average structure in wider regions (Figure 7.1). Although this seems intuitively correct, we discuss the true cause(s) here. We emphasise that the measure of information gain depends on both the posterior and prior distributions. If the posterior pdf is narrower, we are more confident about the value of a parameter and D_{KL} increases, given the same prior distribution. However, the converse is also true: if the posterior pdf is similar for regions of varying thickness, but the prior pdf increases in width, D_{KL} is higher.

We consider the prior and posterior pdfs for ϕ in the five layers of our resolution test (Figure 7.1). Due to the style of parametrisation, which includes successive random perturbations applied to each knot in the earth models (Appendix A), the prior width increases with depth (see for instance the prior for ϕ in Figure A.1). For all five target parameters, the prior distribution of the 5000 test set samples can be represented by a Gaussian distribution. We find that the prior standard deviation for the average ϕ increases from 3.3% anisotropy for the 38 km thick layer to 3.5% for the widest layer (305 km). Simultaneously, the standard deviation in the five posterior pdfs decreases from 1.3% (38 km) to 0.8% (305 km). Thus, the changes in both the posterior and prior pdfs contribute to a higher information gain. The decrease in posterior uncertainties is largest and therefore is the dominant factor underlying the increase in information gain.

The increase in prior width with increasing layer thickness, and its effect on D_{KL}, can be explained by the style of parametrisation. But how are the changes in the posterior distributions related to the changes in the target parameters? The sensitivity of the data to a specific region in the Earth, and the data noise, provide a lower bound on the layer thickness at which we can resolve an arbitrary earth model parameter. Below this limit, there is no signal for this parameter in the noisy data. However, the synthetic data are calculated for the original (smooth) earth models, and not for earth models in which regions are set to the radial averages used as targets for the MDNs. Therefore, a physical effect, such as differences in the sensitivity of the data to different regions in the inner core, cannot explain the increase in D_{KL}, or at least not directly.

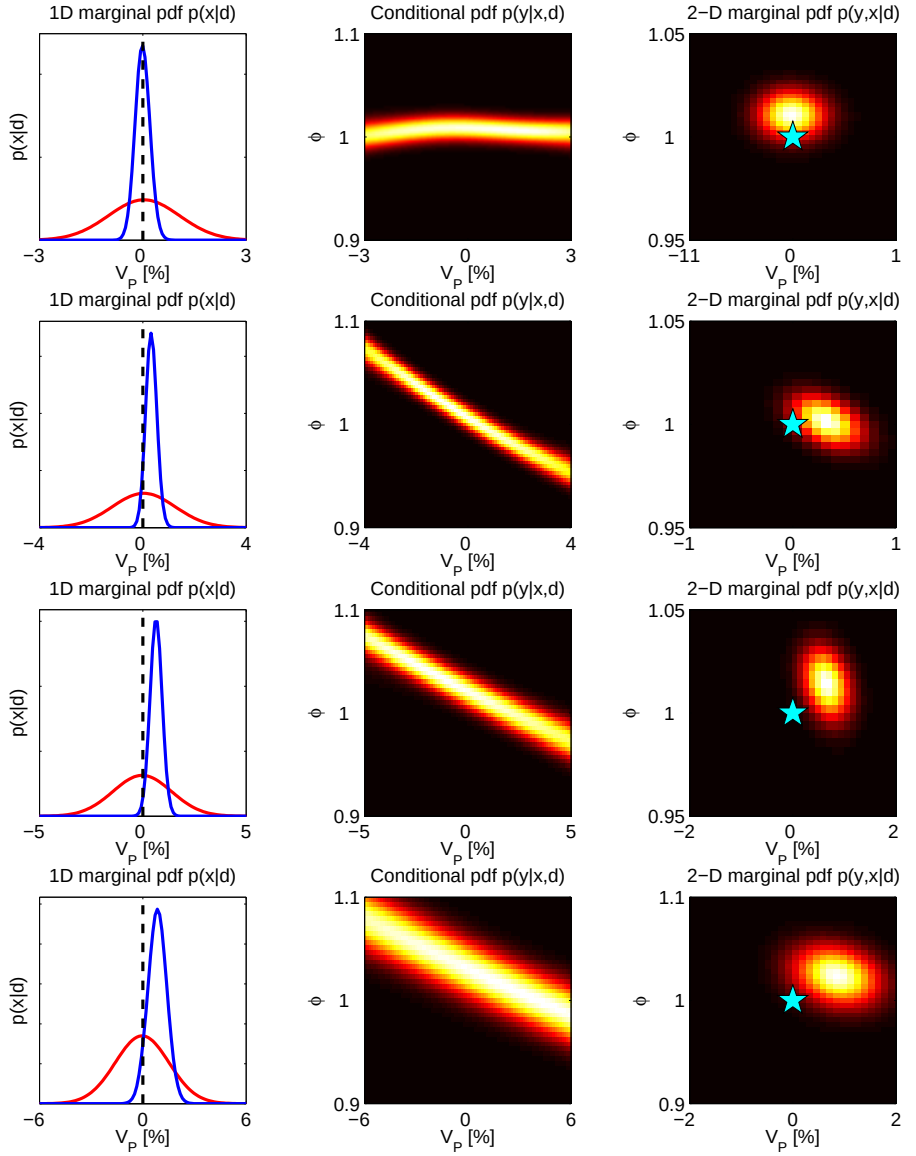


Figure 7.4: Construction of 2-D marginal posterior pdfs (Equation 4.3) for $\bar{V}_P$ and $\bar{\phi}$ in four layers of ∼300 km thickness in the the inner core, corresponding to *radii* 916–1221 km (IC^4_I, top row), 611–916 km (IC^4_{II}, second row), 305–611 km (IC^4_{III}, third row) and 0–305 km (IC^4_{IV}, bottom row). In each row, the three panels show the 1-D marginal (blue) and prior (red) pdfs for $\bar{V}_P$ (left-hand panel), the conditional pdf of $\bar{\phi}$ given $\bar{V}_P$ and the observed data **d** (middle panel) and the 2-D marginal pdf of $\bar{\phi}$ and $\bar{V}_P$ for the observed data **d** (right-hand panel). Lighter colours denote higher probabilities. The corresponding *PREM* values are denoted by the black line (left-hand panel) and the cyan star (right-hand panel).

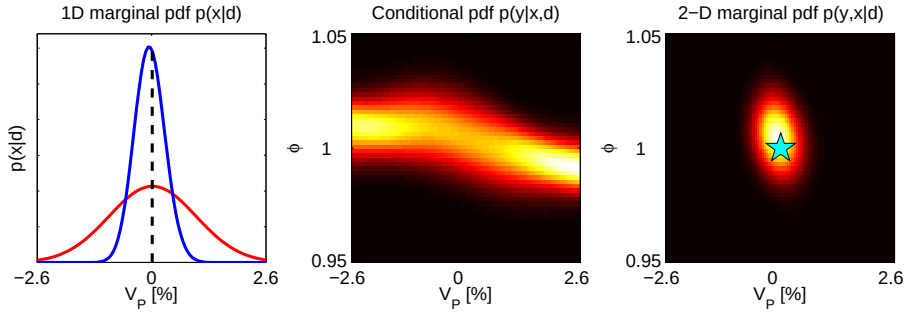


Figure 7.5: Similar to Figure 7.4, but for $\bar{V}_P$ and $\bar{\phi}$ in the top ∼115 km of the inner core.

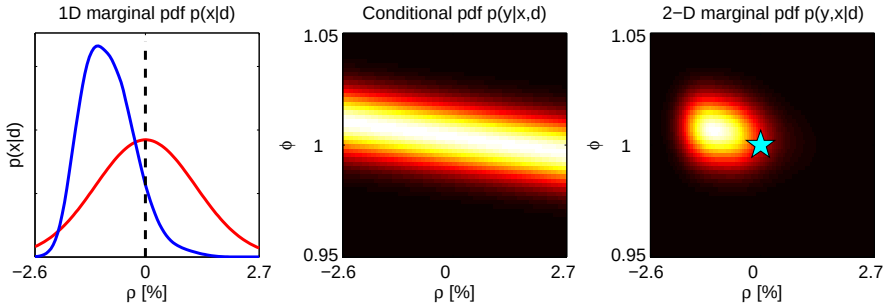


Figure 7.6: Similar to Figure 7.4, but for $\bar{\rho}$ and $\bar{\phi}$ in the top ∼115 km of the inner core.

Besides the change in target parameter, the only other change occurs for the inverse mapping itself. We believe that the averaging procedure partly cancels the random perturbations, which are applied to the individual knots during the generation of the synthetic earth models. What remains in the target parameter is the overall effect of the average radial structure on the data. The average structure over a relatively wide region is bound to have a larger impact on the data than the perturbations in individual knots. In other words, there is less ambiguity in the non-unique inverse problem: there are simply fewer options, i.e. less different values for the average Earth structure, that can explain the seismic observations. Consequently, the MDNs make more accurate and confident predictions, reflected by the narrower posterior distributions.

7.5 Conclusions

We trained MDNs on the radial averages of velocities, density and anisotropic parameters in layers of varying thickness in the inner core. The corresponding information gain served as a measure of the constraint provided by the spheroidal and radial mode data. By a stepwise increase of the layer thickness, we were able to deter-

mine a first-order estimate of the resolving length of the data, within the framework of the chosen earth model parametrisation. P-wave velocity and anisotropy are resolved best and are constrained even for layers ∼38 km thick, while the data contain little information on shear-wave structure, independent of layer depth and thickness. For a minimum layer thickness of ∼115 km, the majority of the parameters, including density, could be resolved. The increased confidence with which we infer the average parameter value over a wide region relates to the reduced ambiguity in the non-unique inverse problem. Phrased differently, there are simply fewer values (hypotheses) for the average Earth structure in a thicker layer that are compatible with the seismic observations.

Second, we studied the average radial structure in the top ∼115 km of the inner core. The average density is likely lower than in *PREM*, while the probability $p(\phi > 1)$, or equivalently $p(V_{PV} > V_{PH})$, is 0.71. Thus, our inferences are compatible with proposed radial anisotropy for which $V_{PV} > V_{PH}$, which could provide constraints on the dynamics of the shallow part of the inner core. However, we should note that such anisotropy is far from conclusive; we cannot rule out an isotropic structure in this region given our 1-D marginal pdfs. We constructed 2-D marginal pdfs to assess model parameter trade-offs, which we found to be negligible between ϕ and both V_P and ρ in the outermost inner core.

Third, we found strong positive anisotropy in ϕ in the deep inner core, increasing with depth. Conversely, the average density in each layer is likely lower than in both *PREM* and *ak135f*. 2-D marginal pdfs show a strong negative trade-off between the radial anisotropy and V_P, for which the positive deviation from *PREM* also increases with depth. An isotropic structure, i.e. $\phi \approx 1$, is compatible with the data only if V_P strongly deviates from *PREM*, i.e. by 1–2%. Thus, trade-offs between parameters (V_P and ϕ) in the deep inner core prevent us from making unambiguous inferences, i.e. obtain narrow 1-D marginals, on the radial anisotropy, despite the relatively strong constraint on these parameters ($D_{KL} \geq 3.9$ bits). A joint inversion of normal mode data and inner core-sensitive body-wave travel times may offer sufficient information to resolve these ambiguities, provided the issues regarding the latter data type, as discussed in Section 4.6.2, are addressed.

An explanation for the observed radial anisotropy in the deep inner core could lie in the nature of the data and the style of parametrisation. A cylindrically anisotropic model, such as has been proposed for the Earth's inner core, has an imprint on degree-zero measurements. Therefore, it is conceivable that we infer an anisotropic inner core from the centre frequencies, although our parametrisation only allows for radial anisotropy. A different parametrisation with a cylindrical symmetry may, for the same centre frequency data, result in the anisotropy being strongest in a different parameter or at different depths. Thus, two options exist: the inferred radial anisotropy either (i) is due to an inner core that is truly radially anisotropic or (ii) relates to the signal of cylindrical anisotropy in the degree-zero measurements. At this stage, we can only conclude that the centre frequencies require anisotropy in the (deep) inner core, in agreement with previous studies of inner core anisotropy, and may provide further evidence for a distinct innermost inner core.

8

An approach to dimensionality reduction for seismic waveform inversion

Abstract

Handling data dimensionality is an important task in machine learning applications. We illustrate one pragmatic approach to data dimensionality reduction, using an autoencoder network. Our method of choice is approximate and should be seen as a first step towards non-linear Bayesian seismic waveform inversion using encoded seismograms. We present three examples of increasing complexity, which illustrate that encoded seismograms can be used to perform a Bayesian seismic waveform inversion for 1-D Earth structure. Moreover, we show that the setup with the encoded waveforms can be advantageous over the time-domain alternative. This advantage may become more prominent if we use more data in future inversions, i.e. use more sources, receivers or seismogram components. Given the nature of seismic waveform data, inverting for 3-D Earth structure or seismic source parameters may be more suitable applications. Such inversions may benefit similarly from the data dimensionality reduction technique investigated here.

8.1 Introduction

In this age of 'big data', the ever-growing supply of data demands a simultaneous development of strategies and techniques to handle the increase in data volume. Due to its relevance to modern data acquisition, processing and inference problems, much work has been done on dimensionality reduction. For instance, one approach that recently has received much attention is that of compressed sensing, e.g. Donoho (2006); Candès and Wakin (2008), a signal processing technique that aims to sample signals at a reduced (sub-Nyquist) rate while minimising signal loss, and its application to the seismic realm, e.g. Herrmann et al. (2012). Seismology forms no exception to this trend of data abundancy, as evidenced by the growing collection of publicly available seismograms in data centres maintained by for instance the Incorporated Research Institutions for Seismology (IRIS). Advanced approaches are required to extract all available information and construct the best estimate possible for both Earth structure and seismic source. Current state-of-the-art techniques, dubbed Full Waveform Inversion (FWI), invert seismic waveforms for 3-D Earth structure using sophisticated forward modelling codes, e.g. Tromp et al. (2008); Virieux and Operto (2009); Fichtner (2010), although it remains a challenge to really use the full waveform. Seismic waveform inversion has also been conducted using artificial neural networks. See the reviews of geophysical neural network applications by van der Baan and Jutten (2000); Poulton (2001). Examples of waveform inversions for seismic structure include Röth and Tarantola (1994); Langer et al. (1996); Fu (2001). More recently, Käufl et al. (2015a) inferred seismic source parameters from GPS displacement waveforms.

The dimensionality of the data is a great challenge in such machine learning approaches. In previous chapters, we made inferences on Earth structure from body wave and free oscillation data. The input (data) vectors were constructed by simply concatenating a few hundred travel time or centre frequency measurements. It is not difficult to imagine that such an approach is prohibitive for seismic waveform inversion. A single seismogram may consist of hundreds of dimensions (time steps). Concatenating recordings at tens or hundreds of receivers, a typical number in modern waveform inversions, will result in an input dimension that will make neural network training computationally infeasible (e.g. $\mathcal{O}(10^2)$ time steps at $\mathcal{O}(10^2)$ stations for $\mathcal{O}(10^2)$ events results in $\mathcal{O}(10^6)$ dimensions). From the perspective of network training, an extreme example is the study by Röth and Tarantola (1994), who used many more free network pararameters ($\sim$135 000) than training patterns (450). In our experience, that this can make network training unstable, despite our use of ensembles (Section 2.4.8). In our view, if one wishes to use neural networks to invert tens or more of waveforms simultaneously, corresponding to an input dimensionality of $\mathcal{O}(10^3 - 10^4)$ at a minimum, a crucial first step is to reduce the dimensionality of the seismic data used as input.

Here, we illustrate one approach to data dimensionality reduction, using a so-called autoencoder, or autoassociator, network (Hinton and Salakhutdinov, 2006). Such a neural network was applied to seismograms by Valentine and Trampert (2012b) to

encode, or compress, the information on Earth structure that is contained in the seismic data and thus reduce data dimensionality. Our method of choice will be approximate, yet practical, and should be regarded as a first step towards non-linear Bayesian seismic waveform inversion using encoded seismograms.

First, we will briefly describe the waveform data and the model parametrisation. We use the same 1-D earth model parametrisation as in Chapter 4. Second, we outline the framework for the intended Bayesian inversion of encoded waveform data. Subsequently, we show three examples of increasing complexity, in which we compare the performance of Mixture Density Networks (MDNs, Section 2.4.4) on both time-domain and encoded waveforms. In the first example, we consider one seismogram for a single seismic event, i.e. one source-receiver pair. The second example extends this setup to four source-receiver pairs (four seismograms) to evaluate whether this results in a higher resolving power. The final example is similar to the second, but for a broader frequency band, again to assess whether more information on the model parameters can be extracted than in the previous case. In the three examples, we focus on shallow Earth structure, i.e. the depths of the 220 km and Moho discontinuities, P- and S-wave velocities and density in the uppermost mantle, and the source depth. Since the goal of this chapter is to present a framework for data dimensionality reduction, we focus on results for synthetic data for the spherically symmetric (1-D) *PREM* model (Dziewoński and Anderson, 1981).

8.2 Setup

8.2.1 Seismic waveform data

We only consider one seismic event, which we selected arbitrarily from the Global Centroid Moment Tensor (CMT) catalogue (Dziewoński et al., 1981; Ekström et al., 2012). The earthquake occurred in Southern China in January 2000 (25.34°N, 101.28°E, 33.0 km depth) and was recorded at a GSN station in Canada (ALE, 82.50°N, 62.35°W). The epicentral distance is 71.87°, $M_W = 5.5$, $M_0 = 1.92 \cdot 10^{24}$ dyn·cm and the source half-duration is 1.3 s.

The corresponding recording is available from the IRIS data centre. We downloaded one hour of recording at station ALE. For simplicity, we only use vertical-component seismograms. As a starting time for the hour-long window we used the event onset time, as given by the CMT solution. Prior to use, the data need to be processed. The details of the filtering and windowing processes are given in the relevant section for each example.

We use station-specific noise to corrupt the synthetic seismograms. The noise model is based on one hour of ambient noise recording at station ALE. Noise vectors are generated using the amplitude spectrum of the noise recording and a random phase spectrum. For simplicity, we use a polynomial fit (of fourth order) to the amplitude spectrum instead of the oscillatory amplitude spectrum itself.

8.2.2 Model parametrisation

The full model is represented by the 1-D earth model and the description for the seismic source.

1-D Earth structure

The model parametrisation is similar to that used in Chapter 4, i.e. the radial (1-D) structure of the Earth is parametrised in terms of V_P, V_S, density (ρ), the anisotropic parameter η and bulk and shear attenuation ($1/Q_\kappa$ and $1/Q_\mu$, respectively). We work with the first of the two anisotropic parametrisations, i.e. the isotropic parametrisation in which only the uppermost mantle is allowed to be radially anisotropic, similar to *PREM* (Dziewoński and Anderson, 1981). The parametrisation is described in more detail in Appendix A.

Source parameters

We allow for variations in the seismic source to take account of the uncertainties in the CMT solutions. We vary the three source location parameters and the magnitude parameter M_0 (Table 8.1). The uniform prior distributions for the source depth, latitude and longitude are centred on the event in Southern China. We impose an additional constraint in terms of the epicentral distance to the receiver, which has to lie in the range 71.5–72.5°. Note that this introduces correlations between latitude and longitude in the prior distribution. For simplicity, we fix the source half-duration and the six moment tensor components to the corresponding values in the CMT solution for the China event and ignore the station elevation.

For the examples in this chapter, this inhibits us from applying the trained network to the recordings for other events, i.e. with other radiation patterns or moment tensors. One of the advantages of neural networks is that—once trained—they can be applied rapidly and repeatedly to new observations. Ideally, all source parameters are varied when generating training sets for future applications. This would enable us to apply our trained neural networks to the seismograms recorded for all events (CMT solutions) that lie within our source prior, as explained by for instance Käufl et al. (2014), and by doing so make efficient use of available computational resources.

Synthetic data

We use the Mineos package (Masters et al., 2011) to compute synthetic seismograms for 100 000 earth models drawn from the prior model distribution. Spheroidal, toroidal and radial modes are summed for radial orders 0–30 and angular orders 1–8000 in the frequency range 0–80 mHz. Using a value of 10^{-7} for the parameter that controls the accuracy of the Runge-Kutta integration scheme, computing seismograms for one 1-D earth model and a single source takes approximately 70 s on a standard desktop computer.

Table 8.1: Prior information on the source parameters. The prior distributions for the source depth, latitude and longitude are centred on the CMT solution for the event in Southern China. The six moment tensor components are fixed to this CMT solution.

Parameter	Range
Depth [km]	28.00 – 33.00
Latitude [°]	25.09 – 25.59
Longitude [°]	101.03 – 101.53
Epicentral distance [°]	71.50 – 72.50
M_0 [10^{24} dyn·cm]	1.50 – 3.50

8.3 Methodology

As in previous chapters, we apply Mixture Density Networks (MDNs) to seismic data to infer probability density functions (pdfs) on earth model parameters (Section 2.4.4). To reduce the data dimensionality we use autoencoder networks, which we will introduce here. A detailed description of autoencoder networks can be found in for instance Hinton and Salakhutdinov (2006); Valentine and Trampert (2012b). Second, we outline the envisioned framework for the dimensionality reduction and subsequent inversion of seismic data. Finally, we specify the configuration for the MDNs and the autoencoders.

8.3.1 Autoencoders

Autoencoders, or autoassociators, are a type of deep neural network, where 'deep' in essence refers to the presence of more than one hidden layer. Recently, deep neural networks have received much attention due to their perceived ability to learn more complex relations than networks with many hidden neurons in a single layer. As such, they may bring us one step closer to mimicking the functioning of the human brain. Current applications are widespread and include autonomous driving and image recognition. See Bengio (2009); Schmidhuber (2014) for an overview and the state-of-the-art of deep learning.

The autoencoder itself is an approach to project high-dimensional data onto a lower dimensional space, while retaining all or most of the information contained in the data. As such, it can be viewed as a tool for non-linear principal component analysis. The idea behind the autoencoder is to encode the input $\mathbf{x}$ into some lower dimensional representation $f(\mathbf{x})$ in such a way that the input can be reconstructed from that representation (Hinton and Salakhutdinov, 2006). Hence, the original high-dimensional data form both the input and the target for the autoencoder. In between the input and output, multiple hidden layers of decreasing size (decreasing number of hidden neurons) lead up to a so-called 'bottleneck', or *encoding*, layer (Figure 8.1). The encoding

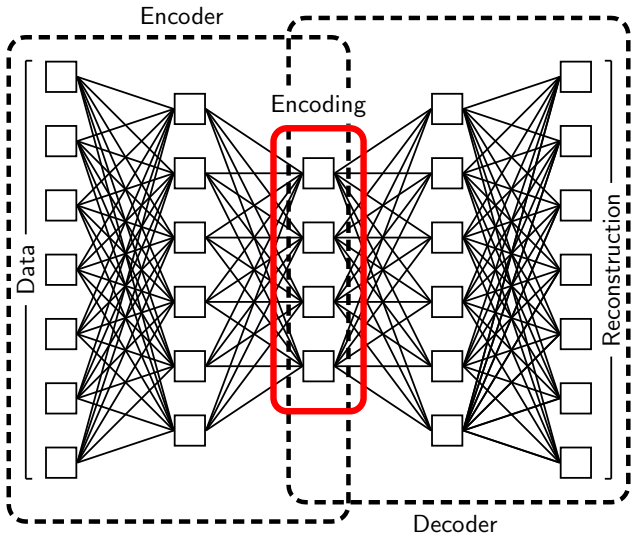


Figure 8.1: A schematic representation of an autoencoder network, as depicted in Valentine and Trampert (2012b). In the first of two steps, the autoencoder *encodes*, or compresses, the input data into a lower-dimensional representation. The second component, the decoder, attempts to reconstruct the input from the encoding. Autoencoder optimisation, or training, involves adjusting the network parameters to minimise the reconstruction error, i.e. the difference between the decoding and the original input, for a given data set.

layer is the smallest in terms of the number of hidden units. The size of successive layers increases again until we reach the output layer, or *decoding*, which has the same dimension as the input layer. The difference between the target, the original data, and the output, the decoded data, is the driving factor in autoencoder optimisation or training. Once this misfit is small, given some pre-defined criterion, we assume that autoencoder training is successful and that the low-dimensional encoding contains the same information as the original data, but represented more efficiently, i.e. in fewer dimensions. Subsequently, we can use the trained autoencoder to encode, or compress, the original data. See Valentine and Trampert (2012b) for an detailed explanation of autoencoder training.

8.3.2 Workflow

We envision the following framework for data dimensionality reduction in the context of seismic waveform inversion.

1. Decide on a source-receiver configuration, i.e. (input) data distribution, based on the availability of real data. Process the observed data (filter, deconvolve instrument response, select time windows).

2. Compute synthetic seismograms using a forward solver (Mineos) for a set of random earth models and point sources, drawn from a prior model distribution. The computational requirements of the forward solver determine to what accuracy the waveforms can be computed. Here, it depends on the time needed by Mineos to sum normal modes up to a certain frequency. Add random noise to each synthetic datum, using a pre-defined noise model. This is necessary to make autoencoder training insensitive to the noise-level variations in the waveforms, i.e. to improve the generalisation capacity of the autoencoder (Section 2.4.6). Subsequently filter and downsample the synthetic data using the same filter as was used for the observed data.

3. Pre-process the synthetic data set to facilitate autoencoder training in the next step. Here, we standardise both the input and the target data, e.g. (Bishop, 1995; de Wit et al., 2013). For each time step, the synthetic samples are (linearly) transformed to have zero mean and unit variance (Section 2.4.7). The same transformation is applied to the synthetic data for *PREM* and the observed data. We emphasise that this transformation per time step does not alter the information content of the data. Network optimisation is driven by the differences between training patterns and not by absolute values; the rescaling due to the standardising does not change the relative differences between these training patterns. Furthermore, we note that the linear tranformation for each element could in theory be equal to the initial value for the network weights in the input layer (Figure 2.1). Such network weights may take any value, and do not alter the information contained in the input data, which is independent of the weights.

4. For each station, train a separate autoencoder network, since every station, or source-receiver pair, corresponds to a different propagation path and thus to a different sampling of, and information on, Earth structure. The autoencoders are trained using the synthetic training set. Alternatively, one could try to train autoencoders on real data, as a form of unsupervised learning (Section 2.3), or by using a mix of real and synthetic data. Such options are not explored in this chapter, however; we focus on the proof of concept for seismogram dimensionality reduction and inversion. Note that the autoencoders are trained on the (pre-processed) filtered and windowed waveforms, not on the original full recordings.

5. Apply the trained autoencoders to both the synthetic and the observed data to obtain the lower-dimensional encodings.

6. Train ensembles of MDNs using the encoded waveforms as the input and the earth model parameters as the target. The MDN input patterns will consist of a concatenation of the encoded waveforms for each source-receiver pair.

8.3.3 Network configuration

Autoencoder training

We train autoencoders using the program developed and described by Valentine and Trampert (2012b). We use 10 000 patterns for both the training and monitoring data set, which are all given unit weight. Prior to autoencoder training, the input data are standardised (Section 2.4.7). Further, the individual hidden layers in the autoencoder are subjected to an initial optimisation, or *pre-training*, procedure before autoencoder training commences. Pre-training of the hidden layers is crucial in deep learning, and many believe this to be a prerequisite for their training and operation to be successful[1], e.g. Hinton and Salakhutdinov (2006); Bengio (2009).

In our setup, the pre-training stage involves 500 iterations to optimise Continuous Restricted Boltzmann Machines (CRBMs). Subsequently, autoencoder training is run for 3000 iterations using the *iRprop+* algorithm (Igel and Hüsken, 2003). We do not add noise during either CRBM or autoencoder training, as we already add realistic station-specific data noise before pre-processing the waveforms.

As an example, we show time-domain input waveforms and the corresponding encoded and decoded waveforms for synthetic data for *PREM* and the observed data at station ALE for the event in Southern China (Figure 8.2). The layers in this particular autoencoder consisted of 87–64–32–16–32–64–87 units, i.e. the input was 87-D and the encoding contained 16 units. Although the input and decoding overlap for the larger part, the decoding is not a perfect reconstruction of the input. For all 100 000 synthetic patterns, we evaluate reconstruction quality using the correlation coefficient R_{rec} between the input $\mathbf{x}_{in}$ and decoding $\mathbf{x}_{dec}$ and their relative misfit, defined as $\psi = \sqrt{\sum_{i=1}^{I}(\mathbf{x}_{in}^i - \mathbf{x}_{dec}^i)^2/(I\sigma_i^2)}$ with the standard deviation σ_i of the noise model for the I input units (87 time steps here, Figure 8.3). For perfect reconstructions, $R_{rec} = 1$ and the misfit is zero. However, we note that a perfect match is not necessary or even possible, as both the synthetic *PREM* and observed waveforms contain noise, which we do not need to recover; for noise-level errors in the reconstruction, i.e. on the order of 1σ–2σ, the misfit ψ is approximately 1–2. Furthermore, we did not attempt to modify the autoencoder training algorithm until it reached optimal performance on the synthetic waveform data set we use here. Rather, we accept a decent performance and focus our efforts on illustrating the dimensionality reduction and waveform inversion framework. As an additional advantage of our application, we can assess the MDN performance on the (encoded) synthetic data for the test set and *PREM*, since we know the target values for these earth models.

MDN training

We train ensembles of 48 MDNs on the waveforms in both the original time and the encoding domain. The number of Gaussian kernels is 5 for all results shown in this

[1] As an aside, we note that autoencoders themselves have been used to pre-train the layers in deep neural networks, e.g. Larochelle et al. (2009).

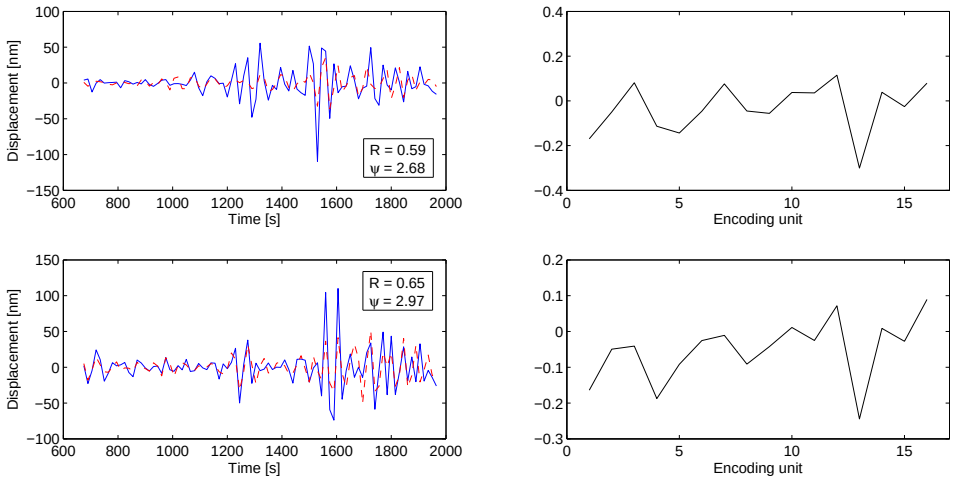


Figure 8.2: Examples of a 87-D time-domain input (blue solid line) and decoded waveform (red dashed line, left-hand panels) and the corresponding 16-D encoding (right-hand panels) for synthetic data for *PREM* (first row) and the observed data (second row) at station ALE and the event in China. Time is given relative to event onset (at 0 s). The quality of the decodings is quantified by the correlation coefficient R_{rec} and the relative misfit ψ (Section 8.3.3). The original seismograms are shown in Figure 8.4. The examples corresponds to the setup used in Section 8.4.1.

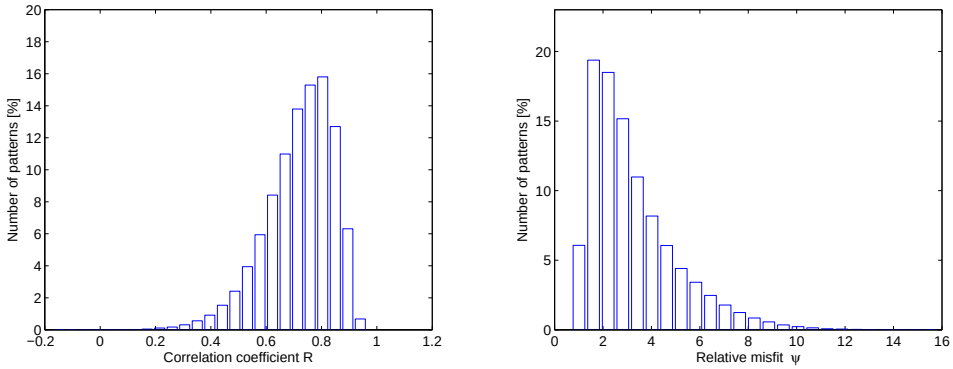


Figure 8.3: Quality of decodings, as expressed by the correlation coefficient R_{rec} (left-hand panel) and the relative misfit $\psi = \sqrt{\sum_{i=1}^{I}(\mathbf{x}_{in}^{i} - \mathbf{x}_{dec}^{i})^2/(I\sigma_i^2)}$ (right-hand panel) between the input $\mathbf{x}_{in}$ and decoding $\mathbf{x}_{dec}$ for the 100 000 samples in the training set. Clearly, the decoding is not a perfect reconstruction of the input, but serves our purpose in this study, as explained in the text (Section 8.3.3). The source-receiver setup is the same as in Figure 8.2, i.e. for the example shown in Section 8.4.1 for an autoencoder trained on an 87-D input and decoding and 16-D encoding at station ALE and the event in China.

chapter; we verified that the precise number of kernels is not crucial to the MDN output. The number of hidden units varies with the size of the input, which varies between examples depending on the number of receivers and the frequency band. As in Chapters 3 to 7, we use a synthetic data set containing 100 000 patterns, of which 80%, 15% and 5% are used for the training, validation and test set, respectively. The MDNs are trained using the Scaled Conjugate Gradient (SCG) algorithm Møller (1993) for a maximum of 5000 iterations.

8.4 Results

We illustrate the approach through three examples. In all cases, we compare the MDN performance for two different inputs: (i) time-domain seismic waveforms and (ii) encodings of these time-domain waveforms, as output by a trained autoencoder. As quantitative proxies for the data constraint on the target parameters, we use the correlation coefficient R for the 5000 test set samples (see for instance Figures 3.9 and 4.2) and the information gain measure D_{KL} for *PREM* (Section 2.5). We consider six 1-D target parameters in the uppermost mantle, that is, the Moho depth, the depth of the "220" discontinuity, the source depth, the radially averaged Voigt average isotropic velocities $\bar{V}_P$ and $\bar{V}_S$ (Equations A.4 and A.5) and the density ($\bar{\rho}$) in the region between the Moho and the "220" ("Moho-220"). In the first example, we consider the data recorded at the station ALE and the event in Southern China. The second example extends this setup to four receivers at different epicentral distances from the same event. In the final example, we use the same source and receivers as in the second setup, but the data are filtered using a broader band-pass. The corresponding increase in frequency content may increase the information available on uppermost mantle structure, at the expense of a larger input dimension (the time steps are smaller).

8.4.1 Example I – One station and narrow frequency band

We add filtered station-specific noise to the synthetic data set and the synthetic waveform computed for *PREM*. The noise is based on an hour-long recording of ambient noise at station ALE. We apply a cosine taper (of normalised width 0.1) and band-pass filter to the seismograms with corner frequencies 10, 12.5, 25 and 33 mHz (periods 100, 80, 40 and 30 s). Figure 8.4 shows the filtered observed and noisy synthetic data for *PREM*, as computed by Mineos, at station ALE for the event in Southern China. As a reference, we show the travel times for the *P* and *S* phases in *PREM*, as computed by the TauP package (Crotwell et al., 1999), which calculates body wave travel times in spherically symmetric earth models. The observed and synthetic waveform are fairly similar, but the surface waves are not recovered by the 1-D forward solver. Note that the displacements prior to the arrival of *P* in both the observed and noisy *PREM* waveform, and in fact the displacements for *P* itself, are on the order of the noise level.

The data are downsampled using a 15.0 s interval, resulting in waveforms consist-

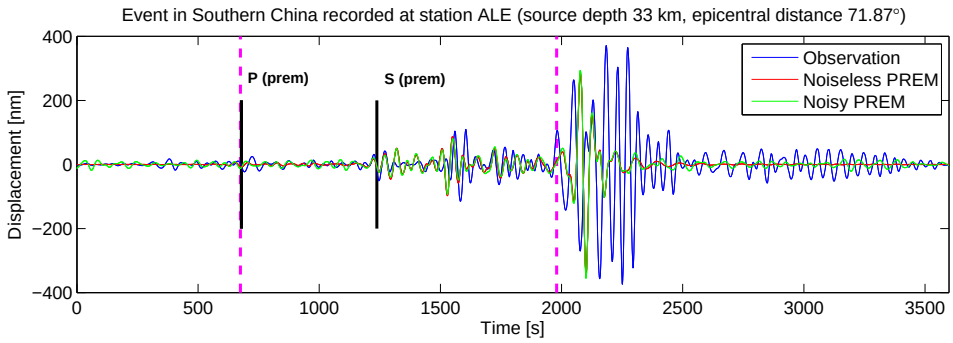


Figure 8.4: Comparison of recording at station ALE for an event in Southern China (blue line) and synthetic data for *PREM*, both with (green line) and without (red line) addition of ambient noise. P and S travel times, computed by TauP, are shown as a reference (black). The time window used as input to the MDNs is marked by the magenta dashed lines.

ing of 240 time steps. As a first step to reduce the data dimensionality, we window the hour-long recordings. The window starts in the time step preceding the arrival of the direct *P* phase, as computed by TauP. For the source-receiver pair China-ALE and *PREM*, TauP gives P = 678.70 s and S = 1238.25 s. For simplicity, the window ends approximately just before the surface waves arrive ($\sim$ 1980 s, Figure 8.4) and reduces the input dimensionality from 240 to 87 time steps. Admittedly, a more conservative window, e.g. one that starts (tens of) seconds before the direct P-wave arrives, would increase the likelihood of including such primary phases and thus their constraint on Earth structure. However, the focus in this chapter lies on illustrating the framework and not on the actual application to 1-D earth models. Furthermore, as we noted above, the amplitude of for instance the *P* phase is similar to that of the noise we add to the synthetic data. Thus, in the current setup it is unlikely that the MDNs will extract information on Earth structure from this particular part of the waveform

For this first example, the layer sizes for the autoencoder are 87–64–32–16–32–64–87, i.e. the encoding consists of 16 units. Thus, we train the autoencoder on the windowed waveforms and not on the original hour-long seismograms. This enables us to make a direct comparison of MDN performance for the time-domain and encoded waveforms. For both input types, we train MDNs with 16 hidden neurons.

Time-domain waveforms

First, we train MDN ensembles on the six 1-D targets using the windowed 87-D time-domain waveforms as input. Except for the Moho depth ($R = 0.92, D_{KL} = 8.0$), the data do not constrain these model parameters, or at most very weakly for $\bar{V}_P$ and $\bar{V}_S$ (Table 8.2). We show 1-D marginal posterior pdfs for the Moho depth and $\bar{V}_S$ (Figure 8.5) for a test set sample, *PREM* and the observed data. The prediction for the Moho depth is very accurate for the test sample and for *PREM*. For the observed data, the marginal indicates a preference for a deeper Moho. Very little or no information

Table 8.2: Correlation coefficient R for the 5000 test set samples and D_{KL} (in bits) for synthetic *PREM* data for station ALE and the event in Southern China for the six 1-D target parameters (Section 8.4.1). For each parameter, the columns show R and D_{KL} for the 87-D time-domain ('time') and 16-D encoded ('enc') input.

Target	R		D_{KL} [bits]	
	'time'	'enc'	'time'	'enc'
"220" depth	0.24	0.05	0.0	0.0
Moho depth	0.92	0.85	8.0	3.9
Source depth	0.19	0.09	0.0	0.0
$\bar{V}_P$	0.63	0.45	0.1	0.2
$\bar{V}_S$	0.70	0.47	0.3	0.1
$\bar{\rho}$	0.18	0.11	0.1	0.1

is available for $\bar{V}_S$, which results in a broad Gaussian-shaped pdf similar to the prior for both the *PREM* and the observed data.

Encoded waveforms

Second, we train an autoencoder to compress the windowed 87-D waveforms into 16-D encodings and use the encodings of the complete synthetic data set (100 000 patterns) to train the MDNs. We find that R is lower than for the 87-D input for all six target parameters (Table 8.2). This indicates a (minor) loss of information due to the compression. Nonetheless, the Moho depth is still accurately resolved for *PREM* (Figure 8.6). The prediction for the observed data again indicates a preference for a greater Moho depth than in *PREM*, but the marginal is wider and assigns a non-zero probability to most of the depths allowed in the prior.

8.4.2 Example II – Four stations and narrow frequency band

To increase the constraint on the model parameters, we extend the previous setup by adding three fictitious stations at epicentral distances of 11.16°, 30.86° and 51.72° with respect to the China event. As we do not have observations for such epicentral distances, all analyses in this and the third example are based on synthetic data for *PREM* only. Figure 8.7 shows synthetic seismograms for the China event and *PREM* at the four receivers. We use the ambient noise recording for station ALE for all four stations. The recordings at the stations closer to the source will have a higher signal-to-noise ratio, as the amplitude of the recorded displacement is larger. The frequency band and other data processing is the same as in the first example.

As before, we window the waveforms, which contain a lot of time steps that record ground motion of near-zero amplitude (Figure 8.7). The windows are defined from the P travel time to approximately the arrival of the surface waves. For *PREM* and the

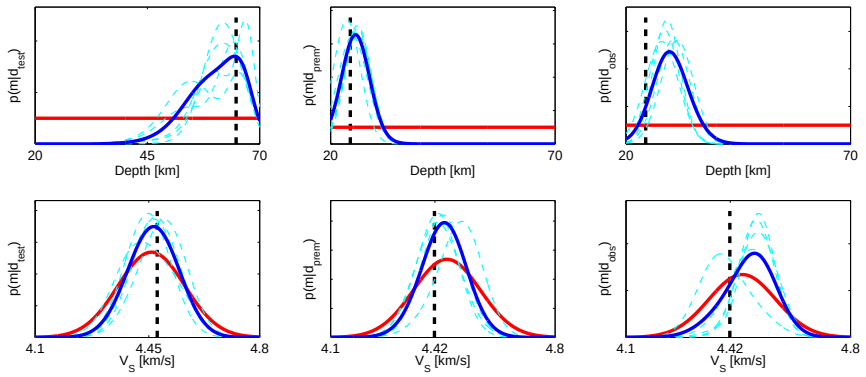


Figure 8.5: Ensemble of networks trained on the Moho depth (first row) and the Voigt average $\bar{V}_S$ in the "Moho-220" region (second row) as target and the 87-D time-domain waveform as input. The columns show 1-D marginal posterior pdf (blue line), prior pdf (red) and the target value (black, dashed) for a test set pattern (left-hand panels), *PREM* (middle panel) and the observed data for the China event (right-hand panel). The normalised pdfs for five individual networks in the ensemble are shown in cyan. Note that the *PREM* value in the right-hand column is shown as a reference and does not represent a target for the observed data.

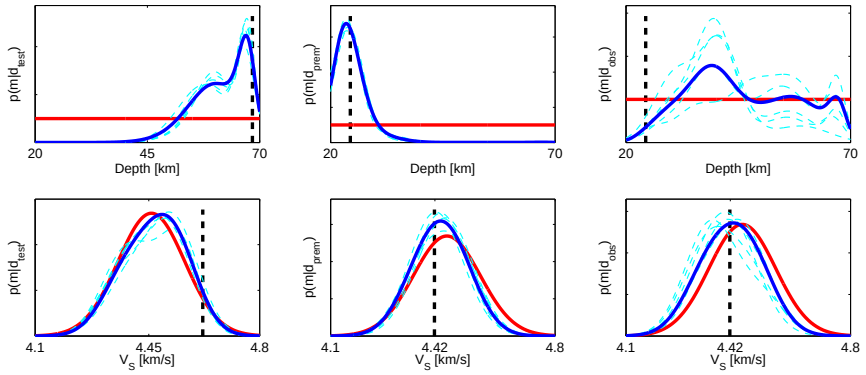


Figure 8.6: Similar to Figure 8.5, but for MDN ensembles trained on the 16-D encodings of the 87-D time-domain waveform.

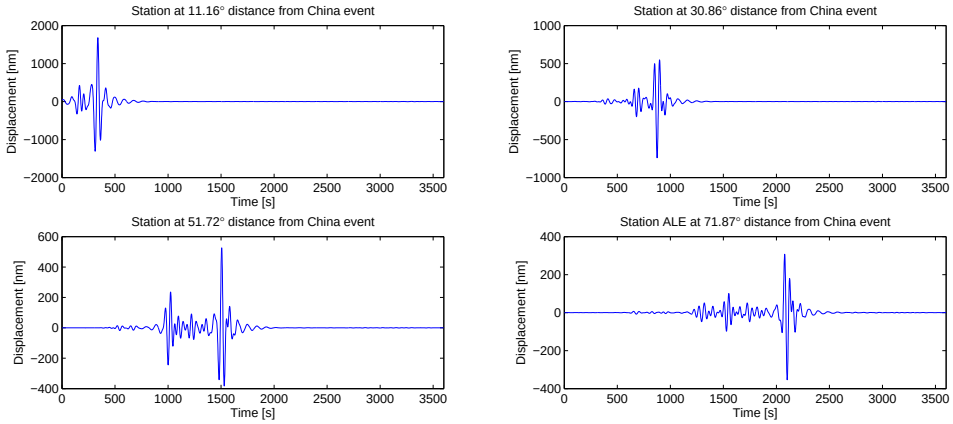


Figure 8.7: Noiseless synthetic data for the China event and *PREM* for the four source-receiver pairs of the second example (Section 8.4.2). The band-pass filter is the same as for the first example and has corner frequencies 10, 12.5, 25 and 33 mHz.

China event, TauP computes a P-wave arrival after 154.9 s (for a distance of 11.16°), 372.9 s (30.86°), 543.4 s (51.72°) and 678.7 s (71.87°). For the same four distances, the end of the time window is set to 285 s, 765 s, 1455 s and 2025 s, respectively. Using the time interval of 15.0 s, the input dimensionality for the four receivers is 10, 28, 61 and 90 (ordered by increasing epicentral distance). Thus, the concatenated input has 189 dimensions. For each source-receiver pair, we train a separate autoencoder on the windowed waveforms. Prior to training the autoencoders, we adjust the size of the encoding layer to the size of the windowed waveform. The concatenated input contains 56 elements: the individual encodings are of length 16, except for the receiver closest to the source (8 units). For both the time-domain and encoding waveforms, MDNs are trained using 40 hidden units to facilitate the comparison of the two input types.

Time-domain waveforms

For all six targets, R and D_{KL} are higher than in the first example (Table 8.3). The increased constraint on the velocity structure in the uppermost mantle is likely due to the addition of the nearby station(s) at $\sim$11° (and $\sim$31°) distance. Density and the "220" depth are not resolved; for the latter, this result is similar to the lack of constraint found using body wave travel times (Chapter 3) and normal mode centre frequencies (Chapter 4).

Encoded waveforms

For all six parameters, performance is comparable to the windowed waveforms in the time domain, although both R and D_{KL} are slightly lower due to some minor loss

Table 8.3: Similar to Table 8.2, but for the setup with four stations (Section 8.4.2). Correlation coefficient R for the 5000 test set samples and D_{KL} (in bits) for synthetic *PREM* data for four stations and the event in Southern China. For each parameter, the columns show R and D_{KL} for the 189-D time-domain ('time') and 56-D encoded ('enc') input vectors.

Target	R		D_{KL} [bits]	
	'time'	'enc'	'time'	'enc'
"220" depth	0.57	0.55	0.1	0.1
Moho depth	0.99	0.98	10.3	9.4
Source depth	0.86	0.79	0.9	0.5
$\bar{V}_P$	0.92	0.91	2.6	2.0
$\bar{V}_S$	0.84	0.84	1.3	1.0
$\bar{\rho}$	0.40	0.47	0.1	0.1

of information during the encoding step (Table 8.3). This is reflected by the posterior pdfs, which are a bit broader than for the time-domain waveforms but indicate accurate predictions for *PREM* (not shown here).

8.4.3 Example III – Four stations and broad frequency band

To increase the constraint on the model parameters, we extend the previous four-station setup by using a broader frequency band. The band-pass filter has corner frequencies 5, 10, 67 and 80 mHz (periods 200, 100, 15 and 12.5 s). The time step for downsampling is 6 s and the resulting hour-long seismograms contain 600 elements. The ambient noise recording at station ALE is filtered accordingly using the same broader band-pass. The time windows are similar to those used for the second example. The concatenated input for the four resulting waveforms is 461-D (the number of inputs for $dt = 6.0$ s is 23, 77, 147 and 314 for the four receivers). Again, a separate autoencoder is trained to encode the windowed waveforms at each of the four stations. The corresponding concatenated encodings are 96-D; the encodings have 32 and 16 elements for the two closest and two furthest receivers, respectively. Note that the larger 32-D encodings are simply meant to reduce the degree of compression, and thus the risk of information loss, for the two longest waveforms, i.e. 147 and 314 time steps compressed to 32 instead of 16 dimensions. For both the time-domain and encoded waveforms, we use 40 hidden units and train MDNs on the same six 1-D targets as in the first two examples.

Time-domain waveforms

For five of the six parameters, the correlation coefficient R is lower (Table 8.4) than in the second example (Table 8.3). This is surprising, as the broader frequency band

Table 8.4: Similar to Table 8.3, but for the waveform data filtered by a broader band-pass (Section 8.4.3)). Correlation coefficient R for the 5000 test set samples and D_{KL} (in bits) for synthetic *PREM* data for four stations and the event in Southern China. For each parameter, the columns show R and D_{KL} for the 461-D time-domain ('time') and 96-D encoded ('enc') input vectors.

Target	R		D_{KL} [bits]	
	'time'	'enc'	'time'	'enc'
"220" depth	0.29	0.57	0.0	0.1
Moho depth	0.98	0.99	7.1	10.6
Source depth	0.74	0.79	0.1	0.3
$\bar{V}_P$	0.94	0.95	2.7	3.2
$\bar{V}_S$	0.81	0.80	0.7	0.7
$\bar{\rho}$	0.30	0.34	0.1	0.2

was expected to contain more information on the model parameters. We can explain this lower constraint in several ways. First, the larger size of the input (461 time steps versus 189 elements in the encodings) results in a larger number of network parameters, while the number of training patterns was fixed (80% of 100 000 patterns). This lower ratio of training patterns versus network weights may have complicated the network training process. Second, we should not forget that in both cases we use 40 hidden units in the MDN. While this facilitates a comparison between the two examples, it causes a relatively strong compression of the information in the 461 input elements (461 time steps to 40 versus 189 encoding elements to 40 units) and may have resulted in information loss. Third, the decreased performance may simply relate to the increase in the assumed data noise; we found that the amplitude of the noise in the real data is larger for the broader frequency band (not shown here).

Encoded waveforms

For five of the six parameters, performance is *better* than for the windowed waveforms in the time domain, in terms of both R and D_{KL} (Table 8.4). This is in contrast with the results for the first and second example. The improved performance may be for the same reasons as we suggested above to explain the poorer performance of the 461-D setup with respect to the 189-D case. First, the network has fewer free parameters for the 96-D input, which facilitates network training and results in a relative abundance of training patterns. Second, the setup for these encoded waveforms involves a 96 to 40 unit compression (cf. 461 to 40). Apparently, any loss of information due to the autencoder compression is outweighted by the enhanced performance of the subsequent MDN training. Performance is comparable to that for the second example (Table 8.3).

8.5 Discussion

8.5.1 Source-receiver configuration

In this chapter, the goal was to illustrate one approach to reduce data dimensionality, using autoencoders, which can facilitate MDN training for relatively large inputs. As an example, we applied our framework to seismic waveform inversion for 1-D Earth structure. The autoencoders themselves were trained using synthetic data. Consequently, the main point of attention is the source-receiver configuration and the source prior (radiation pattern). One requirement is that the final inference or inversion is insensitive to the specific propagation path between sources and receivers, i.e. is not biased by 3-D structural variations along the path or the specifics of the seismic source. This ensures that the inferred model parameters are representative of the average (1-D) Earth and not of a region therein. Can we remove this bias by using more seismograms, i.e. more source–receivers paths, as input? Perhaps we should accept that inferences are biased to the propagation path, and that therefore the actual application of waveform inversion for 1-D Earth structure is not the most useful, as we noted at the beginning of this chapter. Better use of the relatively complex waveform data can be made when inverting for 3-D Earth structure or for seismic source parameters, e.g. (Käufl et al., 2015a). The latter approach has the additional advantage that trained neural networks can be applied to multiple observations, i.e. multiple events, thereby potentially offering advantages in terms of computation time.

One way of testing whether or not our inversion is insensitive to 3-D structure and the source is to apply the trained MDNs to multiple data vectors for a given epicentral distance, akin to the travel time study (Chapter 3). There, we 'inverted' 10 EHB input patterns using the same trained neural networks (Figures 3.10 and 3.11). To do so here, the prior on the source parameters has to be conservative enough to incorporate the variation in all real data (all seismic events) one may wish to invert. Admittedly, data selection for the seismic waveforms would be a bit more tedious than for the existing travel time bulletins. For each waveform used to train the MDNs, we would need to find multiple recordings for both a similar epicentral distance and a source that is incorporated by our source prior. This would require that the moment tensor components are not fixed in the prior distribution, as they were here.

8.5.2 The (in)dependence of dimensionality reduction

In the proposed setup (Section 8.3.2), the optimal autoencoder is the one which minimises the discrepancy between the seismograms and their encoded-decoded equivalent (the output of the autoencoder). During the optimisation process, the underlying earth model was not considered. Thus, the optimal encoder found through autoencoder training may not be the encoder that preserves the maximum possible amount of information on Earth structure, given the allowed dimension of the encoding. If the instruction would be to preserve all possible information on Earth structure, at the expense of information on the seismic source, the optimal autoencoder may be different. To implement such an instruction, we would need to combine the dimensionality re-

duction with the actual seismic inversion. In fact, a classical two-layer MLP, where the hidden layer usually is of a lower dimension than the input layer, does nothing else than encode, or compress, the information in the input data space, the aim being to make accurate predictions for the target parameters.

So why is the explicit separate encoding step necessary to begin with? The alternative would be to concatenate the original seismograms, instead of their encodings, for al receivers, i.e. at all epicentral distances. Clearly, the dimension of such an input vector will easily be at least one order of magnitude larger than that of the encoded input. Consequently, the number of free parameters in an MLP or MDN may be one or more magnitudes larger and may require many more synthetic patterns to train succesfully. Our final example already illustrates this limitation (Section 8.4.3). The concatenated input for the time-domain waveforms has 461 dimensions, as opposed the 96-D encoded input. For the 40 hidden units and the 5 Gaussian kernels, this results in an MLP with 19 095 weights (Equation 4.1), versus 4495 for the 96-D encodings. Furthermore, we note that the dimension reduction in the former case is much stronger: with 40 hidden neurons, a 461-D vector is projected onto a 40-D space. This compresses of the number of dimensions by one order of magnitude, as opposed to the compression from 96 to 40 dimensions in the case of the encodings. Thus, in the former setup we additionally may have to increase the number of hidden units, requiring more training patterns. This will become computationally infeasible quite rapidly. In this light, we emphasise again the insight offered by the third example, in which a loss of information due to a possibly suboptimal or imperfect autoencoder was outweighted by the enhanced performance of the MDN ensembles on the encodings (Table 8.4).

8.6 Conclusions

The relatively simple examples in this chapter are meant to illustrate one approach to dimensionality reduction of seismic data and how the resulting encoded seismograms can be used to solve a Bayesian inverse problem for 1-D Earth structure. To test the method, we compared the inferences on model parameters, i.e. 1-D marginal posterior pdfs, for the time-domain and encoded waveforms. The differences between the results for the two input types seem to be small for the examples shown here, which is promising. Moreover, the third example shows that the setup with the encoded waveforms can be advantageous over the time-domain alternative if the dimension of the input data space becomes large. The loss of information due to the encoding was outweighted by the improved MDN performance.

We conceive that this advantage becomes more prominent if we use more data, i.e. use more sources, receivers or seismogram components. This brings us to the crucial issue of the source-receiver configuration; data selection requires a lot of attention if we want to extend the setup to produce a robust constraint on 1-D Earth structure. Thus, directly inverting for 3-D Earth structure or seismic source parameters may be a more suitable application, given the nature of seismic waveform data. Such applications may also benefit from a data dimensionality reduction technique as we investigated here.

9

Conclusions and perspectives

Ideally, the assessment of solution quality is an integral part of any inverse method. The main motivation for this thesis was to investigate a means to simultaneously infer Earth structure and quantify the uncertainties in our estimates. Applications thereof are contained in Chapters 3 to 8. The results were discussed in detail and summarised at the end of each chapter. These final pages are meant to provide a broader perspective and discuss methodological developments and future applications. First, I address the contribution of the work presented in this thesis, in the context of the objectives formulated in Chapter 1. Second, I highlight the main findings on 1-D Earth structure and the seismic data sets used. As is so often the case, there is room for criticism and improvement. Therefore, I discuss some advantages and limitations of the neural network method with respect to other sampling-based techniques. Finally, I provide some perspective for future seismological studies using neural networks.

9.1 Contributions of the thesis

The solution to a non-linear seismic inverse problem is non-unique. It is often difficult to solve the non-linear problem directly, let alone quantify the uncertainty in the solution. Consequently, most seismic inversions are based upon linear approximations and employ regularisation to stabilise the inverse problem. If the linear assumption is valid and the inverse problem is well-posed, the best-fitting solution, as obtained from a conventional linearised inversion, and the most probable model in a posterior pdf, as obtained using a sampling-based approach, will in general be similar. If not, the best solution may be corrupted due to the invalidity of the linear approximation,

while the regularisation may introduce prior information into the system that has no physical basis and may thus bias the solution.

With the methodology developed in this thesis, no linearisation is required; the full non-linearity of the forward problem can be taken into account. I note that the neural network approach is also usable if the forward problem itself is linear. For such a relatively simple application, the full potential of the method is simply not used. In addition, the final solution is not affected by subjective regularisation criteria. The Bayesian framework enables us to obtain complete statistical information on earth model parameters, represented by posterior probability density functions. The posterior pdfs represent all information that is contained in the data and the prior model distribution. Instead of presenting a single 'best-fitting' earth model, these marginal pdfs provide a means to assess the likelihood of features in the earth model. Equivalently, they can be used to test hypotheses on Earth structure in a robust, quantitative manner. Furthermore, the method facilitates a quantitative comparison of the information contained in measurements of different seismic observables. Similarly, one can assess the constraint on different seismic parameters, e.g. compare the information for the average density at different depths, or for V_P and V_S in the same region. As such, MDNs can be used to analyse the sensitivity of the data to specific features in the earth model. I will discuss this advantage further in Section 9.2.1. In summary, the method presents a quantitative framework for solving non-linear inverse problems, meeting the first two objectives outlined in Chapter 1.

9.1.1 Inferences on radial Earth structure

Spherically symmetric earth models are often used as a reference for earthquake location determination and 3-D seismic tomographic models. Therefore, the quality of the latter is intrinsically linked to the robustness of the former and it becomes crucial to assess the quality of 1-D earth models. However, it is impossible to compare existing radial earth models, since they lack a quantitative estimate of their uncertainty or non-uniqueness. The third objective of this thesis was to obtain complete statistical descriptions of features of radial Earth structure, in terms of elastic and anelastic structure, anisotropy and depths of major discontinuities. In general, I conclude that a lot can still be learned on 1-D Earth structure from seismic data; ideally, we do so prior to tackling the 3-D tomographic problem. I did not construct a new radial earth model that could be used as a reference in subsequent applications. Doing so in a Bayesian framework is challenging using sampling-based methods, given currently available computational facilities. Alternatively, it may be worthwhile to adopt a classical (linearised) joint inversion of all currently available data sets and construct a new full 1-D earth model, while taking note of the probabilistic inferences on radial Earth structure, such as were made here. The main findings can be summarised follows.

In Chapter 3, we assessed the constraint on radial P-wave velocity structure provided by travel times of major body-wave phases from the EHB bulletin. This provides an independent validation of the velocity in the core and lower mantle in the *ak135* model, which is often used in 3-D seismic tomography and earthquake location

algorithms. By contrast, the data contained little or no information on the depths of major discontinuities and the P-wave velocity in the D" layer, the upper mantle and the homogeneous crustal layers, in agreement with the teleseismic nature of the data ($>25^\circ$ epicentral distance). Note that for instance the PKP and PKIKP datasets may not sample the ICB homogeneously, as they are mostly recorded along a few distinct ray paths, e.g. Romanowicz et al. (2003), which may cause a bias in the 1-D earth model.

We extended the analysis to degree-zero spheroidal mode splitting function data and a more complete *PREM*-like parametrisation of 1-D Earth structure, in terms of P-wave and S-wave velocities, density, bulk and shear attenuation (Chapter 4). The analysis of the information content suggests that the free oscillations constrain most parameters better than the body wave travel time data used in Chapter 3. Nonetheless, a joint inversion of normal modes and travel times may be beneficial if we are able to construct a more complete noise model for the travel times and include additional seismic phases, such as *PKKP*, *PcP* and converted phases like *SP* and *ScP*. Incorporating travel times for reflected phases, such as *PKiKP*, *PcP* and *ScS*, should enhance the constraint on the depth of seismic discontinuities. In addition, more information on Earth structure may be available when using travel time picks of higher quality than the data in the ISC (and EHB) catalogues, as these bulletins contain measurements of varying precision.

We assessed the robustness of features in existing reference models, such as *PREM*. The data prefer an ICB that lies in the depth range 5154.7–5165.7 km, i.e. deeper than in existing reference models. Although less pronounced, this result is robust with respect to an anisotropic parametrisation (5149.9—5165.7 km). Similar to the analysis shown in Figure 5.4, we found that the most likely ICB depth is still larger than in *PREM* if the standard deviation of the Gaussian noise model is increased by a factor of $\sim$5. A joint inversion of spheroidal and radial modes for the ICB depth yielded a similar result (Chapter 7). Given the robustness with respect to an anisotropic parametrisation, the number of training patterns and the amplitude of the assumed data noise, we conclude that the signal comes from the centre frequency data and is not an artefact of the method. Whether the ICB truly lies a few kilometres deeper, or whether the data are sensitive to a region of finite width ($\sim$tens of kilometres), is an intriguing question and deserves further attention. We found that the effect of a deeper ICB on the travel time of inner-core–sensitive seismic phases is comparable to the estimated noise in such measurements, and may thus be hard to detect using this data type. One way to improve the sensitivity of the body-wave data to the ICB could be to simply sample measumements at a higher density than in for instance Chapters 3 and 4, where the epicentral distance interval between travel time measurements was relatively wide (2°). An additional advantage of using complete travel time curves, for instance for PKP, is the inclusion of any triplications due to caustics, which result from strong increases or contrasts in velocity and could help to constrain the depth of seismic discontinuities. Finally, including travel times for reflected phases, such as *PKiKP*, may enhance the resolvability of the ICB depth, provided that the error in the measurements is smaller than the effect of the ICB depth shift on the travel times.

The marginal pdfs allow for a complete assessment of the probability for velocity and density contrasts at discontinuities, instead of a single value. The 1-D pdfs for the ICB density jump incorporate all previous estimates, while the presence of the "220" discontinuity could not unambigously be determined. The radially averaged density in the lowermost mantle is higher than in *PREM*, but its depth extent could not be uniquely determined with the spheroidal mode data. Shear attenuation parameters in the mantle deviate from *PREM* in a similar fashion to results from more recent studies. We found a non-homogeneous shear attenuation in the inner core, reinforcing the hypothesis that a distinct 'innermost inner core' may exist. The bulk attenuation in the mantle and the outer core is stronger than in *PREM*.

In Chapter 4, we observed trade-offs between parameters, mainly in regions that are believed to be anisotropic, such as the D" region. This formed the motivation to jointly invert complementary spheroidal and toroidal centre frequency measurements for fully anistropic earth models in Chapters 5 and 6, which resulted in a much stronger constraint, especially for shear-wave structure. In the upper mantle, we only find evidence for P-wave anisotropy in the transition zone and in the anisotropic "220–Moho" region. Our results partly agree and disagree with previous studies on radial anisotropy in 1-D earth models. In future work, complementary data types, such as surface wave phase velocities, body wave travel times and long-period waveforms, may enable us to resolve remaining ambiguities in upper mantle structure.

Using the joint data set, we investigated the radially averaged structure in the lower mantle in Chapter 6. This enabled us to constrain the depth extent of the density excess in the lowermost mantle. Furthermore, we showed, for the first time, that the average lower mantle is anisotropic below 1900 km depth, challenging the consensus that this part of the mantle is isotropic. We can explain our seismic observations with currently available mineral physics data for lower mantle minerals. Our results are compatible with a simple thermochemical model of the lower mantle that is on average about 100–200K colder than commonly-assumed adiabats and that consists of a mixture of about 60–65% perovskite and 35–40% ferropericlase containing 10–15% iron. By doing so, we showed that the observed anisotropy can be the result of LPO, which is most likely caused by deformation through dislocation creep (Karato, 2008). The observed seismic anisotropy does not support purely superplastic flow in the lower mantle and the associated diffusion creep, as this does not lead to the development of LPO in mantle minerals (Karato et al., 1995). Therefore, seismic anisotropy, such as observed here, can provide constraints on mantle flow and deformation mechanisms. However, meaningful geodynamic interpretations require a full 3-D analysis to be made.

We assessed similar hypotheses on anisotropy in the inner core using a joint data set of spheroidal and radial mode centre frequencies (Chapter 7). The only two models that allow for radial anisotropy in the top of the inner core are the heterogeneous growth model by Yoshida et al. (1996) and solidification texturing, i.e. solidication of outer core material at the ICB, related to a predominantly radial direction of maximum heat flux. Proof of radial anisotropy in the top of the inner core can help us discriminate between these and other existing dynamic models, such as aspherical

inner core growth and internal convection due to thermochemical variations, and thus improve our understanding of core dynamics and solidification processes. In the top ∼115 km of the inner core, our inferences are compatible with proposed radial anisotropy, for which $V_{PV} > V_{PH}$, but we cannot rule out an isotropic structure in this region. Further, we find strong positive P-wave anisotropy in the deep inner core, increasing with depth, while the average density is likely lower than in both *PREM* and *ak135f*. Trade-offs between parameters (V_P and ϕ) in the deep inner core prevent us from making unambiguous inferences on the radial anisotropy. Given the nature of the earth model parametrisation used here, we can only conclude that the centre frequencies require anisotropy in the (deep) inner core, in agreement with previous studies, and may provide further evidence for a distinct innermost inner core.

Finally, we illustrated one pragmatic approach to data dimensionality reduction, which is an important challenge in machine learning (Chapter 8). Our method using autoencoder networks is approximate and should be seen as a first step towards non-linear seismic waveform inversion using encoded seismograms. Foremost, we find that the encoded seismograms can be used to perform a Bayesian seismic waveform inversion for 1-D Earth structure. Moreover, we show that the setup with the encoded waveforms can be advantageous over the time-domain equivalent. The relative advantage is expected to increase for inversions using more data, e.g. more stations and sources. Given the nature of seismic waveform data, inverting for 3-D Earth structure or seismic source parameters may be more suitable applications, while benefitting similarly from such dimensionality reduction methods.

9.2 Advantages and limitations

In the above, I outlined the main advantages of non-linear Bayesian sampling-based strategies over conventional linearised inversion techniques. These advantages pertain to both neural network and Monte Carlo methods. Therefore, in this section, I highlight the differences between the two approaches, and discuss the advantages and limitations of the MDN methodology. Limitations shared by all sampling-based techniques are the complexities arising when sampling high-dimensional spaces and the required computational resources, which may be prohibitively large, even for earth models with only a few tens of parameters. For reference, consider constructing an ensemble of 48 MDNs for a given 1-D target parameter. Training time for such an ensemble on 12 processors simultaneously is on the order of 1–2 days for the applications in this thesis. This is a few orders of magnitude more than required by typical linearised inversion schemes.

Admittedly, the comparison of the two types of methods in this thesis has been rather qualitative and intuitive, only highlighting a notion of conservativeness in the pdfs output by the MDNs (Sections 4.6.5 and 5.4.1). A numerical comparison of the performance of both methods on the same problem is required to facilitate a full quantitative analysis of the pros and cons of each. One point in favour of Markov Chain Monte Carlo (MCMC) methods is that they are built on a mathematical proof of con-

vergence, i.e. the search algorithm will ultimately converge to the true posterior pdf, if appropriate rules are followed in the process, e.g. Mosegaard and Sambridge (2002). The MDN method lacks such a proof. To this end, we recently compared the MDN approach (Käufl et al., 2015b), which uses samples from the prior model space, to a more traditional MCMC approach using the Metropolis-Hastings sampling algorithm, which searches for samples of the posterior pdf (Hastings, 1970). We find that the posterior uncertainties output by the MDN, obtained using prior sampling, can be considered conservative estimates of the uncertainties that are obtained by directly sampling from the posterior distribution using Monte Carlo methods. Moreover, we show that the MDN method can provide an unbiased and conservative estimate of the marginal posterior pdf $p(\mathbf{m}'|\mathbf{d}_{obs})$, conditioned on the observed data $\mathbf{d}_{obs}$, in the case of a realistic point source parameter estimation problem. Clearly, this analysis is not equivalent to a proof of convergence, but it supports the notion of conservativeness of MDNs developed throughout this thesis and we feel this is an insightful first step towards a more quantitative comparison of the performance of the two methods.

9.2.1 Advantages

The machine learning approach differs from more common random sampling methods in a few distinct and advantageous ways. First, the neural networks were trained using samples from the prior model space. By contrast, Monte Carlo methods aim to sample from the posterior model space directly. If computationally feasible, the latter approach would be preferable. However, sampling directly from the posterior model space can prove difficult; search algorithms not uncommonly arrive in local minima of a cost function, or fail to converge to a global minimum. If the mapping of interest is sufficiently smooth, the interpolating capacities of neural networks prove useful. A trained neural network represents a continuous function that can interpolate between (input) samples and can thus be applied to new unseen data without the need to resample the model space. Furthermore, in my experience the approach gives a conservative estimate of uncertainties in model parameters. By definition, the prior model space is equal or larger than the posterior model space. Consequently, the average distance between prior samples is larger than between posterior samples, given the same number of models. The network interpolates between these prior points following a smooth mapping. Simple experiments with larger training sets (Sections 4.6.5 and 5.4.1) support this notion of conservativeness; in no case the marginal pdf was wider, or the information gain lower, for the larger training set. As highlighted above, the recent study by Käufl et al. (2015b) arrived at a similar conclusion. This is desirable, since we want to minimise the possibility that a relatively narrow pdf rejects the true Earth as a possible explanation for the observations. In conclusion, the MDN ensemble output reflects both the uncertainty in the earth model parameter and, to first order, the uncertainty in the inverse mapping approximated by the neural networks.

Second, since the training patterns are sampled from the prior model space, the inferred mapping does not depend on the observed datum, which is used as the reference for data misfit evaluations in Monte Carlo methods. Thus, we can investigate

whether or not the seismic data are sensitive to Earth structure, given the assumed data noise level, using only synthetic examples. Given the quantitative nature of the method, such a sensitivity analysis could provide a complete overview of the information in the data on features in the earth model. Furthermore, once the observed data are available, they can readily be inverted, since we already possess a trained neural network that approximates the inverse mapping. Note that the prior sampling and the associated independence of the MDN training from the observed data can also be a disadvantage, as I will discuss in Section 9.2.2.

Third, the approach is flexible, in the sense that we can use the same set of prior samples from the model space to perform multiple 'inversions', i.e. train networks on many different target parameters using the same synthetic data set. By contrast, when using Monte Carlo methods, one first has to sample the posterior model space to obtain the full posterior pdf and subsequently marginalise over all other model parameters. Furthermore, neural networks can be trained on any combination of earth model parameter(s). Thus, if we wish to invert for the average density in five lower mantle layers, we do not need to impose constant-value layers in the earth models, separated by fictitious discontinuities. MDNs can be trained directly on such radial averages, while the underlying 1-D earth models in the training set are realistic (smooth). In addition, by being able to invert for multiple parameters using a single training set, neural networks make efficient use of computational resources, in terms of the required number of solutions to the forward problem.

In relation to the flexibility and independence of the observations, another advantage of neural networks is that—once trained—they can be applied rapidly and repeatedly to multiple observations. They can for instance be used to invert observations for the location and characteristics of many seismic sources, e.g. Käufl et al. (2014, 2015a). Note that the applications in this thesis do not benefit from this feature, as I will discuss in Section 9.3.2. Examples of structural inversions exploiting this repeatibility include Meier et al. (2007b), who applied MDNs to surface wave phase and group velocities in each point on a global 2-D grid, and Shahraeeni and Curtis (2011), who inverted seismic velocity estimates for petrophysical parameters for each element in a 3-D grid. For such applications, only one neural network needs to be trained, after which the actual 'inversions' can be performed in a fraction of a second. By contrast, a Monte Carlo algorithm would need to be run for each grid point, due to the aforementioned dependence of the model space search on the real measurements.

In summary, the main advantages of artificial neural networks relate to their ability to handle non-linearities in the mapping of interest, robustness with respect to data noise and the ability to recognise patterns or extract features, that is, patterns that are not perceptible by direct human analysis or more conventional linear statistical methods. As such, neural networks can be very useful in situations where the forward relation is known, but the inverse mapping is unknown or difficult to establish by more conventional analytical or numerical methods. This situation applies to many geophysical inverse problems. Furthermore, neural networks are flexible and can be applied repeatedly to many observations without requiring retraining. Finally, the number of free parameters in artificial neural networks scales favourably with the size

of the input, compared to statistical modelling techniques that use fixed basis functions (Bishop, 1995). This made it computationally feasible to apply neural networks to the non-linear regression problems addressed in this thesis.

9.2.2 Limitations

One of the most common criticisms of neural networks is their nature as 'black boxes', which have no direct relation to the true mapping being approximated. An example is the classification of letters and digits in handwritten addresses on postcards. A neural network can be successfully trained on this task, but an analytical or intelligible description of the approximate mapping is unavailable. While such criticisms are understandable, I argue that for the type of application considered in this thesis this perceived drawback is manageable. In seismology, the relation between earthquakes and the Earth on the one hand and the recorded seismic data on the other hand is accurately captured by the physical laws of seismic wave propagation. In other words, we have a very good analytical understanding of the forward relation (Equation 2.2) and can generate exact training examples of earth models and seismic data using seismic forward simulation codes. Even though the neural networks are trained directly on the inverse relation, I feel that our understanding of the physics of the problem alleviates the issues of a mysterious 'black box' applied to a data set.

A second drawback relates to the sampling of the prior model space, as opposed to the posterior sampling of Monte Carlo methods. In Section 9.2.1, I argued that the prior sampling and the associated conservativeness of the MDN output can be an advantage. The marginal posterior pdfs do not show an overly optimistic inference; rather, they represent a lower bound on the information contained in the data and prior distribution. However, the argument can be turned in disfavour of neural networks. First, from the sampling-perspective, drawing samples from the prior can be highly inefficient, as we are in general not interested in the whole prior model space. Our main interest goes out to that part of the prior where the model resides that generated the observed data $\mathbf{d}_{obs}$, i.e. the posterior model space $p(\mathbf{m}|\mathbf{d}_{obs})$. Monte Carlo methods try to locate the latter via some guided search of the model space and by doing so make more efficient use of available computational resources. If successful, the resulting posterior pdf may be narrower, i.e. reflect a stronger constraint on model parameters, given the same data as was available to an MDN. In other words, the conservativeness of the neural network approach may prevent us from extracting all available information from the data. In this sense, it becomes important to assess the efficiency of both methods, i.e. compare the difference between the MCMC result and the conservative MDN estimate in light of the difference in the computational resources required for both methods. Such an efficiency analysis can expand on the numerical comparison performed in Käufl et al. (2015b) and should be the scope of future work. A second limitation of the prior sampling is the requirement that the inverse mapping is smooth. It is difficult to construct a representative set of samples from a high-dimensional model space (tens of dimensions or more). While this is true for any sampling-based method, it becomes more relevant in the context of the

interpolation between available samples performed by neural networks. Neural networks represent a smooth mapping that interpolates between available samples, or rather extrapolates to parts of the prior model space where no training samples are available. Arguably, if the true inverse mapping is strongly non-linear, the smooth interpolation between prior samples may not detect the detailed variations in the true mapping and consequently predictions may be inaccurate. However, it is possible to detect potential problems related to the validity of the smoothness assumption. In any application in this thesis, network prediction accuracy was assessed using independent test samples. Obviously, these test samples themselves are samples from the prior model space. Therefore, network performance was also analysed using *PREM*, for which I found predictions to be accurate. In fact, an accurate performance on *PREM* served as a prerequisite for applying the trained MDNs to the observed data.

A third limitation is that an MDN does not produce the full posterior pdf. Although in theory it can be trained on the full earth model as the target, in practice such a network is difficult to train. Alternatively, the full posterior pdf can be found by training networks that take a model parameter as an additional input, and subsequently multiplying the resulting conditional and marginal pdfs (Equation 3.2). However, such a procedure is again quite complicated for the model parametrisation used in Chapters 4 to 8, which contains hundreds of dimensions. Instead, only 1-D and 2-D marginal posterior pdfs were considered in this thesis. The 2-D pdfs allow one to assess trade-offs between two particular parameters, but potential higher-order (i.e. $>$2-D) trade-offs were not assessed. I emphasise that this does *not* mean that any existing trade-offs are not accounted for. Model parameters, such as velocity, density and anisotropic parameters are independently varied (Appendix A) and no fixed scaling relations are imposed between parameters. The 1-D marginal pdfs will incorporate the effect of any trade-offs in the full model space. This causes the 1-D marginals to be wider, reflecting the increased uncertainty due to the existence of trade-offs.

9.3 Perspectives

Some interesting research avenues where not explored in this thesis, but are discussed here. First, I consider potential improvements of the neural network method. Second, I highlight several applications that may exploit more of the method's advantages, while resolving its limitations.

9.3.1 Potential improvements

A first improvement could be made by addressing the choices made regarding the network architecture. In this thesis, the network architecture was fixed *a priori*. From the Bayesian perspective, ideally one allows for variations in the network configuration, such as the number of hidden units, layers or Gaussian kernels. Then, the posterior model pdf $p(\mathbf{m}'|\mathbf{d}, \mathbf{w}, \mathcal{A})$ is conditioned on the data $\mathbf{d}$, the network weights $\mathbf{w}$ and the network architecture $\mathcal{A}$, which represents all choices made regarding the

neural network settings and the training process. By doing so, we reduce the possibility that our posterior inferences are affected by subjective decisions made prior to interrogating the data. If our interest goes out to the marginal posterior model pdf, as was the case in this thesis, we can in theory integrate over 'network architecture' space, similar to the weight space integration in Equation 2.25, i.e.

$$p(\mathbf{m}'|\mathbf{d}) = \int p(\mathbf{m}'|\mathbf{d}, \mathbf{w}, \mathcal{A}) p(\mathbf{w}|D, \mathcal{A}) p(\mathcal{A}|D) \mathrm{d}w \mathrm{d}\mathcal{A}. \tag{9.1}$$

As before, performing such an integration is difficult; how does one define the full space of network architectures $\mathcal{A}$ to begin with? A more pragmatic approach is to replace the full network architecture in Equation 9.1 by so-called *hyperparameters*, such as the width (variance) of the prior weight or data noise distributions, which are used to control the complexity of the network mapping (MacKay, 1992, 2003). The hyperparameters themselves are allowed to vary following some *hyperprior*. Since the values of the hyperparameters control for instance the data noise prior, the approach considers multiple levels of inference and is referred to as *Hierarchical Bayes*. In such a case, the posterior probability distribution reflects for instance the uncertainty in the measurement errors and the specific network configuration. In seismology, such a hierarchical method has been adopted for Bayesian approaches to seismic tomography (Malinverno and Briggs, 2004; Bodin et al., 2012).

In the above, I used the number of hidden neurons as one example of a parameter that can be varied in the setup. Naturally, a similar concept can be applied to the number of earth model parameters, which is treated as a free parameter in so-called transdimensional inversions, e.g. Sisson (2005); Bodin et al. (2012); Sambridge et al. (2013). Such an approach could be implemented in the MDN methodology by varying the number of parameters in the earth models generated for the training set. However, this would increase the size of the prior model space. In light of the limitations of the prior sampling, as were addressed above, care must be taken that the training set is adequately representative of the enlarged prior model space.

A final suggestion relates to the seismic data sets used to solve inverse problems. In this thesis, I used body wave travel times, normal mode centre frequencies and quality factors and, ultimately, seismic waveforms. An obvious improvement, resulting in an increase in information, may be achieved by incorporating complementary data types, such as surface wave phase and group velocities, as suggested for future studies on upper mantle structure in Chapter 5. Furthermore, one may consider the nature of the centre frequency and quality factor data, which relate to the degree-zero of normal mode splitting functions. These splitting function coefficients themselves are the result of an iterative and regularised non-linear inversion of normal mode spectra. From a Bayesian viewpoint, using the centre frequency data as a starting point corresponds to the implicit assumption that the inversion of the mode spectra was exact, which seems unrealistic. Ideally, one inverts these raw spectra directly for Earth structure, at the expense of an increase in the (data) dimensionality of the problem. Whether or not this is feasible computationally could be the topic of future work.

9.3.2 Future applications

One of the aforementioned advantages of neural networks is that—once trained—they can approximate a mapping that can be applied repeatedly to multiple data. The applications of neural networks in this thesis do not involve a repetitive component and thus do not exploit the full potential of neural networks. Since this can be a tremendous advantage of neural networks over Monte Carlo methods, in terms of computational efficiency, future neural network-based studies can benefit greatly if they involve a degree of repeatibility. Successful applications include the aforementioned inversions for seismic sources, e.g. Käufl et al. (2014, 2015a), and Earth structure, e.g. Meier et al. (2007b); Shahraeeni et al. (2012). I envision that a similar benefit could be achieved in 3-D structural inversions using for instance receiver functions, e.g. Bodin et al. (2012); Shen et al. (2013) and Shen et al. (2013). Obviously, the drawbacks of neural networks persist, as outlined above, but may be outweighted by the ability to solve Bayesian non-linear inverse problems rapidly and repeatedly. In this context, a recent suggestion by Walker and Curtis (2014) to allow for spatial variation in prior information may be useful. Often, the amount of available independent information varies throughout a model, i.e. differs per grid point. For instance, direct geological observations may provide us with a greater prior constraint on shallow (crustal) structure. Walker and Curtis (2014) propose to train an MDN using a general, i.e. conservative, prior model distribution and subsequently multiply the MDN output for each grid point with the ratio of the point-specific and general prior distribution. By doing so, one appreciates the spatial variation in the data-independent information.

Surrogate modelling of seismic data misfit functionals

Throughout this thesis, and in Chapter 8 in particular, I have addressed the issue of (a too large) data dimensionality. The preceding chapter showed one pragmatic approach of reducing the dimensionality of seismograms, while preserving their information on Earth structure. Another route to tackling this issue may be found by redefining the problem. That is, instead of inverse mappings, we could approximate the forward problem, or more directly, the data misfit functional using machine learning methods. A way forward (no pun intended) may be offered using *surrogate* or *response surface* models, which are fast approximations to computationally demanding simulations (Forrester et al., 2008; Myers et al., 2009). In fact, the applications in this thesis can also be classified as surrogate models, in the sense that the trained neural networks represent a surrogate for the inverse mapping. When using surrogates, the aim is to simulate the outcome of (physical) experiments, given some input model. As such, a surrogate model is a second-level abstraction, i.e. a model of a simulation model, which itself captures some relation, such as elastic wave propagation in the case of seismology. Applications are widespread and include optimising airplane wing design in aerospace engineering, calibrating groundwater models or reducing costs in manufacturing processes, e.g. (Queipo et al., 2005; Razavi et al., 2012).

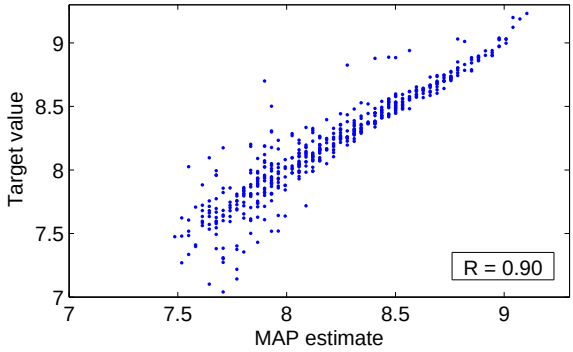


Figure 9.1: Maximum a posteriori (MAP) probability estimate versus the target value for all 500 patterns in the test data set. An ensemble of 10 MDNs was trained on 484-D earth models as input and the base-10 logarithm of the L_2-norm of the data misfit for the 184-D spheroidal mode centre frequency data as the 1-D target. The corresponding correlation coefficient R is given.

In a sampling-based framework, the data misfit is evaluated for each model realisation and thus the forward problem needs to be solved many times. A surrogate model could mimic this process, taking an earth model as input and making a prediction for the corresponding data misfit in a fraction of a second. If successful, the surrogate model can be used to evaluate the misfit for many more earth models and thus may enable us to perform (part of) a grid search of the model space, which would otherwise be infeasible for high-dimensional models, i.e. models with tens of parameters or more. The biggest challenge is to construct a surrogate model that accurately predicts the data misfit for models throughout the prior model space. This typically requires an initial set of samples from the prior model space, similar to the training of neural networks in this thesis. In fact, artificial neural networks are commonly used in for surrogate modelling.

As an example, I trained an ensemble of 10 MDNs using a set of 10 000 synthetic samples. The input to the MDNs consisted of the 484-D earth models for the partly anisotropic parametrisation (Appendix A). The 1-D target was given by the base-10 logarithm of the L_2-norm of the data misfit (Equation 6.1) for the synthetic centre frequencies, with respect to the 184 spheroidal mode measurements and their errors, as used in Chapters 4 to 7. In doing so, I exploit one aforementioned advantage of neural networks, as the input and output have been interchanged with respect to the inverse mappings considered earlier. Figure 9.1 shows the target value for all patterns in the test data set versus the Maximum A Posteriori (MAP) probability estimate of the MDN output, as was done before to test MDN prediction accuracy (e.g. Figures 3.9 and 4.2). As evidenced by the diagonal distribution of the samples and the high correlation coefficient $R = 0.90$, the MDNs make accurate predictions for the data misfit of the test set samples, given the earth models as input. Note that in a standard surrogate

modelling configuration, a conventional MLP may suffice, as we are only interested in a single prediction for the misfit value at this stage.

Naturally, the successful predictions in Figure 9.1 are nothing more than a promising starting point. Constructing an accurate surrogate model is by no means straightforward, and one can expect similar issues as with any other sampling-based methods. First, I note that none of the 500 earth models in the test set generate a misfit that is close to that of for instance *PREM* (5.8), let alone a near-zero misfit, i.e. an accurate prediction of the measured data, given the data noise. Given this limitation of an initial set of prior model samples, an iterative process is commonly adopted in surrogate modelling. After initial training, a surrogate can be used to predict the misfit values for earth models found through a grid search. For a subset or all of these models, depending on the computational budget, we can calculate the true data misfit, which requires running an expensive forward simulation for each. By evaluating the misprediction of the surrogate model, we can identify regions of model space where prediction accuracy is lowest and thus needs to be improved. Subsequently, we would add models from these regions to the training set for the second iteration of surrogate construction. Thus, the performance of the surrogate is iteratively refined. Eventually, the surrogate model may be usable for misfit predictions for the most relevant input earth models, i.e. the region of earth model space that lies close to the true Earth. Clearly, this iterative scheme requires more forward simulations and thus more computational resources. Whether or not the increased computational costs are outweighted by the ability to perform very fast forward calculations using a surrogate model, will depend on the application and demands thorough investigation. Naturally, the performance of the surrogate-based Bayesian inversions has to be compared to alternative sampling-based strategies, such as Monte Carlo methods, that seek a similar representation of the posterior model distribution.

9.3.3 A final remark

The machine learning method adopted in this thesis is a pragmatic approach to solving non-linear Bayesian inverse problems. Ultimately, no sampling-based method or algorithm escapes the aforementioned complexities arising when sampling high-dimensional data and model spaces. The best we can do is to develop algorithms that use the available computational resources as effectively and efficiently as possible when solving our inference problems. Future applications of MDNs should invoke some repeatability of application, if they are to be used as a computationally more efficient replacement of Monte Carlo methods. This would seem to preclude MDNs from being used directly for (global) 3-D structural inversions, which commonly do not require repeated inversions. The only type of '3-D inversions' which exploits the full potential of the method, in my opinion, is the type in which the actual response of the data to the model parameters is local. As mentioned above, this is the case in the applications of for instance Meier et al. (2007b); Shahraeeni and Curtis (2011), which essentially involve repeatedly solving many 1-D inverse problems (one at each grid point in the model), and could be useful for similar 3-D structural inversions using

for instance receiver functions. This is not a feature of seismic waveform inversions, as waveform data are sensitive to many (adjacent) parameters in a 3-D model. Therefore, neural networks may simply be less suitable to perform 3-D seismic waveform inversions, which are the current state-of-the-art in seismology, in a Bayesian framework. Nonetheless, 'one-time' applications, such as were considered in this thesis, are insightful. The trained neural networks have been extensively evaluated throughout and their prediction has been shown to be accurate, facilitating robust and quantitative data sensitivity analyses and hypothesis tests for radial Earth structure, including uncertainty estimates.

In short, the use of artificial neural networks can be advantageous when solving geophysical inverse problems. Their flexibility and interpolation capabilities, in combination with the quantitative Bayesian framework, make MDNs well-suited for data sensitivity analysis and testing hypotheses on Earth structure. Rather than a replacement for Monte Carlo methods, I suggest that in the future they are used as a complementary tool, providing an initial assessment of data sensitivity and a lower bound on the information on model parameters that is contained in the data. As such, one could envision a setup in which an MDN provides an initial conservative estimate of the (marginal) posterior pdf, which can be used as a starting point for a subsequent MCMC search algorithm. If the MDN can extract information from the data, the new or updated prior model space will be smaller than the original prior. This may enhance the speed of convergence of the MCMC search and thus may make more efficient use of computational resources than a pure Monte Carlo approach. In general, neural networks can be used to improve our understanding of a data-generating system by observing the behaviour of the output relative to the input parameters. Naturally, a neural network cannot explain *why* the Earth generated such patterns; it merely detects them, facilitating additional analysis. As such, neural networks should not be considered as a replacement for physics-based (forward) modelling. Rather, a hybrid approach, which uses both physical modelling and machine learning, may be the optimal approach to analyse any physical system and patterns in the associated data.

A

Model parametrisation

This section describes the parametrisation of the 1-D earth models used in Chapters 4 to 6. The work presented in Chapter 3 involves a different style of parametrisation, as outlined in Section 3.2.

A.1 Velocity and density structure

We define V_P, V_S and ρ on a discrete set of 185 grid points (as one of the options in the Mineos package (Masters et al., 2011)), the depths of which range from the Earth's surface to its centre. We parametrise the depths of seven discontinuities in the model: the inner-outer core boundary (ICB) and core-mantle boundary (CMB), the top of the D" layer, the discontinuities around 660, 410 and 220 km depth ("660", "410" and "220", respectively) and the Moho. The lower mantle (LM) represents the region between the top of the D" layer and the "660", while the transition zone (TZ) spans the region between the "660" and the "410". For both velocities and density, the core and mantle are defined at 157 points in the earth model (Table A.1). The remaining 28 points in the earth model represent the crust, which is parametrised by two homogeneous layers. No sediment or water layers are present.

We consider two different parametrisations for anisotropy. In a first case, we include radial anisotropy only in the uppermost mantle between the Moho and the

Components of this appendix have been published, together with Chapter 4, as: de Wit, R. W. L., P. J. Käufl, A. P. Valentine, and J. Trampert, 2014. Bayesian inversion of free oscillations for Earth's radial (an)elastic structure. *Physics of the Earth and Planetary Interiors 237*, 1–17.

"220", as in *PREM* (Dziewoński and Anderson, 1981). A radially anisotropic, or transversely isotropic, medium has hexagonal symmetry with a radial symmetry axis and can be described by five independent parameters (Love, 1927). Here, we parametrise the anisotropy by the velocities of vertically and horizontally propagating P-waves (V_{PV} and V_{PH}), the velocities of vertically and horizontally polarised S-waves propagating horizontally (V_{SV} and V_{SH}) and the anisotropic parameter η on the nine points in this region. In the rest of the model, the wave velocities are isotropic and $\eta = 1$.

Tables A.2, A.3 and A.4 define the prior model distribution $p(\mathbf{m})$ for this first case. Discontinuity depths are independently drawn from uniform priors, as are V_{PV}, V_{SV} and ρ directly below the seven discontinuities and V_{PH}, V_{SH} and η below the Moho (Table A.2). The prior distributions are centred on the corresponding values in *PREM*. For the upper mantle and D", we chose priors similar to those of de Wit et al. (2013). We define narrower priors for the core and lower mantle, taking into account the constraint on these parameters provided by body wave travel times (de Wit et al., 2013). In the "220-Moho" region and directly below the "220" discontinuity, we allowed for the largest variations in our prior, since existing models, such as *PREM* and *ak135f* (Montagner and Kennett, 1996), vary strongly in this region.

To exclude physically implausible models from the prior model space, we introduce correlations between adjacent points (layers) in each region, i.e. between discontinuities. First, we draw the value of the independent point directly below a discontinuity. We then use this value and the local gradient in *PREM* to calculate the value for the underlying point. Subsequently, this value is perturbed, with the amount of perturbation drawn from a uniform prior (Table A.3). This procedure is performed for all the points successively with increasing depth and introduces a correlation between the parameters in each region. In general, the radial velocities and density increase with depth, i.e. the velocity and density gradients are mostly positive. The η profile in the uppermost mantle is constructed in a similar fashion.

To avoid physically unrealistic 1-D density profiles, we constrain the mass and moment of inertia of the earth models using the error estimates reported by Chambat and Valette (2001). A model is discarded whenever its mass or moment of inertia does not lie within $(5.9733 \pm 0.0090) \cdot 10^{24}$ kg or $(8.018 \pm 0.012) \cdot 10^{37}$ kg m^2, respectively. Figures 4.1 and A.1 show the parameter range spanned by the prior model space and a number of existing 1-D reference models. We emphasise that the prior ranges for η, ϕ and ξ, as shown in Figure A.1, are only applicable to the second (fully anisotropic) parametrisation; for the first parametrisation, $\eta = \phi = \xi = 1$, except in the uppermost mantle.

A.2 Radial anisotropy

In the second case of model parametrisation, we allow radial anisotropy in the whole mantle and in the inner core. The parametrisation is the same as for the uppermost mantle in the first case, i.e. in terms of V_{PV}, V_{SV}, V_{PH}, V_{SH} and η. Thus, the profiles for the four velocity components are constructed independently using the prior

Table A.1: Number of grid points L between each pair of adjacent discontinuities in the earth model. The full model is defined on a discrete set of 185 grid points.

Region	L
IC	33
OC	33
D"	5
LM	59
TZ	9
"410–220"	9
"220–Moho"	9
LC	11
UC	17

ranges specified in Tables A.2 and A.3 for V_P and V_S. The outer core is isotropic in both parametrisations. A radially anisotropic medium can be described by hexagonal symmetry with a vertical (radial) symmetry axis, density and the five independent Love coefficients A, C, N, L and F (Love, 1927). Three parameters are commonly used to describe the radial anisotropy: the P-wave anisotropy

$$\phi = \frac{C}{A} = \frac{V_{PV}^2}{V_{PH}^2}, \tag{A.1}$$

the shear-wave anisotropy

$$\xi = \frac{N}{L} = \frac{V_{SH}^2}{V_{SV}^2} \tag{A.2}$$

and

$$\eta = \frac{F}{A - 2L}, \tag{A.3}$$

which corresponds to anisotropy at intermediate incidence angles. For an isotropic medium, $\eta = \phi = \xi = 1$.

In addition to the three anisotropic parameters, we studied the density and the isotropic equivalent of the P- and S-wave velocities, which are given by the Voigt averages (Babuska and Cara, 1991; Panning and Romanowicz, 2006),

$$V_P = \sqrt{\frac{K + \frac{4}{3}G}{\rho}} \tag{A.4}$$

and

$$V_S = \sqrt{\frac{G}{\rho}}, \tag{A.5}$$

Table A.2: Prior information on *independent* model parameters. Prior distributions are uniform over the specified ranges, which are given as percentage perturbations from *PREM*, except for the discontinuity depths and the two crustal layers. V_P, V_S and ρ parameters represent the points located directly below a discontinuity. The tops of the lower mantle (LM) and the transition zone (TZ) are formed by the "660" and "410", respectively.

Discontinuity	Range [km]
ICB	$5129.5 - 5169.5$
CMB	$2871 - 2911$
D" layer (top)	$2721 - 2761$
"660"	$640 - 700$
"410"	$370 - 430$
"220"	$200 - 240$
Moho	$20 - 70$
V_P, V_S, ρ, η	Range [%]
Inner core (IC)	± 2
Outer core (OC)	± 2
D" layer	± 3
Lower mantle (LM)	± 2
Transition zone (TZ)	± 5
"410–220"	
V_P	[-10,+2.5]
V_S	[-10,+5.0]
ρ	[-10,+5.0]
"220–Moho"	± 7
Lower crust (LC)	
V_P [km/s]	$6.4 - 7.4$
V_S [km/s]	$3.6 - 4.1$
ρ [g cm^{-3}]	$2.8 - 3.0$
Upper crust (UC)	
V_P [km/s]	$5.6 - 6.3$
V_S [km/s]	$3.1 - 3.6$
ρ [g cm^{-3}]	$2.6 - 2.8$
Mass (10^{24} kg)	5.9733 ± 0.0090
Moment of inertia (10^{37} kg m^2)	8.018 ± 0.012

Table A.3: Prior information on *dependent* model parameters. Prior distributions are uniform over the specified ranges, which are given as percentage perturbations from the updated model value (see text). The corresponding independent parameters are listed in Table A.2.

V_P, V_S, ρ, η	Range [±%]
IC	0.5
OC	0.5
D"	1
LM	0.5
TZ	1
"410–220"	1
"220–Moho"	2

where the Voigt average bulk and shear moduli, K and G respectively, are defined as

$$K = (C + 4A - 4N + 4F)/9 \tag{A.6}$$

and

$$G = (C + A + 6L + 5N - 2F)/15. \tag{A.7}$$

with the five independent Love coefficients A, C, N, L and F. In Chapters 5 to 8, we always use these exact formulations for the Voigt average velocities, as opposed to the approximate formulations used in Section 4.5.3 (Figure 4.4). Note that this approximation to the Voigt average V_P and V_S is valid for under the assumption of small anisotropy, i.e. $\eta \approx 1$, and as such the results for the Voigt average velocities in the uppermost mantle in Chapter 4 are equally valid.

Finally, we emphasise that the results of the inversion are not affected by the choice between a parametrisation in terms of the Love coefficients or wave velocities and η, as our method is derivative-free. It is straightforward to extract the five Love coefficients from the polarised wave velocities and η in an earth model and calculate the corresponding ϕ, ξ and Voigt average isotropic wave velocities (Equations A.4 to A.7). Figure A.1 shows the prior range and *PREM* for all parameters (the Voigt average V_P and V_S, ρ, the anisotropic parameters η, ϕ and ξ and the attenuation parameters Q_μ and Q_κ.

A.3 Attenuation structure

The bulk and shear attenuation are parametrised by the inverses of Q_κ and Q_μ, respectively. We closely follow the parametrisation of Resovsky et al. (2005) and define the radial bulk and shear attenuation structure by 13 parameters. Q_κ is parametrised as four layers of constant attenuation: the inner core, the outer core, the lower mantle

Table A.4: Prior information on the attenuation parameters. Prior distributions are uniform on a base-10 logarithmic scale over the specified ranges.

Q_μ	Region	Range
	Inner core (IC)	
	6371–5760 km	$10 - 300$
	5760–5150 km	$10 - 300$
	Outer core (OC)	0
	Lower mantle (LM)	
	2891–2157 km	$100 - 1000$
	2157–1428 km	$100 - 1000$
	1428– 670 km	$100 - 1000$
	Transition zone (TZ)	$50 - 400$
	"410–220"	$50 - 400$
	Low-velocity zone (LVZ)	$20 - 200$
	High-velocity lid + crust	$10 - 2200$
Q_κ	Region	Range
	Inner core (IC)	$300 - 300\,000$
	Outer core (OC)	$300 - 100\,000$
	Lower mantle (LM)	$300 - 100\,000$
	Upper mantle (UM)	$300 - 200\,000$

and the upper mantle. The latter two are separated by the "660". Note that the depths of the discontinuities separating these regions are free parameters in the earth model. Q_μ is parametrised by two and three layers of roughly equal thickness in the inner core and the lower mantle, respectively, and is zero in the outer core. We add a second layer to the inner core compared to the parametrisation of Resovsky et al. (2005). The upper mantle consists of four layers, which represent the TZ, the "410–220" region, the low-velocity zone (LVZ) between the "220" and 80 km depth in *PREM* and a layer encompassing both the overlying high-velocity lid and the crust. The prior distributions are given in Table A.4 and shown in Figures 4.1 and A.1. No correlations exist between the 13 parameters and all priors are uniform on a base-10 logarithmic scale.

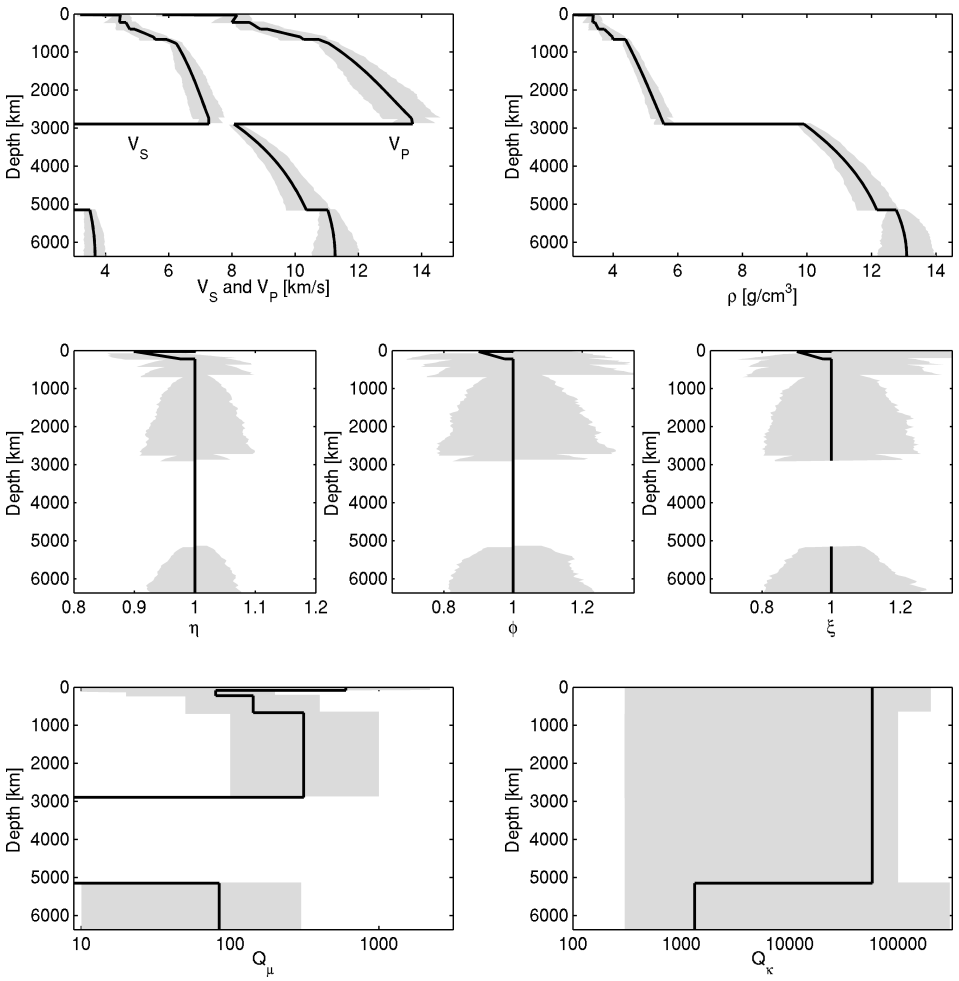


Figure A.1: Radial earth models contained in the model prior. The parameter range spanned by the prior model space is represented by the grey shaded area, along with the 1-D reference model *PREM* (black, solid) for the Voigt average isotropic V_S and V_P (top-left panel), ρ (top-right panel), η (middle-left panel), ϕ (centre panel), ξ (middle-right panel), Q_μ (bottom-left panel) and Q_κ (bottom-right panel). The horizontal scale for the panels showing Q_μ and Q_κ is logarithmic. Note that the outer core is isotropic.

B

Construction of polycrystal aggregates

In Chapter 6, we investigated whether our seismic results could be explained by a simple thermochemical lower mantle model. For this analysis, we constructed a total of 22 491 thermochemical models using a polycrystal aggregate that is to first order representative of the lower mantle. The composition of the lower mantle in pyrolitic models usually constitutes ∼75% orthorhombic Mg-perovskite ($MgSiO_3$), ∼10% cubic Ca-perovskite ($CaSiO_3$) and ∼15% ferropericlase ((Mg,Fe)O) (Ono and Oganov, 2005; Mainprice, 2007). Alternatively, the nearly isotropic Ca-perovskite (Li et al., 2006) is often ignored, leading to a model with ∼80% Mg-perovskite and 20% ferropericlase. Laboratory and first-principles modelling studies show that both minerals are anisotropic under lower mantle conditions (Karki et al., 1997, 2000; Oganov et al., 2001; Wentzcovitch et al., 2004). Furthermore, a substantial amount of iron appears to be present in the lower mantle, partitioning into perovskite and ferropericlase, and is assumed to have a significant influence on the elasticity and other physical properties of these minerals (Mao et al., 1979; Kobayashi et al., 2005; Sinmyo and Hirose, 2013), An additional complexity arises when aluminium-bearing perovskite is added to the equation (Irifune, 1994; Nishiyama et al., 2007). We restricted our analysis to a polycrystal aggregate of aluminium-free and iron-bearing perovskite and ferropericlase and provided a first-order explanation of our seismic observations.

For each of the thermochemical models, we constructed the polycrystal aggregate

The content of this appendix, together with that of Chapter 6, has been submitted to *Earth and Planetary Science Letters* as: de Wit, R. W. L. and J. Trampert, 2015. Robust constraints on average radial lower mantle anisotropy and consequences for composition and texture.

Table B.1: Variation of compositional model parameters and temperature. The temperature is represented by a deviation ΔT from the geotherm by Brown and Shankland (1981), which is given for the centre of the five lower mantle layers in Table B.2. All possible combinations of these four parameters were considered, resulting in a total of 22 491 different thermochemical models.

Parameter	Minimum	Interval size	Maximum	No. of options
X_{Pv} [%]	50	1	100	51
X_{Fe} [%]	5	1	25	21
K_D	0.3	0.1	0.5	3
ΔT [K]	-200	50	100	7

in the following way. First, we varied the volume fraction of perovskite X_{Pv}, the volume fraction of iron X_{Fe} and the iron partitioning coefficient K_D (Table B.1). K_D controls the partitioning of the available iron into perovskite (Pv) and ferropericlase (Fp) and is defined as (Deschamps and Trampert, 2003; Kobayashi et al., 2005)

$$K_D = \frac{x_{Fe}^{Pv}/(1-x_{Fe}^{Pv})}{x_{Fe}^{Fp}/(1-x_{Fe}^{Fp})}, \tag{B.1}$$

where x_{Fe}^{Pv} and x_{Fe}^{Fp} are the volume fractions of iron in perovskite and ferropericlase, respectively. The total iron content is given by

$$X_{Fe} = X_{Pv}x_{Fe}^{Pv} + (1-X_{pv})x_{Fe}^{Fp}. \tag{B.2}$$

Using these two equations and the values given for X_{Pv}, X_{Fe} and K_D, we calculated x_{Fe}^{Pv} and x_{Fe}^{Fp} for each thermochemical model.

Elasticity and density at lower mantle pressures were determined for pure-Mg orthorhombic perovskite (Wentzcovitch et al., 2004) and for cubic periclase (Karki et al., 2000). Wenk et al. (2006) used these estimates and associated temperature and pressure derivatives to obtain the elasticity at specified depths in the lower mantle. To first order, the elasticity and density change linearly with depth. Therefore, we linearly interpolated between the values reported in Table 2 of Wenk et al. (2006) to obtain the approximate elasticity and density at depths corresponding to the centre of our five lower mantle layers. This gave us the properties for the pure-Mg minerals at the five desired depths (Table B.2).

To correct the elasticity and density for iron, we used pure-Fe estimates for the two minerals (Jacobsen et al., 2002; Kiefer et al., 2002). Again, to first order the elasticity and density vary linearly between the Mg and Fe endmembers. Therefore, it is straightforward to calculate the properties of a (Mg,Fe)-mixture for perovskite and ferropericlase using the relation

$$a_{(Mg,Fe)}^{Pv,Fp} = a_{Mg}^{Pv,Fp}(1-x_{Fe}^{Pv,Fp}) + a_{Fe}^{Pv,Fp}x_{Fe}^{Pv,Fp}, \tag{B.3}$$

Table B.2: Elastic constants and density for pure-Mg perovskite and periclase. Listed depths correspond to the centre of our five lower mantle layers. The elasticity at these depths was approximated by linearly interpolating between the values reported in Table 2 of Wenk et al. (2006). Also shown is the temperature along the geotherm by Brown and Shankland (1981), which corresponds to $\Delta T = 0$ K (Table B.1).

$MgSiO_3$-perovskite					
Depth [km]	849	1242	1670	2099	2527
Density [g/cm^3]	4.453	4.658	4.880	5.085	5.276
C_{11} [GPa]	578.2	642.6	706.3	774.9	851.5
C_{12} [GPa]	249.4	309.6	375.4	447.6	524.5
C_{13} [GPa]	228.9	274.4	323.8	372.9	426.6
C_{22} [GPa]	651.7	742.1	840.6	947.4	1062.7
C_{23} [GPa]	249.5	296.2	346.8	400.8	458.2
C_{33} [GPa]	619.8	720.3	823.6	931.4	1044.2
C_{44} [GPa]	202.7	225.7	249.7	273.2	298.6
C_{55} [GPa]	189.6	203.2	217.1	233.4	250.2
C_{66} [GPa]	175.4	202.0	229.3	256.3	284.1
MgO (periclase)					
Depth [km]	849	1242	1670	2099	2527
Density [g/cm^3]	3.981	4.219	4.457	4.668	4.870
C_{11} [GPa]	496.0	635.2	787.5	945.2	1108.5
C_{12} [GPa]	144.6	169.1	195.8	222.7	249.7
C_{44} [GPa]	159.0	169.8	179.4	188.1	196.3
Temperature [K]	1934	2055	2174	2279	2375

where is $a^{Pv,Fp}$ is an element of the elasticity tensor or the density for Pv or Fp and $x_{Fe}^{Pv,Fp}$ is the volume fraction of iron for each of the two minerals, as calculated for each thermochemical model using Equations B.1 and B.2.

Second, we varied the temperature, in terms of a deviation from the Brown-Shankland geotherm (Brown and Shankland, 1981). We updated the elastic properties using the temperature derivatives $\partial a^{Pv,Fp}/\partial T$ estimated for perovskite (Wentzcovitch et al., 2004) and periclase (Karki et al., 2000) by applying the correction

$$a^{Pv,Fp} = a^{Pv,Fp} + T_{diff}\frac{\partial a^{Pv,Fp}}{\partial T}, \tag{B.4}$$

where $a^{Pv,Fp}$ is again an elastic property and T_{diff} is the temperature difference between our desired temperature, which is represented by a deviation ΔT from the Brown-Shankland geotherm (Table B.1) and the temperature used by Wenk et al. (2006) (their Table 2). The temperature along the Brown-Shankland geotherm, which corresponds to $\Delta T = 0$ k, is given for the centre of our five layers in Table B.2.

Third, we rotated the individual perovskite and ferropericlase crystals about the principal axes of their elastic tensors, which were aligned with the Cartesian reference frame (Walker and Wookey, 2012). We considered rotations about the two horizontal axes (x_1 and x_2) in 10 degree angles from 0 (no rotation) to 90 degrees. For each mineral, this gave 10^2 options; since we allowed the crystals to rotate separately, the total number of configurations became 10^4. Further, since rotations are not commutative, i.e. the order of rotation matters, we considered rotating in both orders $x_1 - x_2$ and $x_2 - x_1$, which gave a final number of $2 \cdot 10^4$ configurations for the rotation for each of the 22 491 thermochemical models.

Fourth, we calculated the Voigt average of the rotated elasticity tensors for the two minerals. The Voigt average assumes constant strain in the medium and places an upper bound on the true value of the elasticity (Voigt, 1910; Babuska and Cara, 1991; Mainprice, 2007). To facilitate a comparison with our seismic observations of radial anisotropy, we averaged the tensor for the polycrystal about the vertical axis (x_3) to impose vertical transverse isotropy (VTI), or radial anisotropy, using the approach by Walker et al. (2011); Walker and Wookey (2012).

Finally, we obtained a hexagonally symmetric elastic tensor for the polycrystal. It is trivial to extract η, ϕ, ξ and the Voigt average equivalent isotropic velocities V_P and V_S, as defined above, from the elastic tensor using the five independent Love coefficients A, C, N, L and F and the density ρ (Panning and Romanowicz, 2006; Mainprice, 2007). This enabled us to compare the values for the anisotropic parameters, density and wave velocities for each polycrystal aggregate with the 1-D marginal posterior pdfs that we obtained by solving the seismological inverse problem (Figure 6.2).

Note that for both orders of rotations about x_1 and x_2, we considered rotations about the vertical axis x_3 last; since we subsequently averaged about x_3, we did not need to investigate rotations about this axis. If x_3 would be one of the first two rotation axes, the elasticity tensor would be different. However, the total number of permutations for a 3-D rotation vector is six, and more importantly, the number of configurations per mineral would be 10^3, resulting in a total of $6 \cdot (10^3)^2$ options for each thermochemical model. This was computationally infeasible and we limited ourselves to rotations about two of the three axes. For each of the 22 491 thermochemical models considered, this gave 10^4 configurations for each order of rotation, as noted above.

Figure B.1 shows constraints on the rotation angles of the orthorhombic perovskite and cubic ferropericlase crystals for all the accepted thermochemical models and an order of rotation $x_1 - x_2(-x_3)$. For this order of rotation, no models were accepted in the second, third and fourth lower mantle layer (cf. Figure 6.4). For ease of comparison, we also visualised the accepted orientations using the more conventional Bunge Euler angles (Bunge, 1982), which can be easily derived from any rotation matrix. Since we only varied the rotation angle about the two horizontal axes (x_1 and x_2), we naturally did not consider all combinations of the three Euler angles. Figures B.2 and B.3 show constraints on the rotation, as represented by triplets of Bunge Euler angles, for both minerals and an order of rotation $x_2 - x_1(-x_3)$, corresponding to the constraints shown in Figure 6.4.

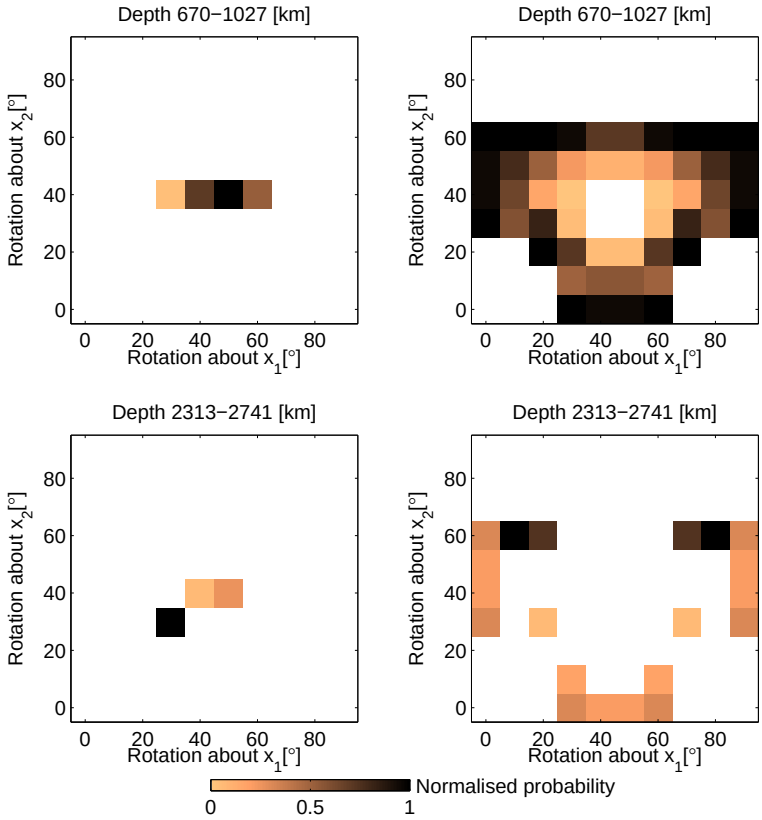


Figure B.1: Constraints on the orientation of the perovskite and ferropericlase crystals, represented by 2-D histograms of rotation angles for all accepted thermochemical models in the five lower mantle layers in D′. In each panel, the rotation angle about two principal (horizontal) axes x_1 and x_2 is shown for the orthorhombic perovskite (left) and cubic ferropericlase (right). The order of rotation was $x_1 - x_2(-x_3)$, where x_3 represents the vertical axis, over which we averaged to impose radial anisotropy. No models were accepted for the second, third and fourth layer for this order of rotation (cf. Figure 6.4). Each of the ten 2-D histograms is normalised, so that the colour indicates the relative number of accepted rotations. Empty cells represent rotation angles for which no thermochemical model fits all six seismic parameters within their uncertainties. The resemblance of ferropericlase in the bottom mantle layer to a piglet or monkey is considered coincidental and may not be related to the physics of the problem.

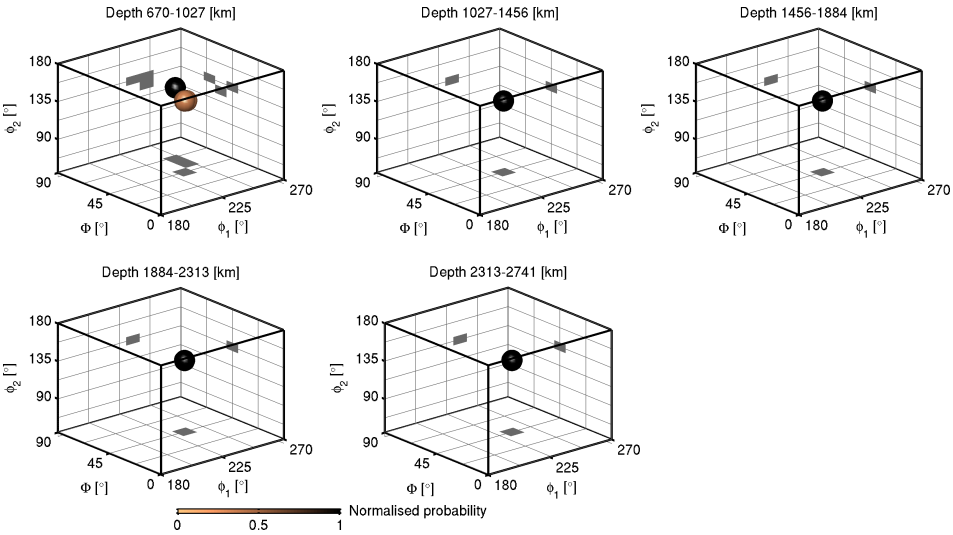


Figure B.2: Constraints on the orientation of perovskite crystals in the five lower mantle layers. Orientations are represented by Bunge Eugler angles (ϕ_1, Φ, ϕ_2, Bunge (1982)). The rotation angles for all accepted thermochemical models are shown as coloured spheres (voxels) and projected in grey, with the colour indicating the relative number of accepted rotations for each layer. Note that we performed the rotations about the two horizontal principal axes x_1 and x_2, and subsequently averaged over the vertical axis (x_3) to impose radial anisotropy (Figure 6.4). The more conventional Bunge Euler angles are only used for visualisation. The five boxes represent the five lower mantle layers, with the depth range given above each box.

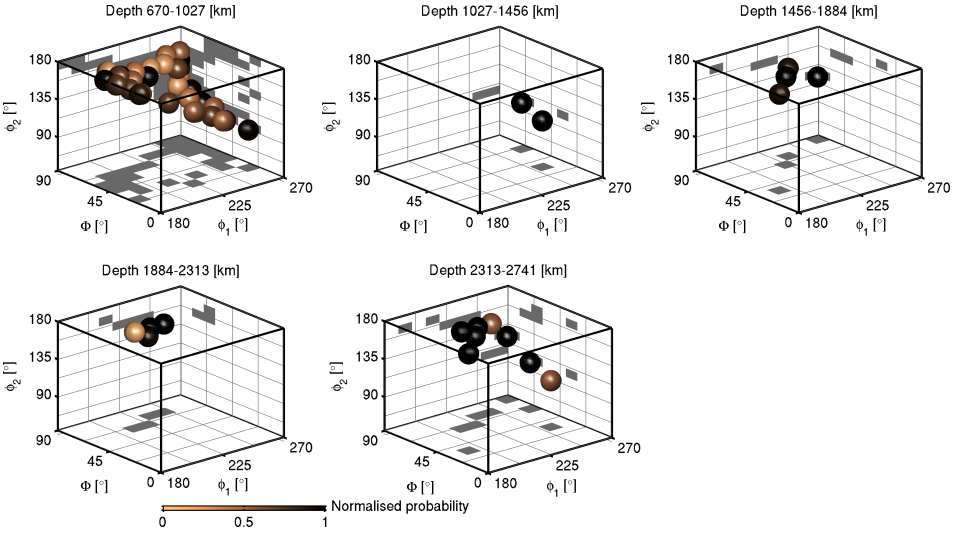


Figure B.3: Constraints on the orientation of ferropericlase crystals. See Figure B.2 for a description.

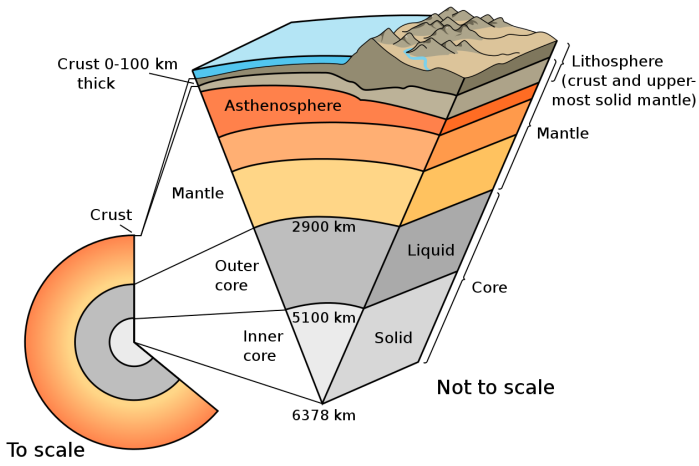
Samenvatting (Summary in Dutch)

Achtergrond

De Aarde heeft een straal van 6371 kilometer en een omtrek van ongeveer 40 000 kilometer. Sinds mensenheugenis reizen we over haar oppervlak, met als belangrijke mijlpaal de eerste complete reis om de wereld door de bemanning van Ferdinand Magellaan in 1522. Waar het navigeren van het aardoppervlak al eeuwenlang mogelijk is, is het nog steeds een grote uitdaging om het binnenste van de Aarde te bereiken. De Russen kwamen het verst in een wedloop met de Amerikanen, parallel aan de ruimterace die de twee naties uitvoerden tijdens de Koude Oorlog. Tussen 1970 en 1994 werkten zij aan de Kola Superdeep Borehole, die tot een diepte van 12,2 kilometer reikt. Maar dit is nog geen twee promille van de straal van de Aarde!

Onze kennis van het binnenste van de Aarde komt dan ook voornamelijk uit indirecte observaties. Hierbij kan men denken aan onder andere geologische observaties, metingen van zwaartekracht en het aardmagneetveld, en seismologische waarnemingen, zoals seismogrammen. Een seismogram is een registratie van een grondbeweging op de locatie van een seismometer, als gevolg van een aardbeving, aardverschuiving of explosie elders op Aarde. De energie die bij een aardbeving vrijkomt reist in de vorm van seismische golven door het binnenste en langs het oppervlak van de Aarde. Onderweg passeert de golf verschillende structuren en materialen. Omdat deze materialen verschillende eigenschappen hebben, zullen ze de seismische golven op verschillende manieren beïnvloeden. Deze informatie over de verschillende structuren en materialen zit dan ook verstopt in het signaal dat we uiteindelijk met een seismometer registreren.

Aan seismologen is het vervolgens de taak om dit signaal uit elkaar te pluizen en te vertalen naar hoe de Aarde er van binnen uitziet. Naar het determineren van een object, in dit geval de Aarde, op grond van indirecte observaties wordt vaak gerefereerd als het oplossen van een 'inverse' probleem. Dit in contrast met het voorwaartse ('forward') probleem, waarin bijvoorbeeld synthetische seismische data gesimuleerd worden. Conceptueel is het seismische inverse probleem vergelijkbaar met een CT-scan, waarbij röntgenstralen worden gebruikt om een menselijk lichaam "in kaart te brengen". Ter illustratie, stel dat we de route, of het pad, kennen van een seismische golf, reizend van een aardbeving in Japan naar een seismometer in Papenveer,



Figuur S.1: Een schematische weergave van het binnenste van de Aarde. Gereproduceerd met toestemming van het U.S. Geological Survey.

alsmede de reistijd (het verschil tussen de aankomsttijd in Papenveer en het moment van de aardbeving). In dat geval kunnen we iets zeggen over de snelheid van de golf en daarmee over de materialen waar de golf doorheen is gereisd. Zo is de geluidssnelheid in lucht 343 m/s, maar is deze veel hoger in bijvoorbeeld water (∼1500 m/s), ijzer (∼5000 m/s) en diamant (∼12 000 m/s).

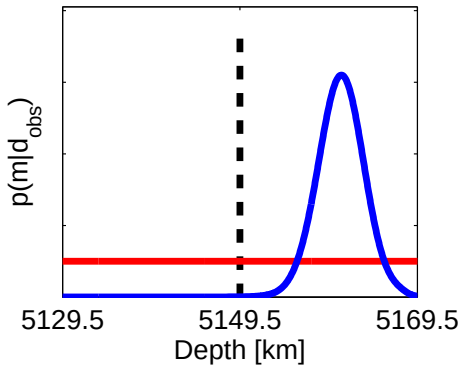
Sinds het begin van de vorige eeuw hebben seismologen de structuur van de Aarde in toenemend detail kunnen beschrijven (Figuur S.1). De Engelse geoloog en seismoloog Richard Oldham (1906) stelde dat de Aarde een gesmolten kern moest hebben. Dit concludeerde hij op basis van observaties, of beter gezegd het gebrek daaraan, van zogenaamde (transversale) S-golven. Dit type golf kan zich namelijk niet voortbewegen in vloeibare materialen, in tegenstelling tot (longitudinale) P-golven. De Deense seismologe Inge Lehmann (1936) verfijnde het beeld van de kern: de Aarde bestaat uit een vaste binnenkern met daaromheen een vloeibare buitenkern. Convectie in de vloeibare buitenkern genereert het aardmagneetveld, vergelijkbaar met de manier waarop een dynamo werkt. Het magneetveld beschermt ons tegen schadelijke zonnewinden en is één van de redenen dat leven op Aarde mogelijk is. Andere belangrijke seismologische observaties zijn die van de Kroatische geofysicus Andrija Mohorovičić (1910, 1992), die de overgang tussen de aardkorst en de mantel identificeerde, en de Duitser Beno Gutenberg (1914), die de diepte van de transitie van kern naar mantel bepaalde. De kern-mantelgrens ligt op 2891 km diepte, de grens tussen binnen- en buitenkern op ongeveer 5150 km. De kern bestaat voornamelijk uit ijzer en nikkel; de temperatuur loopt op van ongeveer 4000 graden Celsius aan de buitenkant van de buitenkern tot ongeveer 5000–6000 graden Celsius in de binnenkern. De mantel bestaat voornamelijk uit silicaten, zoals de mineralen olivijn en pyroxeen.

Volgend op deze ontdekkingen is het detail in (driedimensionale) seismologische aardmodellen door de jaren heen alleen maar toegenomen, een proces dat nu nog gaande is. Dit wordt mogelijk gemaakt door ontwikkelingen op een aantal vlakken. Ten eerste neemt de hoeveelheid data almaar toe; jaarlijks worden er op duizenden seismometers wereldwijd duizenden aardbevingen geregistreerd. De meeste van deze aardbevingen zijn gelukkig onschadelijk, maar kunnen wel nieuwe informatie opleveren over de Aarde. Deze seismische data zijn publiekelijk beschikbaar en het verwerken van deze enorme datavolumes is een uitdaging op zichzelf. Een tweede belangrijke ontwikkeling is de toenemende rekenkracht van computers. Tegenwoordig kan men op een willekeurige desktopcomputer in een aantal minuten of uren het golfveld in een gecompliceerd 3-D medium simuleren, waar dit vroeger dagen of weken kon duren, als het al mogelijk was. Tot slot hebben theoretische en methodologische ontwikkelingen bijgedragen aan het creëren van aardmodellen, middels een beter begrip van de natuurkunde achter golfvoortplanting en het bedenken van nieuwe algoritmes om de relevante informatie uit de beschikbare data te halen.

Het is cruciaal om altijd de kwaliteit van een model aan de kaak te stellen, hoe gedetailleerd (en vaak kleurrijk) dit model ook moge zijn. Seismische aardmodellen worden namelijk in hun nauwkeurigheid door verschillende factoren belemmerd. Ten eerste is de data vaak onregelmatig verdeeld over het aardoppervlak; seismometers staan voornamelijk op continenten, en aardbevingen vinden voornamelijk plaats langs plaatgrenzen. Door deze onregelmatige verdeling worden sommige delen van de Aarde niet voldoende 'belicht' om de structuur nauwkeurig te kunnen bepalen. Bovendien zitten er grote verschillen in de meetnauwkeurigheid van seismische datasets. Ten tweede worden vaak aannames gemaakt om de complexe (niet-lineaire) relatie tussen seismische data en het aardmodel te simplificeren (te lineariseren) en zodoende het mogelijk te maken een aardmodel te construeren binnen de beschikbare tijd en rekenkracht. Vanzelfsprekend hangt de kwaliteit van het aardmodel samen met de betrouwbaarheid van dit soort aannames.

Meestal wordt als resultaat van een onderzoek één specifiek aardmodel gepresenteerd dat de data het best verklaart. Maar gegeven de onzekerheden in de diverse elementen van het proces (data, gemaakte aannames, etc.), rijst altijd de vraag of er nog meer modellen zijn die de data even goed kunnen verklaren. Vaak is dit zo bij seismologische modellen; er kunnen wel duizenden modellen mogelijk zijn, gegeven de beschikbare data. Als alle mogelijke modellen in dit 'ensemble' grote gelijkenissen vertonen, dan kunnen we hieruit afleiden dat de structuren in de modellen 'robuust', of betrouwbaar, zijn. Zijn er daarentegen grote verschillen zichtbaar tussen de modellen, dan bevatten de data blijkbaar niet genoeg informatie over ons studie-object. Deze ambiguïteit ("non-uniqueness"), wordt idealiter meegenomen bij het bepalen van de onzekerheden in aardmodellen. Dit is makkelijker gezegd dan gedaan; hele vakgebieden richten zich op methodes om deze analyses zo goed mogelijk te kunnen uitvoeren. Het kwantificeren van de onzekerheden in seismologische modellen is één van de voornaamste drijfveren van dit proefschrift.

De onzekerheden in een aardmodel kunnen op verschillende manieren bepaald worden. Een eerste klasse van methoden maakt gebruik van lineaire aannames, net



Figuur S.2: 1-D waarschijnlijkheidsverdelingen voor een parameter m. In eerste instantie (*a priori*), dus voordat we een meting hebben gemaakt, weten we dat m op een diepte tussen 5129,5 en 5169,5 km ligt. We hebben geen verdere kennis van m en dus nemen we aan dat de kans dat m tussen deze twee dieptes ligt overal even groot is (de horizontale rode lijn geeft aan alle dieptes dezelfde waarschijnlijkheid). Vervolgens maken we een observatie d_{obs}, die nieuwe informatie bevat over de diepte van m. Onze nieuwe (*a posteriori*) kennis van m wordt gereflecteerd door de blauwe lijn. Zo hebben we bijvoorbeeld geleerd dat m niet kleiner kan zijn dan ongeveer 5149,5 km (de kans is 0%), terwijl de kans dat m op ongeveer 5160 km diepte ligt is toegenomen (de blauwe lijn ligt hoger dan de rode lijn rond deze diepte).

als hierboven, om een eerste-orde idee te krijgen van de grootte van de onzekerheden. Met dit soort methoden kunnen de onzekerheden niet volledig worden gekwantificeerd. Bovendien zijn deze methoden ook afhankelijk van de validiteit van de *a priori* gemaakte aannames. Een alternatief is de klasse van probabilistische, of Bayesiaanse, technieken. In de Bayesiaanse filosofie wordt onze kennis over alle modelparameters beschreven door waarschijnlijkheids- of kansverdelingen. Een kansverdeling reflecteert de onzekerheid in een parameter (en dus een onzekerheid in onze kennis!) op een heel natuurlijke manier. Alle waardes waarvan we weten dat ze onmogelijk zijn, op grond van onze bestaande (*a priori*) inzichten en de geobserveerde data, krijgen een kans van 0% toebedeeld. Alle overige waardes kunnen de data in meer of mindere mate verklaren en krijgen een grotere kans toegedicht. Figuur S.2 illustreert dit Bayesiaanse concept voor een modelparameter m.

De meest gangbare Bayesiaanse methoden maken gebruik van zogenaamde Monte Carlo-technieken, vernoemd naar het gelijknamige casino in Monaco. Dit type algoritme wordt gebruikt om een proces vele malen te simuleren, iedere keer met verschillende (willekeurig gekozen) startcondities, analoog aan de willekeur en kansen bij de spellen in een casino. In de context van de seismologie komt dit neer op het oplossen van de golfvergelijking, i.e. het genereren van synthetische seismische data, voor verschillende aardmodellen en aardbevingen. Deze synthetische seismische data kunnen vergeleken worden met de geobserveerde data. Hoe kleiner de verschillen tussen de

data, hoe groter de kans dat het onderliggende aardmodel representatief is voor de echte Aarde, en vice versa. Door dit proces duizenden malen te herhalen voor verschillende aardmodellen, kan een kansverdeling gevormd worden. Dit geeft ons een idee van de mogelijke structuur van de Aarde en de onzekerheden daarin.

Maar waarom wordt dit soort methodes dan niet altijd gebruikt? Welnu, voor niets gaat de zon op, en ook de Bayesiaanse filosofie kent wat obstakels. Ten eerste maakt deze methode een zeer intensief gebruik van computers. Het is, ook met de huidig beschikbare rekenkracht, lastig of zelfs onmogelijk om de waarschijnlijkheidsverdeling te bepalen voor een volledig aardmodel. De voornaamste reden hiervoor is de exponentiële toename van het aantal mogelijke aardmodellen bij een toename van het aantal dimensies in het model. Bijvoorbeeld, stel dat we een 1-D model hebben in de vorm van een enkele munt, met twee opties: kop of munt. Bij twee munten wordt het aantal opties $2^2 = 4$, bij drie munten $2^3 = 8$, etc. Bij 100 dimensies is het aantal opties enorm, $2^{100} \approx 1.27 \cdot 10^{30}$. Recente aardmodellen bestaan doorgaans uit honderden of duizenden parameters. De 1-D snelheid- en dichtheidsprofielen in Hoofdstuk 4 tot en met 7 zijn bijvoorbeeld opgebouwd uit de snelheid en dichtheid op 185 dieptes, en bevatten dus 185 parameters elk. Gedetailleerde driedimensionale modellen van de Aarde zijn opgedeeld in blokjes (stukken korst, mantel en/of kern) en bevatten soms zelf wel honderdduizenden parameters. Het is derhalve niet mogelijk om voor alle mogelijke aardmodellen synthetische data te genereren. Om dit probleem te omzeilen worden Monte Carlo-algoritmes ontwikkeld die de modelruimte, de ruimte gevormd door alle mogelijke aardmodellen, op een 'slimme' manier doorzoeken. Het idee is om de beschikbare rekenkracht vooral te gebruiken voor modellen die relatief waarschijnlijk zijn, op grond van de geobserveerde data. Ten tweede is het de vraag hoe je onzekerheden visualiseert in zo een hoog-dimensionale ruimte. Het 1-D voorbeeld in Figuur S.2 is (hopelijk) inzichtelijk. Een 2-D voorbeeld is ook goed te begrijpen, maar 3-D wordt ook al een stuk lastiger, zeker op papier. En 100-D? Of meer?

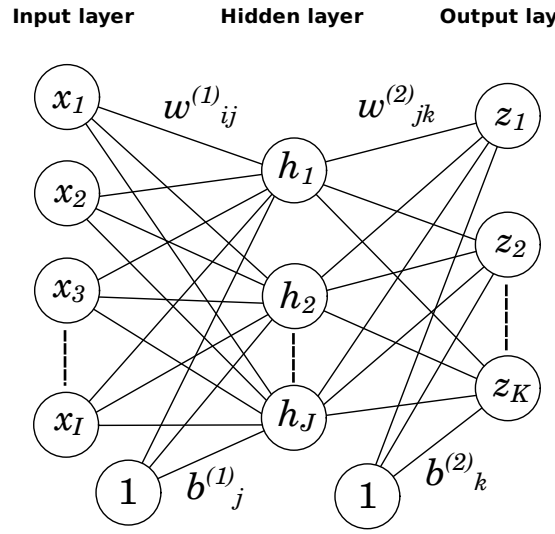
Neurale netwerken

In dit proefschrift heb ik voornamelijk gewerkt aan een alternatieve methode om het Bayesiaanse probleem op te lossen en om het eerste van de twee bovenstaande moeilijkheden te adresseren. Deze methode maakt gebruik van zogenaamde zelf-lerende algoritmes, of "machine learning" in het Engels. Zoals de naam aangeeft, gaat het om een computerprogramma, of machine, die het menselijk leerproces probeert na te bootsen. Dat wil zeggen, in plaats van volledig afhankelijk te zijn van een set voorgeprogrammeerde instructies, is het de bedoeling dat het programma leert van eerdere ervaringen en op grond daarvan beslissingen maakt. In ons dagelijks leven zijn we, zonder het wellicht te merken, omringd door dit soort patroonherkenningstechnieken. Een smartphone zit er vol mee, computers proberen patronen te herkennen in het online-gedrag van consumenten en zelfrijdende auto's proberen lering te trekken uit het gedrag van verschillende objecten op de weg in verschillende situaties.

Er zijn verschillende vormen van leren; in hoofdlijnen zijn deze onder te verdelen in gecontroleerd ("supervised"), ongecontroleerd ("unsupervised") en "reinforcement" leren. In de eerste categorie is er sprake van een 'leraar' die voorbeelden aanlevert, op grond waarvan de leerling (het programma) de relevante patronen moet leren herkennen. In de tweede categorie zijn er in eerste instantie geen voorbeelden voorhanden en wordt het leerproces gedomineerd door eerdere ervaringen. Zo zal een mens, na eenmaal teveel te hebben gedronken, dit in het vervolg van zijn of haar leven wel laten; hij of zij heeft geleerd dat dit simpelweg teveel hoofdpijn oplevert de dag erna. Of? Wellicht illustreert dit voorbeeld vooral hoe slecht sommige mensen leren van hun fouten, maar dat geheel terzijde. De laatste categorie, "reinforcement" leren, is te vergelijken met het opvoeden van een huisdier, waarbij een goede of slechte daad wordt beloond respectievelijk wordt afgestraft. In dit proefschrift vinden alle leerprocessen plaats middels de eerste vorm van gecontroleerd leren.

Er is een grote collectie van verschillende machine learning-technieken; ik maak gebruik van één specifieke categorie, te weten artificiële neurale netwerken. Neurale netwerken zijn, zoals de naam al aangeeft, een netwerk van neuronen en werden oorspronkelijk (vanaf de jaren vijftig van de vorige eeuw) dan ook ontwikkeld om het menselijk brein na te bootsen. Al snel bleek dat deze netwerken bij lange na niet in staat waren de indrukwekkende complexiteit van ons cognitief vermogen te simuleren. Neurale netwerken zijn echter wel zeer geschikt voor patroonherkenning; zo zijn ze zeer flexibel en dus aan te passen aan veel verschillende problemen en kunnen ze efficiënt interpoleren tussen de beschikbare data. In dit proefschrift maak ik gebruik van relatief simpele neurale netwerken (Figuur S.3).

Het idee is als volgt. Net als bij Monte Carlo-methodes worden voor duizenden verschillende aardmodellen synthetische seismische data berekend via computersimulaties. Het neurale netwerk krijgt het eerste synthetische seismische voorbeeld als input te zien en maakt een voorspelling voor de waarde van het bijbehorende aardmodel. Deze voorspelling kan worden vergeleken met de werkelijke waarde van dit aardmodel. In eerste instantie zal deze voorspelling inaccuraat blijken; op grond van het verschil worden de verbindingen (de gewichten w) in het neurale netwerk aangepast. Het netwerk krijgt het tweede voorbeeld te zien, en zo verder. Door de waardes w van de verbindingen aan te passen tijdens het analyseren van de voorbeelden kan het neurale netwerk een relatie of proces 'leren'. Denk als voorbeeld aan de relatie tussen afstand, reistijd en snelheid van een rijdende auto, waarbij we op grond van gemeten reistijden en afstanden de snelheid van een auto willen bepalen. De synthetische dataset (de zogenaamde 'trainingsdata') bevat voorbeelden van lange reistijden met bijbehorende lage snelheden en vice versa. Op grond van deze voorbeelden zou het netwerk moeten leren dat, bij een gelijkblijvende afstand, een lange reistijd correspondeert met een lage snelheid. Het trainingsproces blijft herhaaldelijk de voorbeelden evalueren totdat een (vooraf bepaalde) mate van nauwkeurigheid is bereikt. Voor de resultaten in dit proefschrift lag de trainingstijd van de neurale netwerken in de orde van grootte van uren tot dagen. Als laatste onderdeel van het leerproces wordt de nauwkeurigheid van het getrainde netwerk getest met synthetische voorbeelden die niet gebruikt zijn voor training. Indien het netwerk goed



Figuur S.3: Een voorbeeld van een neuraal netwerk, zoals gebruikt in dit proefschrift. De lijnen representeren de verbindingen tussen de neuronen in het netwerk (de gewichten w en b). De input neuronen x zijn verbonden met zogenaamde 'verborgen' neuronen h, die de input vormen voor de output z. Informatie vloeit alleen van input naar output (*feed-forward*). Door de waardes van de gewichten aan te passen tijdens het analyseren van voorbeelden van input en output kan het neurale netwerk een relatie of proces 'leren'.

presteert op de testvoorbeelden, kan het worden toegepast op de echte geobserveerde seismische data. Als output levert het netwerk dan een voorspelling van de structuur van de echte Aarde. Zodoende kan het seismische inverse probleem opgelost worden. Zoals eerder beschreven, is bij een Bayesiaanse aanpak de gewenste output niet een enkel aardmodel, maar een kansverdeling die de waarschijnlijkheid van verschillende aardmodellen reflecteert. Om dit te bereiken gebruik ik een extensie op het simpele neurale netwerk getoond in Figuur S.3, een zogenaamd Mixture Density Network (Figuur 2.3 in de hoofdtekst); dit type netwerk geeft als output een kansverdeling voor modelparameters.

Het bovenstaande vat in een aantal pagina's de achtergrond van zowel het seismologische inverse (of 'imaging') probleem als de gebruikte methodiek samen. Korter gezegd, het doel van dit proefschrift is om seismische inverse problemen voor de eendimensionale Aarde op te lossen en daarbij de onzekerheden te kwantificeren met behulp van neurale netwerken, als alternatief voor meer gangbare Monte Carlo-methodes. Een uitgebreidere introductie tot de probleemstelling en een behandeling van de gebruikte methodes worden gegeven in Hoofdstukken 1 en 2.

Belangrijkste bevindingen

Het restant van deze samenvatting behandelt de resultaten en de conclusies zoals beschreven in dit proefschrift. De analyses beperken zich tot de radiale (1-D) structuur van de Aarde, dat wil zeggen hoe de structuur van de Aarde verandert met diepte. Dit is gemotiveerd door ten eerste de noodzaak om de neuralenetwerktechniek te testen en ten tweede het simpele feit dat ook deze methode gelimiteerd wordt door de beschikbare hoeveelheid rekenkracht. In Hoofdstukken 3 tot 7 worden verschillende types seismische data geanalyseerd om verschillende aspecten van de 1-D structuur van de Aarde te belichten. Hoofdstuk 3 gebruikt P-golf reis- of looptijden om de P-golfsnelheid in de Aarde te bestuderen. Daarbij beantwoorden we de vraag of en in welke mate dit type data deze golfsnelheden kan bepalen. De voornaamste bevinding is dat deze teleseismische dataset grip heeft op de P-golf structuur in de kern en de ondermantel (dieper dan 660 km), maar niet op die in de korst en de bovenmantel; evenmin kan de data de diepte van seismische discontinuïteiten bepalen.

In Hoofdstuk 4 vergelijken we deze looptijdmetingen met een ander type data, frequenties van de *eigentrillingen* van de Aarde. Ieder eindig object kan trillen op een bepaalde eigenfrequentie, nadat het uit een bepaalde evenwichtssituatie is gehaald, een concept dat resonantie wordt genoemd. Een bekend voorbeeld hiervan is een stemvork. De frequenties van de eigentrillingen hangen af van de eigenschappen van het object. De eigentrillingen van de Aarde kunnen worden geëxciteerd door krachtige aardbevingen. Door de bijbehorende frequenties te meten, kunnen we iets zeggen over de structuur van de Aarde. Onze methode maakt het mogelijk om de informatie in verschillende datasets te kwantificeren. De vergelijking in Hoofdstuk 4 toont aan dat de eigentrillingsdata veel meer informatie bevat dan de P-golf looptijden, die minder nauwkeurig gemeten zijn. We hebben ons vooral gefocust op de diepte van seismische discontinuïteiten en de grootte van de snelheids- en dichtheidscontrasten bij deze overgangen. Een belangrijk resultaat dat volgt uit de analyse van de eigentrillingen is dat de data een andere diepte dan normaliter aangenomen prefereren voor de grens tussen de binnen- en buitenkern (ICB, Tabel 4.2). Daarnaast laat een analyse van attenuatie (dissipatie van energie van seismische golven) een sterk contrast zien tussen de onderste en bovenste helft van de binnenkern, hetgeen eerdere hypotheses over het bestaan van een 'binnenste binnenkern' ondersteunt.

In deze eerste hoofdstukken is aangenomen dat de Aarde isotroop is. Dit betekent dat de eigenschappen van het medium onafhankelijk zijn van de richting waarin een seismische golf door het medium reist. Dit is echter niet altijd correct; delen van de Aarde zijn anisotroop. Als gevolg hiervan reizen golven bijvoorbeeld sneller in een horizontale dan in een vertical richting. Een voorname reden hiervoor is de preferentiële uitlijning van kristallen in anisotrope mineralen, zoals olivijn, perovskiet of periklaas. In dit soort mineralen reizen seismische golven met verschillende snelheden langs de verschillende assen van een kristal. Deze uitlijning van kristallen kan het gevolg zijn van grootschalige deformatieprocessen in de Aarde. Derhalve kunnen we deze deformaties mogelijk beter begrijpen als we de seismische anisotropie in de

Aarde in kaart brengen. In Hoofdstuk 4 breiden we onze modelparametrisatie uit met anisotropie om te testen in hoeverre de aanname van isotropie invloed heeft op onze resultaten. De belangrijkste bevindingen blijven staan, maar we vinden ook hele duidelijke wisselwerkingen tussen verschillende modelparameters. Zo kan de eigentrillingsdata niet alle parameters tegelijk accuraat bepalen in regio's waarvan bekend is dat ze anisotroop zijn, zoals de diepe mantel (de D"-regio).

Om deze anisotrope structuren beter te kunnen bepalen voegen we in Hoofdstuk 6 een tweede dataset met eigentrillingen toe. Deze dataset is complementair aan de eerste door een verschillende gevoeligheid voor het binnenste van de Aarde. De gezamenlijke dataset bevat meer informatie over de anisotrope structuur in de mantel. De belangrijkste bevindingen zijn de vondst van anisotropie in de ondermantel, waarvan tot nu toe gedacht wordt dat deze isotroop is, en een hogere dichtheid dan in bestaande 1-D modellen op dezelfde dieptes. Geodynamische simulaties nemen aan dat dit deel van de Aarde isotroop is; een mogelijk anisotrope structuur zou een sterke invloed kunnen hebben op de nauwkeurigheid van deze modellen. We hebben geverifieerd dat onze observaties van seismische anisotropie te verklaren zijn door een simpel model voor de temperatuur en compositie van de ondermantel (tussen 660 km en $\sim$2741 km diepte). In de toekomst moeten meer geavanceerde (3-D) seismologische en mineralogische modellen gebruikt worden om deze observaties beter te verklaren. Hoofdstuk 7 behandelt de anisotrope structuur in de binnenkern. We vinden anisotropie in de bovenste $\sim$100 km van de binnenkern, consistent met andere studies. Daarnaast observeren we sterke anisotropie in de diepe binnenkern; dit is in lijn met eerdere studies, maar de studies verschillen in de stijl van parametrisatie van de anisotropie. Toekomstig werk moet de definitieve structuur van deze anisotropie bepalen, bijvoorbeeld door gebruik te maken van complementaire P-golf looptijdgegevens. Net als voor de mantel, is het belangrijk om de precieze anisotrope structuur te kennen om betrouwbare geodynamische modellen van de kern te kunnen maken.

Hoofdstuk 8 draait niet om het binnenste van de Aarde, maar richt zich op een meer fundamenteel probleem met de data. Een grote uitdaging in dit tijdperk van 'big data', i.e. ogenschijnlijk eindeloze datavolumes, is het daadwerkelijk analyseren van al deze data. Hiervoor zijn naast meer rekenkracht (grotere computers) ook betere technieken nodig. Eén mogelijkheid is het comprimeren van data: kunnen we het aantal dimensies van de data reduceren, zonder dat hierbij de informatie in de data verloren gaat? Seismogrammen bevatten in theorie alle beschikbare informatie over de Aarde; looptijden zijn bijvoorbeeld maar een klein onderdeel van het gehele seismogram. Idealiter gebruiken we dus het gehele seismogram in onze inversies, maar door hun hoge dimensie (iedere tijdsstap is één dimensie) is dit vaak niet mogelijk. In Hoofdstuk 8 reduceren we de dimensie van seismogrammen met behulp van een andere neuralenetwerktechniek, een zogenaamd 'auto-encoder'-netwerk. Vervolgens gebruiken we deze encoderingen om, net als in eerdere hoofdstukken, voorspellingen te maken voor de structuur van de Aarde. Deze studie is bedoeld als een eerste aanzet tot seismische inversies die gebruik maken van zulke auto-encoders. We richten ons dan ook op het vergelijken van de resultaten voor de originele en de gecomprimeerde

seismogrammen. De resultaten zijn vergelijkbaar; dit is bemoedigend en de methode kan dan ook in de toekomst worden uitgebreid naar een grotere en meer complete opzet.

Tot slot vat Hoofdstuk 9, naast de resultaten voor de 1-D structuur van de Aarde, de belangrijkste bevindingen over de gebruikte methode samen. De in dit proefschrift ontwikkelde neuralenetwerktechniek is flexibel en kan efficiënt omgaan met beschikbare rekenkracht, omdat meerdere modelparameters geanalyseerd kunnen worden met dezelfde synthetische dataset. Dit is een belangrijk verschil met Monte Carlo-methoden, die voor iedere parameter afzonderlijk synthetische voorbeelden moeten genereren en op deze manier meer rekenkracht nodig kunnen hebben. Daarnaast geeft de neuralenetwerktechniek altijd een conservatieve schatting van de onzekerheid in een modelparameter, i.e. de onzekerheid is altijd aan de voorzichtige (hoge) kant, ten opzichte van Monte Carlo-technieken. Het nadeel hiervan is dat de methode minder informatie uit de data haalt dan wellicht beschikbaar is, informatie die Monte Carlo-technieken eventueel wel kunnen achterhalen. Samenvattend kent de neuralenetwerktechniek zowel sterke als zwakke punten; idealiter worden de voordelen van neurale netwerken in de toekomst geëxploiteerd en de nadelen vermeden in een gecombineerde aanpak. Zo zou de beschikbare rekenkracht efficiënter gebruikt kunnen worden door met neurale netwerken een initiële conservatieve schatting van modelparameters te maken. Vervolgens kan deze schatting als startpunt dienen voor een Monte Carlo-algoritme voor verdere analyse van de informatie in de data.

Curriculum Vitae

Personal information

Date of Birth	August 25, 1986
Place of Birth	Leiderdorp, The Netherlands
Citizenship	Dutch

Experience

2015	**Intern**, Roland Berger Strategy Consultants, Amsterdam, The Netherlands.
2011–2015	**PhD researcher**, *Seismology*, Utrecht University, Utrecht, The Netherlands.
2011–2014	**PhD representative and board member**, PrOUt, PhD network of Utrecht University.
2012–2013	**Chairman**, PrOUt, PhD network of Utrecht University.
2011–2013	**PhD representative**, UGG, PhD network of Faculty of Geosciences, Utrecht University.
2008–2013	**Teaching assistant**, Utrecht University, Utrecht, The Netherlands.
2007–2008	**Data processing assistant**, TNO, Utrecht, The Netherlands.
2007	**Data processing assistant**, KNMI, Utrecht, The Netherlands.

Education

2008–2010	**Master of Science *(cum laude)***, *Geophysics*, Utrecht University, The Netherlands. Thesis: Towards quantifying uncertainties in travel-time tomography using the null space shuttle
Nov–Dec 2009	**Visiting scholar**, Massachusetts Institute of Technology, Cambridge, USA.
2004–2007	**Bachelor of Science**, *Earth Sciences*, Utrecht University, The Netherlands.
1998–2004	**Secondary School**, Stedelijk Gymnasium Leiden, Leiden, The Netherlands.

A curriculum vitae is outdated the moment it is printed on paper. To transcend this limitation and advance into the modern digital age, feel free to scan the QR-code below at any time in the future, which will (hopefully) lead you to my up-to-date LinkedIn-profile. De techniek staat voor niets!

List of publications

De Wit, R. W. L., J. Trampert and R. D. van der Hilst, 2012. Toward quantifying uncertainty in travel time tomography using the null-space shuttle. *Journal of Geophysical Research 117*, B03301.

De Wit, R. W. L., A. P. Valentine, and J. Trampert, 2013. Bayesian inference of Earth's radial seismic structure from body-wave traveltimes using neural networks. *Geophysical Journal International 195*, 408–422.

De Wit, R. W. L., P. J. Käufl, A. P. Valentine, and J. Trampert, 2014. Bayesian inversion of free oscillations for Earth's radial (an)elastic structure. *Physics of the Earth and Planetary Interiors 237*, 1–17.

De Wit, R. W. L. and J. Trampert, 2015. Robust constraints on average radial lower mantle anisotropy and consequences for composition and texture. *submitted to Earth and Planetary Science Letters*.

Käufl, P. J., A. P. Valentine, R. W. L. de Wit and J. Trampert, 2015. Robust and fast probabilistic source parameter estimation from near-field displacement waveforms using pattern recognition. *accepted for publication in the Bulletin of the Seismological Society of America*.

Käufl, P. J., A. P. Valentine, R. W. L. de Wit and J. Trampert, 2015. Solving probabilistic inverse problems rapidly with prior samples. *submitted to Geophysical Journal International*.

Bibliography

Agnew, D. C. (2002). History of seismology. In W. H. K. Lee, H. Kanamori, P. Jennings, and C. Kisslinger (Eds.), *IASPEI international handbook of earthquake engineering seismology*, pp. 3–13. Academic Press.

Aki, K., A. Christoffersson, and E. S. Husebye (1977). Determination of the three-dimensional seismic structure of the lithosphere. *J. Geophys. Res. 82*, 277–296.

Alboussière, T. and R. Deguen (2012). Asymmetric dynamics of the inner core and impact on the outer core. *Journal of Geodynamics 61*(0), 172 – 182.

Alpaydin, E. (2004). *Introduction to Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press.

An, M. (2012). A simple method for determining the spatial resolution of a general inverse problem. *Geophysical Journal International 191*(2), 849–864.

Anderson, D. L. and R. S. Hart (1978). Q of the Earth. *J. Geophys. Res. 83*(B12), 5869–5882.

Andrews, J., A. Deuss, and J. Woodhouse (2006). Coupled normal-mode sensitivity to inner-core shear velocity and attenuation. *Geophysical Journal International 167*(1), 204–212.

Babuska, V. and M. Cara (1991). *Seismic anisotropy in the Earth*, Volume 10. Springer.

Backus, G. E. (1962). Long-wave elastic anisotropy produced by horizontal layering. *Journal of Geophysical Research 67*(11), 4427–4440.

Backus, G. E. and F. Gilbert (1968). The Resolving Power of Gross Earth Data. *Geophys. J. R. Astron. Soc. 16*, 169–205.

Backus, G. E. and F. Gilbert (1970). Uniqueness in the Inversion of Inaccurate Gross Earth Data. *Phil. Trans. R. Soc. Lond. 266*, 123–192.

Baker, J. A., P. J. Kornguth, J. Y. Lo, M. E. Williford, and C. E. Floyd Jr. (1995). Breast cancer: prediction with artificial neural network based on BI-RADS standardized lexicon. *Radiology 196*, 817–822.

Bayes, T. (1763). An Essay Towards Solving a Problem in the Doctrine of Chances. By the late Rev. Mr. Bayes, communicated by Mr. Price, in a letter to John Canton, M. A. and F. R. S. *Philosophical Transactions of the Royal Society of London 53*, 370–418.

Beghein, C. and J. Trampert (2003). Robust Normal Mode Constraints on Inner-Core Anisotropy from Model Space Search. *Science 299*(5606), 552–555.

Beghein, C., J. Trampert, and H. J. van Heijst (2006). Radial anisotropy in seismic reference models of the mantle. *J. Geophys. Res. 111*, B02303.

Bellman, R. E. (1961). *Adaptive Control Processes*. New Jersey, USA: Princeton University Press.

Ben-Menahem, A. (1995). A concise history of mainstream seismology: Origins, legacy, and perspectives. *Bulletin of the Seismological Society of America 85*(4), 1202–1225.

Bengio, Y. (2009). Learning deep architectures for AI. *Foundations and trends® in Machine Learning 2*(1), 1–127.

Bergman, M. I. (1997). Measurements of electric anisotropy due to solidification texturing and the implications for the Earth's inner core. *Nature 389*(6646), 60–63.

Bernoulli, J. (1713). *Ars conjectandi*. Basel: Impensis Thurnisiorum.

Betancourt, M., S. Byrne, S. Livingstone, and M. Girolami (2014). The Geometric Foundations of Hamiltonian Monte Carlo. *arXiv preprint arXiv:1410.5110*.

Bishop, C. M. (1995). *Neural networks for pattern recognition*. New York: Oxford University Press.

Bishop, C. M. (1996). Neural Networks: A Pattern Recognition Perspective. In E. Fiesler and R. Beale (Eds.), *Handbook of Neural Computation*. Oxford University Press and IOP.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York: Springer.

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

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

Bodin, T., M. Sambridge, H. Tkalčić, P. Arroucau, K. Gallagher, and N. Rawlinson (2012). Transdimensional inversion of receiver functions and surface wave dispersion. *Journal of Geophysical Research: Solid Earth 117*(B2), n/a–n/a.

Bolt, B. A. (1977). The detection of PKIIKP and damping in the inner core. *Annals of Geophysics 30*(3-4).

Boschi, L. (2003). Measures of resolution in global body wave tomography. *Geophys. Res. Lett. 30*.

Boschi, L. and A. Dziewoński (1999). High and low resolution images of the Earth's mantle - Implications of different approaches to tomographic modeling. *J. Geophys. Res. 104*, 25,567–25,594.

Breiman, L. et al. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science 16*(3), 199–231.

Brown, J. M. and T. J. Shankland (1981). Thermodynamic parameters in the earth as determined from seismic profiles. *Geophysical Journal International 66*(3), 579–596.

Buffett, B. A. and H.-R. Wenk (2001). Texturing of the Earth's inner core by Maxwell stresses. *Nature 413*(6851), 60–63.

Bullen, K. E. (1949). Compressibility-pressure hypothesis and the earth's interior. *Geophysical Supplements to the Monthly Notices of the Royal Astronomical Society 5*(9), 335–368.

Bunge, H. J. (1982). *Texture analysis in materials science: mathematical methods*. Butterworth's London.

Cammarano, F., A. Deuss, S. Goes, and D. Giardini (2005). One-dimensional physical reference models for the upper mantle and transition zone: Combining seismic and mineral physics constraints. *Journal of Geophysical Research: Solid Earth 110*(B1).

Cammarano, F., S. Goes, A. Deuss, and D. Giardini (2005). Is a pyrolitic adiabatic mantle compatible with seismic data? *Earth Planet. Res. Lett. 232*, 227 – 243.

Cammarano, F. and B. Romanowicz (2008). Radial profiles of seismic attenuation in the upper mantle based on physical models. *Geophysical Journal International 175*(1), 116–134.

Cammarano, F., P. Tackley, and L. Boschi (2011). Seismic, petrological and geodynamical constraints on thermal and compositional structure of the upper mantle: global thermochemical models. *Geophys. J. Int. 187*, 1301–1318.

Candès, E. J. and M. B. Wakin (2008). An introduction to compressive sampling. *Signal Processing Magazine, IEEE 25*(2), 21–30.

Cao, A. and B. Romanowicz (2004). Constraints on density and shear velocity contrasts at the inner core boundary. *Geophys. J. Int. 157*(3), 1146–1151.

Cao, A. and B. Romanowicz (2007). Test of the innermost inner core models using broadband PKIKP travel time residuals. *Geophysical Research Letters 34*(8), n/a–n/a.

Chambat, F. and B. Valette (2001). Mean radius, mass and inertia for reference Earth's models. *Phys. Earth Planet. Inter. 124*, 237–253.

Chang, S.-J., A. M. Ferreira, J. Ritsema, H. J. van Heijst, and J. H. Woodhouse (2014). Global radially anisotropic mantle structure from multiple datasets: A review, current challenges, and outlook. *Tectonophysics 617*, 1 – 19.

Cobden, L., S. Goes, F. Cammarano, and J. A. D. Connolly (2008). Thermochemical interpretation of one-dimensional seismic reference models for the upper mantle: evidence for bias due to heterogeneity. *Geophysical Journal International 175*(2), 627–648.

Cobden, L., S. Goes, M. Ravenna, E. Styles, F. Cammarano, K. Gallagher, and J. A. D. Connolly (2009). Thermochemical interpretation of 1-D seismic data for the lower mantle: The significance of nonadiabatic thermal gradients and compositional heterogeneity. *Journal of Geophysical Research: Solid Earth 114*(B11).

Cormier, V. F. and A. Stroujkova (2005). Waveform search for the innermost inner core. *Earth and Planetary Science Letters 236*, 96 – 105.

Cornford, D., I. T. Nabney, and C. M. Bishop (1999). Neural Network-Based Wind Vector Retrieval from Satellite Scatterometer Data. *Neural Computing & Applications 8*(3), 206–217.

Creager, K. C. (1992). Anisotropy of the inner core from differential travel times of the phases PKP and PKIKP. *Nature 356*, 309–314.

Crotwell, H. P., T. J. Owens, and J. Ritsema (1999). The taup toolkit: Flexible seismic travel-time and ray-path utilities. *Seismol. Res. Lett. 70*, 154–160.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signalds and Systems 2*, 304–314.

Dahlen, F. and J. Tromp (1998). *Theoretical global seismology*. Princeton university press.

Dai, H. and C. MacBeth (1997). Application of back-propagation neural networks to identification of seismic arrival types. *Phys. Earth Planet. Inter. 101*, 177–188.

Dawson, M. R. W. (2013). *Mind, Body, World: Foundations of Cognitive Science*. Canada: AU Press.

de Wit, R. W. L., P. J. Käufl, A. P. Valentine, and J. Trampert (2014). Bayesian inversion of free oscillations for earth's radial (an)elastic structure. *Phys. Earth Planet. Inter.*, doi:10.1016/j.pepi.2014.09.004.

de Wit, R. W. L., J. Trampert, and R. D. van der Hilst (2012). Toward quantifying uncertainty in travel time tomography using the null-space shuttle. *J. Geophys. Res. 117*, B03301.

de Wit, R. W. L., A. P. Valentine, and J. Trampert (2013). Bayesian inference of earth's radial seismic structure from body-wave traveltimes using neural networks. *Geophys. J. Int. 195*, 408–422.

Deguen, R. (2012). Structure and dynamics of earth's inner core. *Earth and Planetary Science Letters 333–334*(0), 211 – 225.

Deguen, R., P. Cardin, S. Merkel, and R. A. Lebensohn (2011). Texturing in earth's inner core due to preferential growth in its equatorial belt. *Physics of the Earth and Planetary Interiors 188*(3), 173–184.

Delsanto, S., M. Griffa, and L. Morra (2006). Inverse Problems and Genetic Algorithms. In *Universality of Nonclassical Nonlinearity*, pp. 349–366. Springer.

Deschamps, F. and J. Trampert (2003). Mantle tomography and its relation to temperature and composition. *Phys. Earth Planet. Inter. 140*, 277–291.

Deschamps, F. and J. Trampert (2004). Towards a lower mantle reference temperature and composition. *Earth and Planetary Science Letters 222*(1), 161 – 175.

Deschamps, F., J. Trampert, and P. J. Tackley (2007). Thermo-Chemical Structure of the Lower Mantle: Seismological Evidence and Consequences for Geodynamics. In D. Yuen, S. Maruyama, S.-I. Karato, and B. Windley (Eds.), *Superplumes: Beyond Plate Tectonics*, pp. 293–320. Springer Netherlands.

Deuss, A. (2014). Heterogeneity and Anisotropy of Earth's Inner Core. *Annual Review of Earth and Planetary Sciences 42*(1), 103–126.

Deuss, A., J. Andrews, and E. Day (2013). Seismic Observations of Mantle Discontinuities and Their Mineralogical and Dynamical Interpretation. In *Physics and Chemistry of the Deep Earth*, pp. 295–323. John Wiley & Sons, Ltd.

Deuss, A., J. C. E. Irving, and J. H. Woodhouse (2010). Regional Variation of Inner Core Anisotropy from Seismic Normal Mode Observations. *Science 328*(5981), 1018–1020.

Deuss, A., J. Ritsema, and H. van Heijst (2013). A new catalogue of normal-mode splitting function measurements up to 10 mhz. *Geophys. J. Int. 193*(2), 920–937.

Donoho, D. L. (2006). Compressed sensing. *IEEE Transactions on Information Theory 52*(4), 1289–1306.

Duane, S., A. D. Kennedy, B. J. Pendleton, and D. Roweth (1987). Hybrid Monte Carlo. *Physics letters B 195*(2), 216–222.

Duda, R. O., P. E. Hart, and D. G. Stork (2001). *Pattern Classification*. New York, USA: Wiley.

Durek, J. J. and G. Ekström (1996). A radial model of anelasticity consistent with long-period surface-wave attenuation. *Bull. Seismol. Soc. Am. 86*, 144–158.

Dziewoński, A. M. and D. L. Anderson (1981). Preliminary reference Earth model. *Phys. Earth Planet. Inter. 25*, 297–356.

Dziewoński, A. M., T.-A. Chou, and J. H. Woodhouse (1981). Determination of earthquake source parameters from waveform data for studies of global and regional seismicity. *Journal of Geophysical Research: Solid Earth (1978–2012) 86*(B4), 2825–2852.

Dziewoński, A. M., A. L. Hales, and E. R. Lapwood (1975). Parametrically simple earth models consistent with geophysical data. *Phys. Earth Planet. Inter. 10*, 12–48.

Dziewoński, A. M. and J. H. Woodhouse. Studies of the seismic source using normal-mode theory. In *Earthquakes: Observations, theory and intepretation*.

Ekström, G., M. Nettles, and A. Dziewoński (2012). The global CMT project 2004–2010: centroid-moment tensors for 13,017 earthquakes. *Physics of the Earth and Planetary Interiors 200*, 1–9.

Ellenberger, F. (1999). *History of geology*, Volume 2. Taylor & Francis.

Engdahl, E. R., R. D. van der Hilst, and R. Buland (1998). Global teleseismic earthquake relocation with improved travel times and procedures for depth determination. *Bull. Seismol. Soc. Am. 88*, 722–743.

Fichtner, A. (2010). *Full seismic waveform modelling and inversion*. Heidelberg, Germany: Springer-Verlag.

Fichtner, A., B. L. Kennett, and J. Trampert (2013). Separating intrinsic and apparent anisotropy. *Physics of the Earth and Planetary Interiors 219*, 11–20.

Fichtner, A., B. L. N. Kennett, H. Igel, and H.-P. Bunge (2009). Full waveform tomography for upper-mantle structure in the Australasian region using adjoint methods. *Geophys. J. Int. 179*(3), 1703–1725.

Fienberg, S. E. (2006). When did Bayesian inference become "Bayesian"? *Bayesian analysis 1*(1), 1–40.

Forrester, A., A. Sobester, and A. Keane (2008). *Engineering design via surrogate modelling: a practical guide*. John Wiley & Sons.

Fu, L.-Y. (2001). Caianiello neural network method for geophysical inverse problems. In M. M. Poulton (Ed.), *Computational Neural Networks for Geophysical Data Processing*, pp. 187–215. Amsterdam, The Netherlands: Pergamon.

Gallagher, K. and M. Sambridge (1994). Genetic algorithms: a powerful tool for large-scale nonlinear optimization problems. *Computers & Geosciences 20*(7), 1229–1236.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (1995). *Bayesian Data Analysis*. London: Chapman & Hall.

Geman, S., E. Bienenstock, and R. Doursat (1992). Neural networks and the bias/variance dilemma. *Neural computation 4*(1), 1–58.

Gilbert, F., D. A. (1975). An application of normal mode theory to the retrieval of structural parameters and source mechanisms from seismic spectra. *Phil. Trans. R. Soc. Lond. 278*, 187–269.

Gubbins, D., G. Masters, and F. Nimmo (2008). A thermochemical boundary layer at the base of Earth's outer core and independent estimate of core heat flux. *Geophys. J. Int. 174*(3), 1007–1018.

Gutenberg, B. (1914). Über erdbebenwellen. vii a. beobachtungen an registrierungen von fernbeben in göttingen und folgerung über die konstitution des erdkörpers (mit tafel). *Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse 1914*, 125–176.

Hadamard, J. (1902). Sur les problèmes aux dérivées partielles et leur signification physique. *Princeton university bulletin 13*(49-52), 28.

Halley, E. (1692). An account of the causes of the change of the variation of the magnetical needle; with an hypothesis of the structure of the internal parts of the Earth. *Philosophical Transactions of the Royal Society of London 195*, 208–221.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika 57*(1), 97–109.

He, X. and J. Tromp (1996). Normal-mode constraints on the structure of the Earth. *Journal of Geophysical Research: Solid Earth (1978–2012) 101*(B9), 20053–20082.

Hebb, D. O. (1949). *The Organization of Behavior*. New York, USA: Wiley.

Herrmann, F. J., M. P. Friedlander, and O. Yilmaz (2012). Fighting the curse of dimensionality: compressive sensing in exploration seismology. *Signal Processing Magazine, IEEE 29*(3), 88–100.

Hinton, G. E. and R. R. Salakhutdinov (2006). Reducing the Dimensionality of Data with Neural Networks. *Science 313*, 504–507.

Hochreiter, S. and J. Schmidhuber (1997). Long Short-Term Memory. *Neural Computation 9*, 1735–1780.

Hornik, K., M. Stinchcombe, and H. White (1989). Multilayer Feedforward Networks are Universal Approximators. *Neural Networks 2*, 359–366.

Igel, C. and M. Hüsken (2003). Empirical evaluation of the improved Rprop learning algorithms. *Neurocomputing 50*, 105–123.

Irifune, T. (1994). Absence of an aluminous phase in the upper part of the earth's lower mantle. *Nature 370*, 131–133.

Irving, J. C. and A. Deuss (2011). Stratified anisotropic structure at the top of earth's inner core: A normal mode study. *Physics of the Earth and Planetary Interiors 186*(1–2), 59 – 69.

Irving, J. C. E., A. Deuss, and J. Andrews (2008). Wide-band coupling of Earth's normal modes due to anisotropic inner core structure. *Geophysical Journal International 174*(3), 919–929.

Irving, J. C. E., A. Deuss, and J. H. Woodhouse (2009). Normal mode coupling due to hemispherical anisotropic structure in Earth's inner core. *Geophysical Journal International 178*(2), 962–975.

ISC (2008). Summary of the Bulletin of the International Seismological Centre.

Ishii, M. and A. M. Dziewoński (2002). The innermost inner core of the earth: Evidence for a change in anisotropic behavior at the radius of about 300 km. *Proceedings of the National Academy of Sciences 99*(22), 14026–14030.

Jacobsen, S. D., H.-J. Reichmann, H. A. Spetzler, S. J. Mackwell, J. R. Smyth, R. J. Angel, and C. A. McCammon (2002). Structure and elasticity of single-crystal (mg,fe)o and a new method of generating shear waves for gigahertz ultrasonic interferometry. *Journal of Geophysical Research: Solid Earth 107*(B2), ECV 4–1–ECV 4–14.

Jain, A. K., R. P. W. Duin, and J. Mao (2000). Statistical pattern recognition: A review. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 22*(1), 4–37.

James, W. (1892). *Psychology: Briefer Course*. New York, USA: Henry Holt.

Jaynes, E. (2003). *Probability Theory: The Logic of Science*. Cambridge University Press.

Jeanloz, R. and H.-R. Wenk (1988). Convection and anisotropy of the inner core. *Geophysical Research Letters 15*(1), 72–75.

Jeffreys, H. (1939). *Theory of Probability* (1 ed.). Oxford: Clarendon Press.

Jeffreys, H. and K. E. Bullen (1940). *Seismological Tables*. London, UK: British Association for the Advancement of Science.

Jiang, X. and H. Adeli (2005). Dynamic Wavelet Neural Network Model for Traffic Flow Forecasting. *J. Transp. Eng. 131*, 771–779.

Karato, S.-I. (1993). Inner Core Anisotropy Due to the Magnetic Field-induced Preferred Orientation of Iron. *Science 262*(5140), 1708–1711.

Karato, S.-I. (1998). Seismic anisotropy in the deep mantle, boundary layers and the geometry of mantle convection. In J. Plomerová, R. C. Liebermann, and V. Babuška (Eds.), *Geodynamics of Lithosphere & Earth's Mantle*, Pageoph Topical Volumes, pp. 565–587. Birkhäuser Basel.

Karato, S.-I. (1999). Seismic anisotropy of the Earth's inner core resulting from flow induced by Maxwell stresses. *Nature 402*(6764), 871–873.

Karato, S.-I. (2008). *Deformation of Earth materials: An introduction to the Rheology of Solid Earth*. Cambridge, UK: Cambridge University Press.

Karato, S.-I., S. Zhang, and H.-R. Wenk (1995). Superplasticity in Earth's Lower Mantle: Evidence from Seismic Anisotropy and Rock Physics. *Science 270*(5235), 458–461.

Karki, B., L. Stixrude, S. Clark, M. Warren, G. Ackland, and J. Crain (1997). Elastic properties of orthorhombic mgsio3 perovskite at lower mantle pressures. *American Mineralogist 82*(5), 635–638.

Karki, B. B., R. M. Wentzcovitch, S. de Gironcoli, and S. Baroni (2000, Apr). High-pressure lattice dynamics and thermoelasticity of MgO. *Phys. Rev. B 61*, 8793–8800.

Käufl, P. J., A. Fichtner, and H. Igel (2013). Probabilistic full waveform inversion based on tectonics regionalisation - Development and application to the Australian upper mantle . *Geophys. J. Int. 193*, 437–451.

Käufl, P. J., A. P. Valentine, R. W. L. de Wit, and J. Trampert (2015a). Robust and fast probabilistic source parameter estimation from near-field displacement waveforms using pattern recognition. *submitted to Bulletin of the Seismological Society of America*.

Käufl, P. J., A. P. Valentine, R. W. L. de Wit, and J. Trampert (2015b). Solving probabilistic inverse problems rapidly with prior samples. *submitted to Geophysical Journal International*.

Käufl, P. J., A. P. Valentine, T. B. O'Toole, and J. Trampert (2014). A framework for fast probabilistic centroid-moment-tensor determination—inversion of regional static displacement measurements. *Geophysical Journal International 196*(3), 1676–1693.

Kellogg, L. H., B. H. Hager, and R. D. van der Hilst (1999). Compositional Stratification in the Deep Mantle. *Science 283*(5409), 1881–1884.

Kennett, B. (1998). On the density distribution within the Earth. *Geophys. J. Int. 132*, 374–382.

Kennett, B. (2006). On seismological reference models and the perceived nature of heterogeneity. *Physics of the Earth and Planetary Interiors 159*, 129 – 139.

Kennett, B. and E. R. Engdahl (1991). Travel times for global earthquake location and phase association. *Geophys. J. Int. 105*, 429–465.

Kennett, B., E. R. Engdahl, and R. Buland (1995). Constraints on seismic velocities in the Earth from travel times. *Geophys. J. Int. 122*, 108–124.

Kiefer, B., L. Stixrude, and R. M. Wentzcovitch (2002). Elasticity of (mg,fe)sio3-perovskite at high pressures. *Geophysical Research Letters 29*(11), 34–1–34–4.

Kirkpatrick, S., C. D. Gelatt, and M. P. Vecchi (1983). Optimization by Simulated Annealing. *Science 220*(4598), 671–680.

Kobayashi, Y., T. Kondo, E. Ohtani, N. Hirao, N. Miyajima, T. Yagi, T. Nagase, and T. Kikegawa (2005). Fe-mg partitioning between (mg, fe)sio_3 post-perovskite, perovskite, and magnesiowüstite in the earth's lower mantle. *Geophysical Research Letters 32*(19), n/a–n/a.

Koelemeijer, P., A. Deuss, and J. Ritsema (2013). Observations of core-mantle boundary Stoneley modes. *Geophys. Res. Lett. 40*(11), 2557–2561.

Koelemeijer, P. J. (2014). *Normal mode studies of long wavelength structures in Earth's lowermost mantle*. Ph. D. thesis, University of Cambridge, Cambridge, UK.

Koper, K. and M. Dombrovskaya (2005). Seismic properties of the inner core boundary from PKiKP/P amplitude ratios. *Earth Planet. Res. Lett. 237*, 680–694.

Kroonenberg, S. (2013). *Why Hell Stinks of Sulfur: Mythology and Geology of the Underworld*. Reaktion Books.

Kullback, S. and R. A. Leibler (1951, 03). On information and sufficiency. *Ann. Math. Statist. 22*(1), 79–86.

Kustowski, B., G. Ekström, and A. M. Dziewoński (2008). Anisotropic shear-wave velocity structure of the Earth's mantle: A global model. *Journal of Geophysical Research: Solid Earth 113*(B6).

Langer, H., G. Nunnari, and L. Occhipinti (1996). Estimation of seismic waveform governing parameters with neural networks. *Journal of Geophysical Research: Solid Earth 101*(B9), 20109–20118.

Laplace, P. S. Mémoire sur la probabilité des cause par les événments. *Mémoires de l'Académie royale des sciences de Paris (Savants étrangers) 6*, 359–378.

Laplace, P. S. (1812). *Théorie analytique des probabilités*. Paris: Courcier Imprimeur.

Larochelle, H., Y. Bengio, J. Louradour, and P. Lamblin (2009). Exploring strategies for training deep neural networks. *The Journal of Machine Learning Research 10*, 1–40.

LeCun, Y., L. Bottou, G. Orr, and K. Muller (1998). Efficient BackProp. In G. Orr and K. Muller (Eds.), *Neural Networks: Tricks of the trade*. Springer.

Lee, S., J.-H. Ruy, J.-S. Won, and H.-J. Park (1998). Determination and application of the weights for landslide susceptibility mapping using an artificial neural network. *Engineering Geology 71*, 289–302.

Lehmann, I. (1936). P'. *Publications du Bureau Central Séismologique International A14*(3), 87–115.

Lekić, V., J. Matas, M. Panning, and B. Romanowicz (2009). Measurement and implications of frequency dependence of attenuation. *Earth and Planetary Science Letters 282*, 285 – 293.

Li, C., R. D. van der Hilst, E. R. Engdahl, and S. Burdick (2008). A new global model for P wave speed variations in Earth's mantle. *Geochem. Geophys. Geosyst 9*(5).

Li, L., D. J. Weidner, J. Brodholt, D. Alfè, G. D. Price, R. Caracas, and R. Wentzcovitch (2006). Elasticity of $CaSiO_3$ perovskite at high pressure and high temperature. *Physics of the Earth and Planetary Interiors 155*(3–4), 249 – 259.

Li, X. and V. F. Cormier (2002). Frequency dependent attenuation in the inner core: Part I. A viscoelastic interpretation. *J. Geophys. Res. 107*.

Lin, J.-F., S. Speziale, Z. Mao, and H. Marquardt (2013). Effects of the electronic spin transitions of iron in lower mantle minerals: Implications for deep mantle geophysics and geochemistry. *Reviews of Geophysics 51*(2), 244–275.

Love, A. E. H. (1927). *A Treatise on the Mathematical Theory of Elasticity*. New York: Cambridge University Press.

Lythgoe, K., A. Deuss, J. Rudge, and J. Neufeld (2014). Earth's inner core: Innermost inner core or hemispherical variations? *Earth and Planetary Science Letters 385*(0), 181 – 189.

MacKay, D. (2003). *Information Theory, Inference, and Learning Algorithms*. Cambridge Univ Press.

MacKay, D. J. C. (1992). Bayesian interpolation. *Neural computation 4*(3), 415–447.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, Volume 1, pp. 281–297. Oakland, CA, USA.

Mainprice, D. (2007). 2.16 - seismic anisotropy of the deep earth from a mineral and rock physics perspective. In G. Schubert (Ed.), *Treatise on Geophysics*, pp. 437 – 491. Amsterdam: Elsevier.

Mainprice, D., G. Barruol, and W. Ben Ismaïl (2000). The anisotropy of the Earth's mantle : From single crystal to polycrystal. In S. I. Karato, A. M. Forte, R. C. Liebermann, G. Masters, and L. Stixrude (Eds.), *Earth's deep interior. Mineral physics and tomography from the atomic to the global scale*, Volume 117 of *Geophysical Monograph Series*, pp. 237–264. Washington, D.C.: AGU.

Malinverno, A. and V. A. Briggs (2004). Expanded uncertainty quantification in inverse problems: Hierarchical Bayes and empirical Bayes. *Geophysics 69*(4), 1005–1016.

Mao, H., P. Bell, and T. Yagi (1979). Iron-magnesium fractionation model for the earth. *Carnegie Inst. Wash. Yearb 78*, 621–625.

Masters, G., M. Barmine, and S. Kientz (2011). *Mineos: user manual version 1.0.2*. Pasadena, CA: Calif. Inst. of Tech.

Masters, G. and D. Gubbins (2003). On the resolution of density within the Earth. *Phys. Earth Planet. Inter. 140*, 159–167.

McCulloch, W. and W. Pitts (1943). A logical calculus of ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics 5*, 115–133.

McLachlan, G. J. and K. E. Basford (1988). *Mixture Models: Inference and Applications to Clustering*. New York: Marcel Dekker.

McNamara, A. K., P. E. van Keken, and S.-I. Karato (2002). Development of anisotropic structure in the earth's lower mantle by solid-state convection. *Nature 416*(6878), 310–314.

Meade, C., P. G. Silver, and S. Kaneshima (1995). Laboratory and seismological observations of lower mantle isotropy. *Geophysical Research Letters 22*(10), 1293–1296.

Meier, U., A. Curtis, and J. Trampert (2007a). Fully nonlinear inversion of fundamental mode surface waves for a global crustal model. *Geophys. Res. Lett. 34*.

Meier, U., A. Curtis, and J. Trampert (2007b). Global crustal thickness from neural network inversion of surface wave data. *Geophys. J. Int. 169*, 706–722.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics 21*(6), 1087–1092.

Minsky, M. and S. Papert (1969). *Perceptrons*. Cambridge, MA, USA: MIT Press.

Mohorovičić, A. (1910). Potres od 8. X. 1909. *Godišnje izvješć zagrebačkog meteorološkog opervatorija za godinu 1909*, 1–56.

Mohorovičić, A. (1992). "Earthquake of 8 October 1909" (English translation). *Geofizika 9*(1), 3–55.

Møller, M. F. (1993). A scaled conjugate gradient algorithm for fast supervised learning. *Neural Networks 6*, 525–533.

Montagner, J.-P. (1994). Can seismology tell us anything about convection in the mantle? *Reviews of Geophysics 32*(2), 115–137.

Montagner, J.-P. and D. L. Anderson (1989). Petrological constraints on seismic anisotropy. *Physics of the Earth and Planetary Interiors 54*(1-2), 82 – 105.

Montagner, J.-P. and B. L. N. Kennett (1996). How to reconcile body-wave and normal-mode reference earth models. *Geophys. J. Int. 125*, 229–248.

Morelli, A., A. M. Dziewonski, and J. H. Woodhouse (1986). Anisotropy of the inner core inferred from PKIKP travel times. *Geophysical Research Letters 13*(13), 1545–1548.

Mosegaard, K. and M. Sambridge (2002). Monte Carlo analysis of inverse problems. *Inverse Problems 18*, R29–R54.

Mosegaard, K. and A. Tarantola (1995). Monte Carlo sampling of solutions to inverse problems. *J. Geophys. Res. 100*, 12431–12447.

Mosegaard, K. and A. Tarantola (2002). Probabilistic approach to inverse problems. *International Geophysics Series 81*(A), 237–268.

Mosegaard, K. and P. D. Vestergaard (1991). A Simulated Annealing Approach to Seismic Model Optimization with Sparse Prior Information. *Geophysical Prospecting 39*(5), 599–611.

Müller, B., J. Reinhardt, and M. T. Strickland (1995). *Neural Networks: An Introduction* (2 ed.). Springer.

Murakami, M., Y. Ohishi, N. Hirao, and K. Hirose (2012, MAY 3). A perovskitic lower mantle inferred from high-pressure, high-temperature sound velocity data. *Nature 485*(7396), 90–94.

Myers, R. H., D. C. Montgomery, and C. M. Anderson-Cook (2009). *Response surface methodology: process and product optimization using designed experiments*, Volume 705. John Wiley & Sons.

Nabney, I. T. (2002). *Netlab: Algorithms for Pattern Recognition*. Advances for Pattern Recognition. London, UK: Springer-Verlag.

Neal, R. M. (1994). *Bayesian Learning for Neural Networks*. Ph. D. thesis, University of Toronto, Toronto, Canada.

Neal, R. M. (2011). MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo 2*.

Nishiyama, N., T. Yagi, S. Ono, H. Gotou, T. Harada, and T. Kikegawa (2007). Effect of incorporation of iron and aluminum on the thermoelastic properties of magnesium silicate perovskite. *Physics and Chemistry of Minerals 34*(3), 131–143.

Nolet, G. (2008). A breviary of seismic tomography: imaging the interior of the earth and sun. *Cambridge University Press*, 344pp.

Nowacki, A., J. Wookey, and J.-M. Kendall (2011). New advances in using seismic anisotropy, mineral physics and geodynamics to understand deformation in the lowermost mantle. *Journal of Geodynamics 52*(3-4), 205 – 228.

Odom, M. and R. Sharda (2002). A neural network model for bankruptcy prediction. In *Proceedings of the international joint conference on neural networks*, Volume 2, pp. 163–168. Alamitos, CA, USA: IEEE Press.

Oganov, A. R., J. P. Brodholt, and G. D. Price (2001). The elastic constants of MgSiO_3 perovskite at pressures and temperatures of the Earth's mantle. *Nature 411*, 934–937.

Okal, E. A. and B.-G. Jo (1990). Q measurements for phase X overtones. *Pure Appl. Geophys. 132*, 331–362.

Oldham, R. D. (1906). The constitution of the interior of the Earth, as revealed by earthquakes. *Quarterly Journal of the Geological Society 62*(1-4), 456–475.

Ono, S. and A. R. Oganov (2005). In situ observations of phase transition between perovskite and cairo_3-type phase in mgsio_3 and pyrolitic mantle composition. *Earth and Planetary Science Letters 236*(3–4), 914–932.

Panning, M. and B. Romanowicz (2004). Inference on flow at the base of Earth's mantle based on seismic anisotropy. *Science 303*, 351–353.

Panning, M. and B. Romanowicz (2006). A three-dimensional radially anisotropic model of shear velocity in the whole mantle. *Geophysical Journal International 167*(1), 361–379.

Parker, R. L. (1994). *Geophysical Inverse Theory*. Princeton University Press.

Popper, K. R. (1963). *Conjectures and refutations: The growth of scientific knowledge*. New York],: Routledge.

Poulton, M. M. (2001). *Computational Neural Networks for Geophysical Data Processing*. Amsterdam, The Netherlands: Pergamon.

Poulton, M. M. (2002). Neural networks as an intelligence amplification tool: A review of applications. *Geophysics 67*, 979–993.

Poupinet, G., R. Pillet, and A. Souriau (1983). Possible heterogeneity of the Earth's core deduced from PKIKP travel times. *Nature 305*(5931), 204–206.

Press, F. (1968). Earth models obtained by Monte Carlo Inversion. *J. Geophys. Res. 73*(16), 5223–5234.

Queipo, N. V., R. T. Haftka, W. Shyy, T. Goel, R. Vaidyanathan, and P. K. Tucker (2005). Surrogate-based analysis and optimization. *Progress in aerospace sciences 41*(1), 1–28.

Rawlinson, N., S. Pozgay, and S. Fishwick (2010). Seismic tomography: A window into deep Earth. *Phys. Earth Planet. Inter. 178*, 101–135.

Razavi, S., B. A. Tolson, and D. H. Burn (2012). Review of surrogate modeling in water resources. *Water Resources Research 48*(7).

Reference Earth Model web pages (2001). http://igppweb.ucsd.edu/~gabi/rem.dir/surface/tmodes.list.

Resovsky, J. and J. Trampert (2003). Using probabilistic seismic tomography to test mantle velocity-density relationships. *Earth Planet. Res. Lett. 215*, 121–134.

Resovsky, J., J. Trampert, and R. D. van der Hilst (2005). Error bars for the global seismic Q profile. *Earth Planet. Res. Lett. 230*, 413–423.

Resovsky, J. S. and M. H. Ritzwoller (1999). Regularization uncertainty in density models estimated from normal mode data. *Geophysical Research Letters 26*(15), 2319–2322.

Ritsema, J., A. Deuss, H. J. van Heijst, and J. H. Woodhouse (2011). S40RTS: a degree-40 shear-velocity model for the mantle from new Rayleigh wave dispersion, teleseismic traveltime and normal-mode splitting function measurements. *Geophys. J. Int. 184*, 1223–1236.

Rochester, N., J. Holland, L. Haibt, and W. Duda (1956, September). Tests on a cell assembly theory of the action of the brain, using a large digital computer. *Information Theory, IRE Transactions on 2*(3), 80–93.

Romanowicz, B. (2003). Global mantle tomography: Progress status in the past 10 years. *Ann. Rev. Earth Planet. Sci. 31*, 303–328.

Romanowicz, B. and B. Mitchell (2007). Q in the Earth from crust to core. In G. Schubert (Ed.), *Seismology and the Structure of the Earth*, Volume 1, pp. 731–774. Elsevier, Amsterdam.

Romanowicz, B., H. Tkalčić, and L. Bréger (2003). *On the Origin of Complexity in PKP Travel Time Data*, pp. 31–44. American Geophysical Union.

Rosenblatt, F. (1958). The Perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review 64*, 386–408.

Röth, G. and A. Tarantola (1994). Neural networks and inversion of seismic data. *J. Geophys. Res. 99*, 67536768.

Rowley, H. A. (1998). Neural network-based face detection . *IEEE Transactions on Pattern Analysis and Machine Intelligence 20*, 23–38.

Rumelhart, D. E., G. E. Hinton, and R. J. Williams (1986). Learning internal representations by error propagation. In *Parallel distributed processing: explorations in the microstructure of cognition*, Volume 1: foundations, pp. 318–362. Cambridge, MA, USA: MIT Press.

Russell, S. and P. Norvig (2009). *Artificial Intelligence: A Modern Approach* (3rd ed.). Upper Saddle River, NJ, USA: Prentice Hall Press.

Sabatier, P. C. (2000). Past and future of inverse problems. *Journal of Mathematical Physics 41*(6), 4082–4124.

Sambridge, M. (1999a). Geophysical Inversion with a Neighbourhood Algorithm -I. Searching a parameter space. *Geophys. J. Int. 138*, 479–494.

Sambridge, M. (1999b). Geophysical Inversion with a Neighbourhood Algorithm -II. Appraising the ensemble. *Geophys. J. Int. 138*, 727–746.

Sambridge, M., T. Bodin, K. Gallagher, and H. Tkalčić (2013). Transdimensional inference in the geosciences. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 371*(1984), 20110547.

Sambridge, M., K. Gallagher, A. Jackson, and P. Rickwood (2006). Trans-dimensional inverse problems, model comparison and the evidence. *Geophysical Journal International 167*(2), 528–542.

Sambridge, M. and K. Mosegaard (2002). Monte Carlo methods in geophysical inverse problems. *Rev. Geophys. 40*, 1–29.

Scales, J. A., M. L. Smith, and S. Treitel (2001). Introductory geophysical inverse theory.

Scales, J. A. and R. Snieder (1997). To Bayes or not to Bayes? *Geophysics 62*(4), 1045–1046.

Scales, J. A. and R. Snieder (1998). What is noise? *Geophysics 63*(4), 1122–1124.

Scales, J. A. and R. Snieder (2000). The anatomy of inverse problems. *Geophysics 65*(6), 1708–1710.

Scales, J. A. and L. Tenorio (2001). Prior information and uncertainty in inverse problems. *Geophysics 66*(2), 389–397.

Schmidhuber, J. (2014). Deep Learning in Neural Networks: An Overview. *arXiv preprint arXiv:1404.7828*.

Shahraeeni, M. S. and A. Curtis (2011). Fast probabilistic nonlinear petrophysical inversion. *Geophysics 76*, 45–58.

Shahraeeni, M. S., A. Curtis, and G. Chao (2012). Fast probabilistic petrophysical mapping of reservoirs from 3D seismic data. *Geophysics 77*, 1–19.

Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal 27*, 379–423, 623–656.

Shearer, P. and G. Masters (1990). The density and shear velocity contrast at the inner core boundary. *Geophys. J. Int. 102*(2), 491–498.

Shearer, P. M. (1994). Constraints on inner core anisotropy from PKP(DF) travel times. *Journal of Geophysical Research: Solid Earth 99*(B10), 19647–19659.

Shen, W., M. H. Ritzwoller, and V. Schulte-Pelkum (2013). A 3-D model of the crust and uppermost mantle beneath the Central and Western US by joint inversion of receiver functions and surface wave dispersion. *Journal of Geophysical Research: Solid Earth 118*(1), 262–276.

Shen, W., M. H. Ritzwoller, V. Schulte-Pelkum, and F.-C. Lin (2013). Joint inversion of surface wave dispersion and receiver functions: a Bayesian Monte-Carlo approach. *Geophys. J. Int. 192*(2), 807–836.

Sinmyo, R. and K. Hirose (2013). Iron partitioning in pyrolitic lower mantle. *Physics and Chemistry of Minerals 40*(2), 107–113.

Sisson, S. A. (2005). Transdimensional Markov chains: A decade of progress and future perspectives. *Journal of the American Statistical Association 100*(471), 1077–1089.

Sivia, D. S. (1996). *Data Analysis: A Bayesian Tutorial*. Oxford: Clarendon (Oxford Univ. Press).

Skilling, J. (2006). Nested sampling for general Bayesian computation. *Bayesian Analysis 1*(4), 833–859.

Snieder, R. and J. Trampert (1999). *Inverse problems in geophysics*. Springer.

Song, X. and D. V. Helmberger (1993). anisotropy of earth's inner core. *Geophysical Research Letters 20*(23), 2591–2594.

Song, X. and D. V. Helmberger (1995). Depth dependence of anisotropy of Earth's inner core. *Journal of Geophysical Research: Solid Earth 100*(B6), 9805–9816.

Stein, S. and M. E. Wysession (2003). *An introduction to seismology, earthquakes, and Earth structure*. Wiley-Blackwell.

Stixrude, L. and R. E. Cohen (1995). High-Pressure Elasticity of Iron and Anisotropy of Earth's Inner Core. *Science 267*(5206), 1972–1975.

Su, W.-J. and A. M. Dziewonski (1995). Inner core anisotropy in three dimensions. *Journal of Geophysical Research: Solid Earth 100*(B6), 9831–9852.

Tarantola, A. (2005). *Inverse Problem Theory and Methods for Model Parameter Estimation*. SIAM.

Tarantola, A. and B. Valette (1982). Inverse Problems = Quest for Information. *Journal of Geophysics 50*, 159–170.

Tateno, S., K. Hirose, Y. Ohishi, and Y. Tatsumi (2010). The structure of iron in Earth's inner core. *Science 330*(6002), 359–361.

Tkalčić, H. and B. L. N. Kennett (2008). Core structure and heterogeneity: a seismological perspective. *Australian Journal of Earth Sciences 55*(4), 419–431.

Tkalčić, H., B. L. N. Kennett, and V. F. Cormier (2009). On the inner-outer core density contrast from PKiKP/PcP amplitude ratios and uncertainties caused by seismic noise. *Geophys. J. Int. 179*(1), 425–443.

Trampert, J. (1998). Global seismic tomography: the inverse problem and beyond. *Inverse Problems 14*(3), 371.

Trampert, J., F. Deschamps, J. Resovsky, and D. Yuen (2004). Probabilistic tomography maps chemical heterogeneities throughout the lower mantle. *Science 306*, 853–856.

Trampert, J., A. Fichtner, and J. Ritsema (2013). Resolution tests revisited: the power of random numbers. *Geophysical Journal International 192*(2), 676–680.

Trampert, J. and R. van der Hilst (2005). Towards a quantitative interpretation of global seismic tomography. In R. D. van der Hilst, J. D. Bass, J. Matas, and J. Trampert (Eds.), *Earth's Deep Mantle: Structure, Composition, and Evolution*, Volume 160, pp. 47–62. Washintgon, D.C., USA: AGU.

Tromp, J. (1993). Support for anisotropy of the Earth's inner core from free oscillations. *Nature 366*, 678–681.

Tromp, J. (1995). Normal-mode splitting due to inner-core anisotropy. *Geophysical Journal International 121*(3), 963–968.

Tromp, J., D. Komattisch, and Q. Liu (2008). Spectral-element and adjoint methods in seismology. *Communications in Computational Physics 3*(1), 1–32.

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

Valentine, A. P., L. M. Kalnins, and J. Trampert (2013). Discovery and analysis of topographic features using learning algorithms: A seamount case study. *Geophysical Research Letters 40*(12), 3048–3054.

Valentine, A. P. and J. Trampert (2012a). Assessing the uncertainties on seismic source parameters: Towards realistic error estimates for centroid-moment-tensor determinations. *Phys. Earth Planet. Inter. 210–211*, 36–49.

Valentine, A. P. and J. Trampert (2012b). Data-space reduction, quality assessment and searching of seismograms: Autoencoder networks for waveform data. *Geophys. J. Int. 189*, 1183–1202.

Valentine, A. P. and J. H. Woodhouse (2010). Approaches to automated data selection for global seismic tomography. *Geophys. J. Int. 182*, 1001–1012.

van der Baan, M. and C. Jutten (2000). Neural networks in geophysical applications. *Geophysics 65*, 1032–1047.

Vasco, D. W., L. R. Johnson, and O. Marques (2003). Resolution, uncertainty, and whole Earth tomography. *J. Geophys. Res. 108*(B1), 2022.

Verhoeven, O., A. Mocquet, P. Vacher, A. Rivoldini, M. Menvielle, P.-A. Arrial, G. Choblet, P. Tarits, V. Dehant, and T. Van Hoolst (2009). Constraints on thermal state and composition of the Earth's lower mantle from electromagnetic impedances and seismic data. *Journal of Geophysical Research: Solid Earth 114*(B3).

Vinnik, L., B. Romanowicz, and L. Breger (1994). Anisotropy in the center of the inner core. *Geophysical Research Letters 21*(16), 1671–1674.

Virieux, J. and S. Operto (2009). An overview of full-waveform inversion in exploration geophysics. *GEOPHYSICS 74*(6), WCC1–WCC26.

Visser, K., J. Trampert, S. Lebedev, and B. Kennett (2008). Probability of radial anisotropy in the deep mantle. *Earth and Planetary Science Letters 270*(3–4), 241 – 250.

Vočadlo, L., D. Alfe, M. J. Gillan, and G. D. Price (2003). The properties of iron under core conditions from first principles calculations. *Physics of the Earth and Planetary Interiors 140*(1), 101–125.

Voigt, W. (1910). *Lehrbuch der kristallphysik:(mit ausschluss der kristalloptik)*, Volume 34. BG Teubner.

Walker, A. M., A. M. Forte, J. Wookey, A. Nowacki, and J.-M. Kendall (2011). Elastic anisotropy of D¨ predicted from global models of mantle flow. *Geochemistry, Geophysics, Geosystems 12*(10).

Walker, A. M. and J. Wookey (2012). Msat: new toolkit for the analysis of elastic and seismic anisotropy. *Computers & Geosciences 49*(0), 81 – 90.

Walker, M. and A. Curtis (2014). Varying prior information in Bayesian inversion. *Inverse Problems 30*(6).

Wang, T., X. Song, and H. H. Xia (2015). Equatorial anisotropy in the inner part of Earth's inner core from autocorrelation of earthquake coda. *Nature Geoscience*.

Waszek, L. and A. Deuss (2013). A low attenuation layer in the Earth's uppermost inner core. *Geophys. J. Int.*.

Waszek, L., J. Irving, and A. Deuss (2011). Reconciling the hemispherical structure of Earth's inner core with its super-rotation. *Nature Geoscience 4*, 264–267.

Wenk, H.-R., S. Speziale, A. McNamara, and E. Garnero (2006). Modeling lower mantle anisotropy development in a subducting slab. *Earth and Planetary Science Letters 245*(1-2), 302 – 314.

Wentzcovitch, R. M., B. B. Karki, M. Cococcioni, and S. de Gironcoli (2004, Jan). Thermoelastic Properties of $MgSiO_3$-Perovskite: Insights on the Nature of the Earth's Lower Mantle. *Phys. Rev. Lett. 92*, 018501.

Werbos, P. J. (1974). *Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences*. Ph. D. thesis, Harvard University, Cambridge, MA, USA.

Widmer, R. (1991). *The large-scale structure of the deep Earth as constrained by free oscillation observations*. Ph. D. thesis, University of California, San Diego, CA, USA.

Widmer, R., G. Masters, and F. Gilbert (1991). Spherically symmetric attenuation within the Earth from normal mode data. *Geophys. J. Int. 104*, 541–553.

Widrow, B. and M. E. Hoff Jr. (1960). Adaptive Switching Circuits. In *IRE WESCON Convention Record*, pp. 96–104. Institute of Radio Engineers.

Wiggins, R. A. (1969). Monte Carlo inversion of body-wave observations. *Journal of Geophysical Research 74*(12), 3171–3181.

Williams, P. M. (1996). Using neural networks to model conditional multivariate densities. *Neural Computation 8*, 843–854.

Wood, R. and A. Curtis (2004). Geological prior information and its applications to geoscientific problems. *Geological Society, London, Special Publications 239*(1), 1–14.

Woodhouse, J. H. and A. M. Dziewonski (1984). Mapping the upper mantle: Three-dimensional modeling of Earth structure by inversion of seismic waveforms. *Journal of Geophysical Research: Solid Earth (1978–2012) 89*(B7), 5953–5986.

Woodhouse, J. H., D. Giardini, and X.-D. Li (1986). Evidence for inner core anisotropy from free oscillations. *Geophysical Research Letters 13*(13), 1549–1552.

Yoshida, S., I. Sumita, and M. Kumazawa (1996). Growth model of the inner core coupled with the outer core dynamics and the resulting elastic anisotropy. *Journal of Geophysical Research: Solid Earth (1978–2012) 101*(B12), 28085–28103.

Zhu, H., E. Bozdağ, D. Peter, and J. Tromp (2012). Structure of the European upper mantle revealed by adjoint tomography. *Nature Geoscience 5*, 493–498.